Year : 2023

Modern Statistical and Causal Approaches to Psychology

Vowels Matthew

Vowels Matthew, 2023, Modern Statistical and Causal Approaches to Psychology

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_7E8014AE56941

UNIL | Université de Lausanne

FACULTÉ DES SCIENCES SOCIALES ET POLITIQUES

INSTITUT DE PSYCHOLOGIE

Modern Statistical and Causal Approaches to Psychology

THÈSE DE DOCTORAT

présentée à la

Faculté des sciences sociales et politiques
de l'Université de Lausanne

pour l'obtention du grade de

Docteur en mathématiques appliquées aux sciences humaines et sociales

par

Dr. Matthew Vowels

Directeur de thèse:
Prof. Dr. Peter Hilpert

Jury:
Prof. Dr. André Berchtold
Prof. Dr. Paolo Ghisletta

LAUSANNE
2023

UNIL | Université de Lausanne

FACULTÉ DES SCIENCES SOCIALES ET POLITIQUES

INSTITUT DE PSYCHOLOGIE

Modern Statistical and Causal Approaches to Psychology

THÈSE DE DOCTORAT

présentée à la

Faculté des sciences sociales et politiques
de l'Université de Lausanne

pour l'obtention du grade de

Docteur en mathématiques appliquées aux sciences humaines et sociales

par

Dr. Matthew Vowels

Directeur de thèse:
Prof. Dr. Peter Hilpert

Jury:
Prof. Dr. André Berchtold
Prof. Dr. Paolo Ghisletta

LAUSANNE
2023

UNIL | Université de Lausanne
Faculté des sciences sociales et politiques

IMPRIMATUR

Le Décanat de la Faculté des sciences sociales et politiques de l'Université de Lausanne, au nom du Conseil et sur proposition d'un jury formé des professeurs

- M. Peter HILPERT, Professeur, Directeur de thèse
- M. André BERCHTOLD, Professeur à l'Université de Lausanne
- M. Paolo GHISLETTA, Professeur à l'Université de Genève

autorise, sans se prononcer sur les opinions du candidat, l'impression de la thèse de Monsieur Matthew VOWELS, intitulée :

« Modern Statistical and Causal Approaches to Psychology. »

Nicky LE FEUVRE
Doyenne

Lausanne, le 13 juin 2023

Abstract in English: Psychology and social science face a number of challenges: the inherent complexity of the phenomena of interest, the replication crisis, the theory crisis, and functional and structural misspecification. The confluence of these challenges poses a serious threat to the validity and meaningfulness of research in these domains, and brings into question the direction that researchers in these fields should take. If this direction is to be effective with respect to improvement, I believe it is important for psychologists and social scientists to engage with the meta-research surrounding areas of possible analytical and statistical improvements. In this thesis I present four contributions which are strongly motivated by the problems of misspecification and complexity, an provide recommendations to researchers. The proposals include the use of more powerful, data-adaptive techniques for function approximation (such as those tools from the domain of machine learning), as well as the use of techniques from the domain of causality (such as causal Directed Acyclic Graphs) and information theory. I demonstrate how these techniques can (a) help us to match the level of complexity of our modeling to the inherent complexity of the phenomenon under study, (b) reduce ambiguity with respect to theory specification, and make our assumptions and modeling choices more transparent, (c) reduce the complexity of a mathematical representation of a theory without impacting the validity of any downstream estimates, (d) improve the efficiency of data collection methodologies, (e) highlight the critical nature of causality even when otherwise powerful, exploratory machine learning techniques are used, (f) highlight the strange, unintuitive behaviour of datasets with more than four dimensions, (g) undertake outlier detection in a way that is robust to this aforementioned strange behaviour. Whilst these approaches do not solve the problems in a finite sense, they represent relatively low-cost stepping stones en route to a better way to undertake research in psychology and social science. Indeed, the nature of at least some of the problems would seem to encourage an optimistic interpretation: That there presently exists a tremendous opportunity to modernise the current approach to research, simply by assimilating recent advances and developments from other domains such as engineering, machine learning, and statistics. Psychology and social science are complex domains, full of rich and nuanced phenomena. They deserve to be represented using research methodologies which are flexible enough to reflect this complexity.

Résumé en Français : La psychologie et les sciences sociales sont confrontées à un certain nombre de défis : la complexité des phénomènes auxquels celles-ci s'intéressent, la crise de la réplication, la crise de la théorie et les erreurs de spécification fonctionnelles et structurelles. La confluence de ces défis constitue une menace sérieuse pour la validité de la recherche et pour sa capacité à donner du sens; cela remet de plus en question la direction que les chercheurs en psychologie et en sciences sociales devraient suivre. Si ce cheminement de la recherche doit être efficace en ce qui concerne les améliorations dans ces domaines, il s'agit ici de soutenir qu'il est important que les chercheurs en psychologie et en sciences sociales s'engagent dans une méta-recherche balisant les améliorations analytiques et statistiques possibles. Cette thèse présente quatre travaux fortement motivés par la volonté de résoudre les problèmes de complexité et d'erreur de spécification, et elle s'engagera aussi à faire des recommandations relatives à ces questions aux chercheurs. Les propositions incluent l'utilisation de techniques pour l'approximation de fonction plus puissantes (tels que les outils du domaine de l'apprentissage automatique), l'utilisation de techniques issues du domaine de la causalité (tels que les graphes acycliques dirigés causaux) et enfin des idées relatives à la théorie de l'information. Il s'agira tour à tour de démontrer comment ces techniques peuvent: A) nous aider à faire correspondre le niveau de complexité de notre modélisation à la complexité inhérente au phénomène étudié; B) réduire l'ambiguïté liée à la spécification de la théorie et rendre nos hypothèses et nos choix de modélisation plus transparents; C) réduire la complexité de la représentation mathématique d'une théorie sans affecter la validité des estimations de cette dernière; D) améliorer l'efficacité des méthodologies de collecte de données; E) mettre en exergue la nature critique de la causalité même lorsque de puissantes techniques exploratoires d'apprentissage automatique sont utilisées; F) mettre en évidence le comportement étrange et non intuitif des ensembles de données de plus de quatre dimensions; G) en cas de comportement étrange ou non, entreprendre la détection des valeurs aberrantes d'une manière qui soit robuste. Bien que ces approches ne résolvent pas définitivement les problèmes, elles représentent des étapes peu coûteuses et faciles à franchir pour réaliser une meilleure façon d'entreprendre des recherches. La nature de certains problèmes au moins, telle que l'omniprésence de méthodologies de recherche et de méthodes d'analyse peu sophistiquées dans le paradigme actuel, semble en effet encourager une interprétation optimiste : il existe actuellement de formidables opportunités de moderniser l'approche actuelle de la recherche. La psychologie et les sciences sociales sont des domaines complexes, riches en phénomènes dynamiques. Ceux-ci méritent d'être étudiés à l'aide de méthodologies de recherche suffisamment flexibles pour refléter toute cette complexité.

Extended Abstract

Psychology and social science face a number of challenges: the inherent complexity of the phenomena of interest, the replication crisis, the theory crisis, and as I discuss, functional and structural misspecification. The confluence of these challenges poses a serious threat to the validity and meaningfulness of research in these domains, and brings into question the direction that researchers in these fields should take. If this direction is to be positive with respect to improvements in the domains, I believe it is extremely important for psychologists and social scientists to engage with the meta-research surrounding areas of possible analytical and statistical improvements. In this thesis I present four contributions which are strongly motivated by the problems of misspecification and complexity, an provide recommendations to researchers.

In the first contribution, I focus on three issues that deserve more attention. Namely, the use of models with limited functional form, the use of misspecified causal models (misspecified either due to limited functional form, or incorrect structure), and unreliable interpretations of results. I demonstrate a number of consequences relating to these issues via simulation, and provide recommendations for researchers to improve their research practice, such as the use of techniques from the domains of machine learning and causality; engaging with experts in statistics, causality, and machine learning; being more transparent about the methodological and analytical approach; and be concise and not overambitious in the specification of their research questions and hypotheses.

Following the recommendations made in this first contribution, I also include two example applications of these recommendations. The first application involves the use of machine learning techniques to explore the relationships between partner support and relational and individuals variables. The second involves the use of causal discovery and causal inference tools (which themselves derive from the domain at the confluence of machine learning and causality) to explore the links between attachment styles and mental health during the COVID-19 pandemic. The purpose of these applications is to demonstrate that the recommendations made are not simply hypothetical, but can be readily applied.

Furthermore, and given my specific recommendation that researchers engage with techniques from the domain of causality, I make a second methodological/statistical contribution by exploring how causal graphs can be used to improve the efficiency of data collection process. On the one hand, it is important that we collect data in a way that maximises the validity of what we are measuring, which may involve the use of long scales with many items. On the other hand, collecting a large number of items across multiple scales results in participant fatigue, and expensive and time consuming data collection. It is therefore important that we use the available resources optimally. I consider how the representation of a theory as a causal/structural model can help us to streamline data collection and analysis procedures by not wasting time collecting data for variables which are not causally critical for answering the research question. This not only saves time and enables us to redirect resources to attend to other variables which are more important, but also increases research transparency and the reliability of theory testing. To demonstrate the benefits of this streamlining, I review the relevant concepts and present a number of didactic examples, including a real-world example.

In turn, given the recommendation that researcher engage with tools from the domain of machine learning techniques, in the third (technical) contribution I explore to what extent machine learning techniques are sensitive to the underlying causal structure in the data. Indeed, machine learning explainability techniques have been proposed as a means for psychologists to 'explain' or interrogate a model in order to gain an understanding about a phenomenon of interest. Researchers may be motivated to use machine learning algorithms in conjunction with explainability techniques, as part of exploratory research, with the goal of identifying important variables which are associated with / predictive of an outcome of interest. However, and as I demonstrate, machine learning algorithms are highly sensitive to the underlying causal structure in the data. The consequences of this are that predictors which are deemed by the explainability technique to be unrelated/unimportant/unpredictive, may actually be highly associated with the outcome. Rather than this being a limitation of explainability techniques *per se*, we show that it is rather a consequence of the mathematical implications of regression, and the interaction of these implications with the associated conditional independencies of the underlying causal structure. I provide some alternative recommendations for psychologists wanting to explore the data for important variables.

In the final contribution, I explore the unintuitive behaviour of datasets which attempt to accommodate the inherently high-dimensional complexity of psychological phenomena. In particular, I consider what the notion of 'normality' implies in high-dimensional settings. Normality, in the colloquial sense, has historically been considered an aspirational trait, synonymous with harmony and ideality. The arithmetic average has often been used to characterize normality, and is often used as a blunt way to characterize samples and outliers. I demonstrate that even for datasets with as few as four dimensions, data start to exhibit a number of peculiarities which become progressively severe as the number of dimensions increases. I show that normality can be better characterized with 'typicality', an information theoretic concept relating to entropy. An application of typicality to both synthetic and real-world data reveals that in multi-dimensional space, to be normal (or close to the mean) is actually to be highly atypical. This motivates us to update our working definition of an outlier, and we demonstrate typicality for outlier detection as a viable method which is consistent with this updated definition. In contrast, whilst the popular Mahalanobis based outlier detection method can be used to identify points far from the mean, it fails to identify those which are too close. Typicality can be used to achieve both, and performs well regardless of the dimensionality of the problem.

Whilst the proposals made in these four contributions do not solve the problems I identify in a finite sense, they represent relatively low-cost stepping stones en route to a better way to undertake research in psychology and social science. Indeed, the nature of at least some of the problems, such as the ubiquity of unsophisticated research methodologies and analytical methods in the current paradigm, would seem to encourage an optimistic interpretation: That there presently exists a tremendous opportunity to innovate and modernise the current approach to research, simply by assimilating recent advances and developments from other domains such as engineering, machine learning, and statistics. Psychology and social science are complex domains, full of rich and nuanced phenomena. They deserve to be represented and studied using research methodologies which are flexible enough to reflect this complexity.

Acknowledgements

My wife is the most important person who I thank and acknowledge in reference to this thesis. And not only in the cliché 'thanks to my wife for supporting' kind of way, but more in the 'I literally never thought I'd ever study psychology until I met you' kind of way. Without you I would have never started pursuing psychology in parallel to my engineering passion, and nor would I have noticed the significant opportunities and synergies that could exist at the intersection of engineering, machine learning, and psychology. I thank you also for your patience in discussing some of the finer and arguably more philosophical points about what can and can't be answered with respect to psychology, as well as your patience in dealing with my frustration with the state of academia as a whole. Sorry for putting you through a *second* lot of PhD ramblings...!

On a similar note, I'd like to thank Prof. Nathan Wood for his encouraging curiosity and support. It was actually this curiosity which acted as the catalyst between my wife, psychology as a domain of study, and my application to UKY.

On yet another similar note - I'd like to thank my supervisor Peter for his unwavering patience and understanding, as well as his support in creating opportunities (such as the one to move to Switzerland - which is a great privilege). If it weren't for this patience and support, I'd have probably stayed in engineering, in spite of the detour I took for my first (psychology-related) masters in the U.S., and in spite of our common vision for the future of analysis and research in psychology.

As always, I thank my Uncle for just being who he is. You've been with me since I was a child and you define much of who I am. I also thank my parents for always supporting and being there for me.

Finally, acknowledging the often critical nature of the transcripts in this thesis, I thank in advance the reader for open-minded patience and understanding, as I attempt to deconstruct the current *status quo* in psychology, as well as to present a different way of approaching what interests us as psychologists.

Contents

List of Figures

List of Tables

Declaration

This thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images or text resulting from the work of others (whether published or unpublished) are fully identified as such within the work and attributed to their originator in the text, bibliography or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification.

Signature:

Contributions of this thesis and affiliation of particular publications to chapters are specified in Chapter 1. Note that this thesis follows a 'by publication' format, such that the technical chapters are either presented in their original published form, or as they were submitted for review. The publications supporting this thesis are given below.

Chapter 2:

Vowels, M.J., 2021. Misspecification and Unreliable Interpretations in Psychology and Social Science. *Psychological Methods*. DOI: 10.1037/met0000429.

Chapter 3:

Vowels, L.M., **Vowels, M.J.**, Carnelley, K.B., Kumashiro, M., 2022. A machine learning approach to predicting perceived partner support from relational and individual variables. *Social Psychological and Personality Science*, DOI: 10.1177/19485506221114982.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

Chapter 4:

Vowels, L.M., **Vowels, M.J.**, Carnelley, K.B., Millings, A. Miller, J.G., Under Review. Toward a Causal Link between Attachment Styles and Mental Health during the COVID-19 Pandemic.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

Chapter 5:

Vowels, M.J., 2023. Prespecification of Structure for the Optimization of Data Collection and Analysis. *Collabra: Psychology*. DOI: 10.1525/collabra.71300

Chapter 6:

Vowels, M.J., Under Review. Trying to Outrun Causality with Machine Learning. DOI: TBC.

Chapter 7:

Vowels, M.J., Under Review. Typical Yet Unlikely: Using Information Theoretic Approaches to Identify Outliers which Lie Close to the Mean. DOI: TBC.

Note that, in parallel with the presentation, development, and formalisation of these methods/proposals for new approaches to analysis, the methods have been implemented in various forms as part of research presented in the applied articles below. For each article I also provide information about my own contribution. Due to the resulting length of the thesis, these articles have not been included in the thesis directly, the focus of which I have decided to keep as non-applied / theory-based / commentary-based. Nonetheless, we hope that the works below serve as evidence for the real-world relevance of the work presented at length in this thesis.

Accepted:

Biggiogera, J., Boateng, G., Hilpert, P., **Vowels, M.J.**, Bodenmann, G., Neysari, M., Nussbeck, F., & Kowatsch, T., 2021. BERT meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples' Conflict Interactions. *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pp.385-389. DOI: 10.1145/3461615.3485423.

Contribution: Supervision of analysis and manuscript editing.

—

Vowels, L.M., **Vowels, M.J.**, Mark, K.P., 2021. Uncovering the most important factors for predicting sexual desire using explainable machine learning. *The Journal of Sexual Medicine*, *18*(7), DOI: 10.1016/j.jsxm.2021.04.010.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

—

Hilpert, P., Brick, T.R., Flückiger,C., **Vowels, M.J.**, Ceulemans, E., Kuppens, P., Sels, L., 2020. What can be learned from couple research: Examining emotional co-regulation processes in face-to-face interactions. *Journal of Counseling Psychology*, *67*(4), DOI: 10.1037/cou0000416.

Contribution: Manuscript editing, software design and coding.

See emoTVrater - `https://github.com/matthewvowels1/emotvrater`.

—

Vowels, L.M., **Vowels, M.J.**, Mark, K.P., 2022. Is Infidelity Predictable? Using Explainable Machine Learning to Identify the Most Important Predictors of Infidelity. *The Journal of Sex Research*, *59*(2), DOI: 10.1080/00224499.2021.1967846.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

—

Vowels, L.M., **Vowels, M.J.**, Mark, K.P., 2022. Identifying the strongest self-report predictors of sexual satisfaction using machine learning. *Journal of Social and Personal Relationships*, *39*(5), DOI: 10.1177/02654075211047004.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

Under Review:

Leistner, C., Vowels, L.M., **Vowels, M.J.**, Mark, K.P., Under Review. Associations Between Daily Positive Communication and Sexual Desire and Satisfaction: A Comparison Between Traditional Analyses and Machine Learning.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

—

Hilpert, P., **Vowels, M.J.**, Sels, L., Mestdagh, M., Under Review. Emotion Dynamics Between Intimate Partners: Using Random Forests and Cross-Spectral Techniques to Predict Breakup Two Years Later.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

—

Vowels, L.M., **Vowels, M.J.**, Miller, J.G., Under Review. Identifying the causal link between obesity and mental health outcomes during the COVID-19 pandemic using causal discovery.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

Glossary of Terms/Concepts

Provided below is a list of working definitions for key concepts and terms used in this thesis.

The Big Assumption The assumption that the complexity of any particular social or psychological phenomenon can be adequately represented using a 'simple' / human-interpretable model (in the form, for example, of a Directed Acyclic Graph or set expert-specified structural equations).

Causal Discovery The task of discovering causal links between variables in a data-driven manner. The output is often (but not necessarily) a putative causal graph. See M. Vowels, N. Camgoz, and Bowden (2022) for more detail.

Causal Inference The task of estimating particular causal estimands (such as the average treatment effect). The success of the estimation (in terms of, for instance, the unbiasedness) depends on the identification of the estimand in terms of factors of the observed joint distribution. See Pearl (2012), J. Peters, Janzing, and Scholkopf (2017), and Morgan and Winship (2015) for more detail.

Cross-Validation A model fitting and evaluation process whereby the data used to estimate model parameters (where the parameters may represent coefficients in a multiple linear regression, or decision boundaries in a decision tree) are different from the data used to evaluate or test the model. See K. P. Murphy (2012) for more detail.

Data Generating Process (DGP) The underlying and usually inaccessible causal process which leads to a set of observable states in the world.

Deep Learning Deep learning is a data-driven Machine Learning method that has been applied to numerous applications including computer vision, natural language processing, and general predictive tasks. Deep learning techniques are usually based on types of artificial neural networks and are a multivariate, nonlinear, statistical machine learning method, allowing dependent variables to be related to independent variables via learned, complex, nonlinear relationships (I. Goodfellow, Bengio, and Courville, 2016). These relationships are learned via an automated computational process known as optimisation, such as backpropagation, whereby a loss function is minimised in order to calculate optimal network parameters known as weights and biases (I. Goodfellow, Bengio, and Courville, 2016; Rumelhart, Hinton, and R. J. Williams, 0323; Rumelhart, Hinton, and R. J. Williams, 1985). The weights and biases parameterise 'layers' in a neural network, and when the

number of layers is large (*e.g.*, above 50), the network may be considered to be 'deep'. See also (Artificial) Neural Network below and I. Goodfellow, Bengio, and Courville (2016) for more detail.

Directed Acyclic Graph (DAG) A mathematical object specified as a graph comprising a set of vertices / nodes / variables and directed edges between these vertices. The graph represents the factorisation of the joint distribution, and the edges may be used to represent causal directionality in a causal-DAG. The acyclicity prohibits the existence of cycles. See J. Peters, Janzing, and Scholkopf (2017) for more detail.

Double-Dipping A practice (intentional or otherwise) involving the reuse of data in such a way that inflates the apparent performance or success of the model. For example, fitting data-adaptive model to a data sample to maximise fit, and then failing to use a different sample to evaluate the model will inflate the apparent success of the model. See also overfitting, and Button (2019) and Kriegeskorte et al. (2009) for more detail.

(Information) Entropy A measure of uncertainty or surprise associated with a distribution. For example, the Bernoulli distribution with a parameter of 0.5 (as in the case of a flip from a fair coin) has maximum entropy because the outcome of the trial is maximally uncertain with respect to the space of possible realisations (heads or tails). The entropy would be 0, on the other hand, if the coin were maximally biased towards a probability of 1 for either heads or tails. See MacKay (2018) and Cover and Thomas (2006) for more detail.

(Model) Explanation The task of explaining the decision process or predictions of a model. For example, the coefficients of a multiple linear regression model provide an explanation for why the model makes a certain prediction given a certain input. In contrast, see *interpretation*.

Functional Form The functions used to describe the relationships between variables. For example, in $Y = f(\mathbf{X})$, f is the function relating the set of input variables $\mathbf{X}$ to the outcome Y. In a multiple linear regression, f is a weighted linear sum: $\beta_0 X_0 + \beta_1 X_1 ... \beta_K X_K$. In a random forest, on the other hand, f may be highly complex and non-linear. In contrast, see *structural form*.

Functional Form The functions used to describe the relationships between variables (either the nature of the true, real-world functions, or those used to model these real-world functions). For example, in $Y = f(\mathbf{X})$, f is the function relating the set of input variables $\mathbf{X}$ to the outcome Y. In a multiple linear regression, f is a weighted linear sum: $\beta_0 X_0 + \beta_1 X_1 ... \beta_K X_K$. In a random forest, on the other hand, f may be highly complex and non-linear. In contrast, see *structural form*.

Importances In relation to a set of predictor variables for a model, the importances tell us to what extent each of these predictors impacts model output. The importances thereby tell us which of the input variables is most predictive of the outcome.

Information Theory A branch of mathematics concerned with the formalisation of the notion of 'information'. See seminal work by Shannon and Weaver (1949), as well as MacKay (2018) and Cover and Thomas (2006) for more information.

(Model) Interpretation The task of using a model to interpret relationships which exist in the real-world. For example, in a multiple linear regression model for which one of the coefficients represents an identified causal effect, this coefficient can be interpreted in relation to an average causal effect (it tells us how much the effect changes with respect to a change in the variable associated with the particular coefficient). The requirements for a model to be interpretable are significant, and include correct functional and structural specification. See, in contrast, model *explanation*.

Linearity / Linear Models We use the term linearity to describe the situation whereby an outcome Y can be represented as a linear function (*i.e.*, a weighted linear sum) of either a set of raw input variables $\mathbf{X}$, or a set of variables which have themselves been transformed using some linearising function. In the latter case, for instance, the variables can be projected into a new 'space' whereby the outcome can still be represented as a weighted linear sum of these projected variables. An example, of this is the use of the quadratic or cubic functions of input variables, $\beta_1 X_1^2 + \beta_2 X_2^3$, such that Y is expressed as a weighted linear sum of these polynomial functions of the raw variables $\mathbf{X}$. We also include those models with a simple link function, and the definition therefore subsumes the class of Generalised Linear Models (which includes, for example, the logistic regression model).

Misspecification Misspecification occurs when a researcher specifies a model insufficiently correctly either in terms of its functional or its structural form. For a discussion on what is meant by '(in)sufficiency', see Chapter 8. For further information on misspecification, see Chapter 2.

MultiLayer Perceptron (MLP) See (Artificial) Neural Network below and I. Goodfellow, Bengio, and Courville (2016) for more detail.

Mutual Information Mutual information is an information theoretic measure of the degree to which information associated with one variable (or set of variables) is shared by another variable (or set of variables). It is a more general form of correlation which is a measure of statistical association that assumes linear forms of dependence. See Cover and Thomas (2006) and MacKay (2018) for further detail.

(Artificial) Neural Network (NN) Neural Networks are a multivariate statistical machine learning technique, which comprise a set of weights and biases (often millions thereof) which are optimized to achieve a particular goal via an optimization process known as gradient descent. An example of a commonly used goal is the minimization of the mean squared error in a regression task. The weights and biases parameterise a set of (often simple) functions called 'layers', which are stacked in a sequential fashion (although there exist variations on the arrangement of these functions). In the case of the classic MultiLayer Perceptron (which is a relatively small/simple neural network), there may exist between 2 and 10 layers, although deeper networks are possible (see Deep Learning above), and each layer comprises a set of neurons, each of which takes on a scalar value

and which is determined by a generalised linear function of all neurons in the previous layer. For more information, see I. Goodfellow, Bengio, and Courville (2016) for more detail.

Overfitting When a modeling technique fails to generalise to new data drawn from the same distribution as that with which it was trained/fit, in spite of apparently good performance on the training data, it may be said to be 'overfitting' the training data. This can happen, for instance, if the model is too complex and not sufficiently regularized, such that it learns not only to approximate the function relating the input and output variables, but also solves for the noise particular to the training sample. See also cross-validation and double-dipping, as well as Yarkoni and Westfall (2017) for more detail.

Outlier An outlier is a datapoint which deviates markedly from the sample to which it supposedly belongs. For a more detailed discussion, as well as alternative definitions, see Grubbs (1969) and Leys, Delacre, et al. (2019).

Random Forest A random forest is a type of data-adaptive decision tree that trains on bootstrapped sub-samples of the data to avoid overfitting. The tree can model highly non-linear relationships in the data, and therefore represents a significantly more flexible model than a linear regressor. For further details see Breiman (2001a).

SHapley Additive exPlanations (SHAP) SHAP is a unified framework for undertaking model explanation, and derives from the seminal game theoretic work of Lloyd Shapley Shapley (1953). The framework conceives of predictors as collaborating agents seeking to maximize a common goal (i.e., the regressor performance). The approach involves systematically evaluating changes in model performance in response to including or restricting the influence from different combinations of predictors. SHAP provides estimations of the per-datapoint, per-predictor impact on model output, as well as the average predictor impacts. These estimations are called 'explanations' because they explain why a particular regressor performs the way it does. The results are provided as feature importances, which describe how important the variable is for the model outcome and how much it changes the outcome.

Structural Form Structural form indicates whether or not certain variables or phenomena are able to influence one another (causally), regardless of the functional form underlying these influences. See above for Data Generating Process and, in contrast, functional form.

Structural Equation Modeling (SEM) A statistical modeling technique which represents structural/causal relationships between variables in a set of structural equations. The functions underlying the relationships between these variables is assumed to be linear (see above for the definition of linearity). See Kline (2005) for further details.

Typicality An information theoretic approach to characterising whether samples are 'normal' or not, in a general sense (*i.e.*, not in reference to specific, low order statistics like the mean, for example). A sample is typical sample if it falls within a certain margin defined by the entropy of a distribution. See Cover and Thomas (2006) for further details.

CHAPTER 1

Introduction

> *"To question the foundations of a discipline or a practice is not necessarily to deny its value, but rather to stimulate a judicious and balanced appraisal of its merits."*
>
> Ashcroft and ter Meulen (2004)

> *"In several fields of investigation, including many areas of psychological science, perpetuated and unchallenged fallacies may comprise the majority of the circulating evidence."*
>
> Ioannidis (2012)

Meta-researchers have increasingly drawn attention to the replicability crisis affecting psychology and social science (Oberauer and Lewandowsky, 2019; Botella and Duran, 2019; Aarts et al., 2015; Stevens, 2017; Marsman et al., 2017; Shrout and Rodgers, 2018; Yarkoni, 2019). In addition, the domains have come under heavy criticism for poor theory specification (Scheel et al., in press; Oberauer and Lewandowsky, 2019) to the extent that most research findings in psychology have been described as "not even wrong" (Scheel, 2022). Furthermore, the complicated nature of most psychological phenomena raises questions as to the feasibility of building realistic models which can deal with the complexity of human behaviour and social interaction, even in principle. Meehl coined the term 'crud factor' (Meehl, 1990; Orben and Lakens, 2020),

which alludes to the point that null-effects are practically non-existent in social phenomena because "everything [in social science] correlates to some extent with everything else". This complexity, in turn, has called into question the appropriateness of the fields' relatively blunt approaches to analysis (Bryan, Tipton, and Yeager, 2021; Freedman, 1985). To add to these problems, measurement in psychology is notoriously difficult, and the associated challenges are often not taken seriously. This leads to meta-research with titles such as "Measurement Schmeasurement", which is a valuable article discussing how research in psychology is often undermined by a range of problems, including "a lack of transparency, ignorance, negligence, or misrepresentation of the evidence." (Flake and Fried, 2020)

Table 1.1 summarises some of the open issues. As such, psychology and social science face a number of serious challenges: the inherent complexity of the phenomena of interest and the statistical methods employed to model them, the replication crisis, the theory crisis, and as I discuss below, functional and structural misspecification. The confluence of these challenges poses a serious threat to the validity and meaningfulness of research in these domains, and brings into question the direction that researchers in these fields should take. If this direction is to be positive with respect to improvements in the domains, then it is important that psychologists and social scientists engage with the meta-research surrounding areas of possible theoretical, analytical, and statistical improvements.

In this thesis I present four contributions, as well as two application examples, which are strongly motivated by the problems of misspecification and complexity, and provide recommendations to researchers. The proposals include the use of modern, powerful, data-adaptive techniques for function approximation (such as those tools from the domain of machine learning), as well as the use of techniques from the domain of causality (such as causal Directed Acyclic Graphs) and information theory. I demonstrate how these techniques can (a) help us to match the level of complexity of our modeling to the inherent complexity of the phenomenon under study, (b) reduce ambiguity with respect to theory specification, and increase the transparency of our assumptions and modeling choices, (c) reduce the complexity of a mathematical representation of a theory without impacting the validity of any downstream estimates, (d) improve the efficiency of data collection methodologies, (e) highlight the critical nature of causality even

Table 1.1: Non-exhaustive list of some challenges facing psychologists and social scientists.

A lack of understanding about and misuse of p-values and statistical tests	(Cassidy et al., 2019; Gigerenzer, 2018; Gigerenzer, 2004; Colquhoun, 2014; Colquhoun, 2017; Colquhoun, 2019; McShane et al., 2019)
Overly generous claims and warped interpretations	(Yarkoni, 2019; Spellman, 2015; Scheel et al., in press)
Issues relating to the testing of theory	(Scheel, 2022; Oberauer and Lewandowsky, 2019; Muthukrishna and Henrich, 2019a)
Immature theories	(Scheel et al., in press)
Misunderstandings about statistical power and low sample sizes	(Sassenberg and Ditrich, 2019; Baker et al., 2020; Correll et al., 2020)
Measurement problems	(Flake and Fried, 2020)
A lack of meta-analyses	(Schmidt and Oh, 2016)
A lack of assumptions testing	(Ernst and Albers, 2017)
Pressure to publish	(Shrout and Rodgers, 2018; DeDeo, 2020)
Double-dipping and overfitting	(Kassraian-Fard et al., 2016; Kriegeskorte et al., 2009; Mayo, 2013; Yarkoni and Westfall, 2017)
A failure to consider the consequences of aggregation and non-ergodicity	(Fisher, Medaglia, and Jeronimus, 2018; O. Peters and Werner, 2017)
Academia and research being a strategy game with unscientific incentives	(Gigerenzer, 2018; DeDeo, 2020)
A reluctance of journals to publish replications	(G. Martin and Clarke, 2017; Gernsbacher, 2019)
Issues with the peer review process	(Heesen and Bright, 2020)
Reporting errors	(Nuijten et al., 2016)
A lack of research practice standardization	(Tong, 2019)
The conflation of predictive and causal approaches and interpretations	(Grosz, Rohrer, and Thoemmes, 2020; Yarkoni and Westfall, 2017; Shmueli, 2010)
General scientific misconduct	(Stricker and Günther, 2019)

when otherwise powerful, exploratory machine learning techniques are used, (f) highlight the strange, unintuitive behaviour of datasets with more than four dimensions, and (g) undertake outlier detection in a way that is robust to this aforementioned strange behaviour.

In the remaining part of this Chapter, I start by discussing the problems of misspecification, complexity, and general modeling challenges in more detail, and provide an overview of the structure of the thesis with a summary of each of the contributions.

1.0.1 Functional and Structural Misspecification

The problems and challenges outlined above seem to exist to a greater or lesser extent in different subdomains of psychology and social science, and they are somewhat unsurprising given (a) the inherent complexity of humans as subject matter, and (b) the relatively unsophisticated approaches to modeling and theory building in the fields in general. Indeed, whilst there exist a multitude of powerful, highly adaptive, and general methods developed in the domain of, for example, engineering, psychologists are reluctant to use anything other than highly constrained, highly reductionist, linear models (Blanca, Alarcon, and Bono, 2018). This leads to a statistical modeling culture with assumptions which are "so unrealistic... [that] everybody agrees they are known to be false." (M. van der Laan, 2015)

Sometimes, it is not possible to use anything other than linear models, owing to limitations in data collection methodologies, measurement challenges, etc. For example, consider spectral analysis, which is typically included in the syllabi of most bachelors courses in engineering, and can be used to model and analyse a wide variety of phenomena (M. J. Vowels, L. M. Vowels, and N. Wood, 2021).[1] If one wishes to use this technique in psychological applications, it is necessary (amongst other things) to have time-series/longitudinal data collected at sufficiently consistent and regular intervals (M. J. Vowels, K. Mark, et al., 2018; M. J. Vowels, L. M. Vowels, and N. Wood, 2021). Unfortunately, longitudinal studies are both expensive, and subject to participant dropout. In contrast, cross-sectional data, which are likely to be high in abundance (leading to higher statistical power), may simply not be able to be used to answer the same research questions. Regardless, psychologists rarely, if ever, use this approach, and it would seem that this is more because such methods do not form part of psychologists' syllabi, than it is due to a justifiable choice (M. J. Vowels, L. M. Vowels, and N. Wood, 2021).

The result of a limited set of tools (linear models) with which to deal with wide variety of (arbitrarily complex) problems leads to two types of misspecification. The first I refer to as *functional misspecification*: Fitting a linear function to a non-linear phenomenon can lead to arbitrarily biased estimates (M. J. Vowels, 2021). Frequently, and in addition to problems with

[1] Technically, spectral analysis is a type of linear decomposition, but it enables us to look at non-linear trajectories over time (amongst other things). As such, I distinguish it from, say, linear regressions using only first order functions of the variables.

functional misspecification, researchers also misunderstand the way their analyses (linear or otherwise) interact with the Data Generating Process (DGP) which led to their observations/data. As with functional misspecification, this problem also leads to arbitrarily biased effect sizes which, even under optimal conditions (reliable measurement, sufficiently complex representation of the phenomenon, linearity, etc.), bare no relation to the target quantities the methods are intended to estimate. This I refer to as *structural misspecification* (M. J. Vowels, 2021). One example of structural misspecifcation concerns popular 3-4 variable mediation models. Such simple causal structures are very unlikely to reflect the phenomenon with a necessary degree of complexity and, again, the results will be arbitrarily biased.

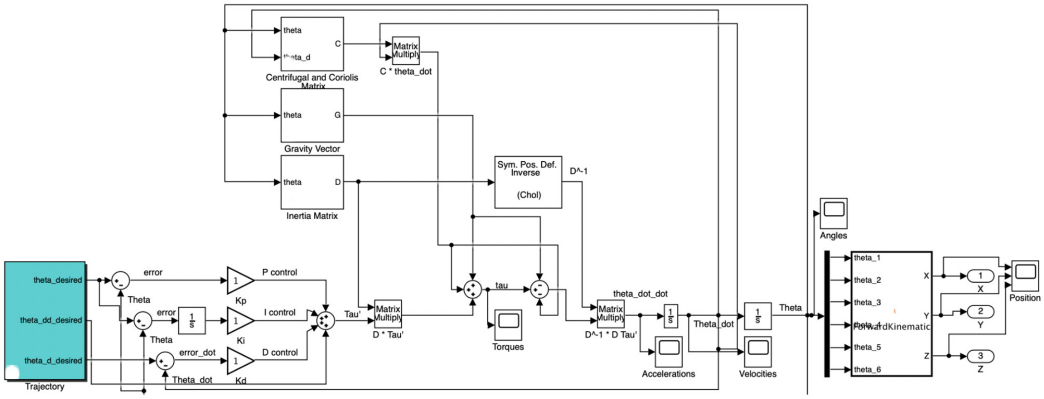
Rarely does it seem to be the case that researchers in psychology and social science take the time to really ascertain whether the research question is even approximately answerable using their chosen methodology. Arguably, if researchers *did* manage to establish an appropriate match between their theoretical, analytical, and statistical techniques and the phenomena under study, we would see a different state of affairs in the meta-research literature. Some authors have even described psychometrics as a pathology of science, on the basis that (at least significant portions of) hypotheses are accepted without serious attempts to test them, and that this problem is never questioned or rectified (Michell, 2016). In terms of the focus of this thesis, the result of functional and structural misspecification, is that replication issues resulting from (amongst other things) underpowered analyses represent only the tip of the iceberg.

1.0.2 Complexity - The Big Assumption

To help establish some perspective for the complexity issue, and why it matters for research and analysis, I designed and implemented a model for controlling a standard, workbench-mounted, six-degree-of-freedom robotic arm.[2] This model is far from innovative and is based

[2]The specific choice of this example is somewhat arbitrary. Another compelling example involves the way engineers in transducer design use both the lumped-parameter method for adequate/rough-and-ready modeling of the low-frequency behaviour of transducers, compared with the highly-parameterised finite-element approach for increased precision over a wider operational bandwidth (J. Wright, 1998; Nielsen et al., 2020). Both methods are significantly more complex than theories in psychology, but the latter is highly parameterised and not directly interpretable (it is thus analogous to the large-language model approach to generating written text. The authors practical experience (following nine years spent designing loudspeakers for commercial applications) is that the choice between the lumped parameter and the finite element methods comes down to a choice between two levels of

Figure 1.1: Simple model for controlling a robotic arm.



Note. Simple controller for a robotic arm. Corresponding expression is presented in Equation 1.1.

on rudimentary / standard recommendations in robotics textbooks - see works, for example, by Fu, R. Gonzalez, and C. Lee (2018) and Ellery (2018). The corresponding block/computational diagram is shown in Figure 1.1, and the expression corresponding with the torque of the manipulator for this model is given below (Ellery, 2018; Fu, R. Gonzalez, and C. Lee, 2018):

$$\boldsymbol{\tau} = D(\theta)\left(\ddot{\theta}^d + K_p\left(\theta^d - \theta\right) + K_i\int\left(\theta^d - \theta\right)dt + K_d\left(\dot{\theta}^d - \dot{\theta}\right)\right) + C(\theta,\dot{\theta})\dot{\theta} + G(\theta) \tag{1.1}$$

Here, the control law is represented in the terms multiplied by the coefficients K_p (proportional control gain coefficient), K_i (integral control gain coefficient) and K_d (derivative control gain coefficient). There exist what are referred to as 'disturbance terms' $D(\theta)$, which is the inertia matrix, $G(\theta)$, which is the gravity matrix, and $C(\theta,\dot{\theta})$, which is the Coriolis effect matrix. These terms are intended to model external factors which interfere with our ability to accurately position the arm. The θ terms are the joint angles which are produced by the control system, in contrast to the θ_d terms which are the desired/target joint angle terms. The dots over the top of certain terms indicate first derivative / velocity (single dot) and second derivative / acceleration (double dot).

simulation accuracy. The former allows good and fast approximation over a limited bandwidth of the audio spectrum, whilst the latter allows us to extend this useful bandwidth significantly, but at the cost of computational expense and modeling detail.

The main point here is that even a visual inspection of the block diagram (Figure 1.1) and the corresponding control law (Equation 1.1) would indicate a level of complexity which exceeds that of most psychological models (see also: Navarro, 2021). And yet, we know that the complexity of human psychology and behaviour exceeds that of a six-degree-of-freedom robotic arm. Indeed, even though linguistic (as opposed to mathematical) representations of psychological phenomena are often deep and nuanced, the contrasting simplicity of the downstream manifestation of these theories as statistical models suggests gross levels of under-specification. These problems with the mathematical representation of theories in psychology (or rather, the general lack thereof) constitute a large part of the focus of articles such as those by Scheel (2022), Scheel et al. (in press), Borsboom et al. (2021), Robinaugh, Hoekstra, et al. (2020), Haslbeck et al. (2021), van Rooij and Blokpoel (2020), Eronen and Romeijn (2020), and Navarro (2021), who argue in favour of the mathematical formalisation of psychological theory. However, in my view, the proposals for *how* to fomalise and specify otherwise verbal theories (the latter of which are sometimes referred to as 'proto-theories' in the literature) are either too non-specific to evaluate for general applicability, or have only been applied to very few and limited examples (see Robinaugh, Haslbeck, et al., 2019 for an example of a computational model designed to model panic disorder). Traditionally linguists took similar kinds of approaches to modeling language as recommended in these works - by breaking down sentences into parts of speech, parameterising grammar, phonetics, phonemics, and using basic statistical models to predict the next word in a sequence, they were able to build top-level mathematical representations of human language (Jelinek, 1997). However, the domain has by-and-large, at least for written language, abandoned these hand-engineered, expert-designed approaches in favour of those at the other end of the complexity spectrum: Highly overspecified, 'black-box', Large Language Models (LLMs). For example, the BERT LLM model (Devlin et al., 2019) comprises between 108 million and 1270 million parameters (depending on whether the small or extra-large model is used). This paradigm shift reminds us of the infamous quote "Every time we fire a phonetician/linguist, the performance of our system goes up" (circa. 1988) .[3] The sentiment of this quote seems to have aged well: LLMs provide state-of-art performance

[3]The origins of the quote are not clear, but see `https://quotepark.com/quotes/1777032-fred-jelinek-every-time-i-fire-a-linguist-the-performance-of-o/` for a non-academic discussion which references Jurafsky and J. H. Martin (2009).

by an incredible margin and produce output which is difficult to discern from that generated by a human, all without requiring hand-design by a team of expert linguists. Even the application of these modern language models as part of analyses concerning imperfect/challenging datasets indicates worthwhile increases in predictive performance margins over otherwise comprehensive, if nonetheless traditional, expert-designed methods such as Linguistic Inquiry and Word Count (LIWC) (Biggiogera et al., 2021; Body et al., 2022).

If we accept, firstly, that the theory required to achieve only approximate control of a simple robotic arm is already more complex/involved than the typical specification of most psychological theories; and secondly, that most psychological and behavioural phenomena are driven by processes which are more complex than - or even subsume those underlying - language, which requires at least hundreds of millions of parameters to model well, then the question raises its head: To what extent can psychological phenomena be represented mathematically, such that they are also useful, sufficiently complex and sufficiently accurate, in principle? Framed in a different way: If it takes 108 million parameters to model written language well (and we already have better models which have well over *175 Billion* parameters, such as GPT-3; Brown et al., 2022), how many parameters are needed to model human psychology and behaviour? Even though models like BERT and GPT-3 are *generative*, in that their principal mode of functioning involves the generation of new text (*vis-à-vis* null-hypothesis significance testing), understanding the complexity of the model required to produce human-like text provides some perspective on the scale of the problem as a whole.

Unfortunately, the focus of this thesis is not intended to answer these questions (for a related perspective, see: Freedman, 1985). Nonetheless, they are worth posing in light of the Chapters presented herein, so that the reader can bear them in mind. I pose them both in the interests of transparency - critiquing a field for its use of overly simplistic models and then proposing by way of solution even more approaches which are at least similar in their levels of tractability is not ideal - but also because I consider ways to address complexity (low/practicable levels thereof) in Chapters 5 and 7. Indeed, some of the methods I propose assume *a priori* that a realistic model of the phenomenon can be and has been specified by the researcher. Whether or not the model is correct ‘enough’ (or could ever be so, even in principle) to yield meaningful and interpretable

results is difficult to answer, and therefore the relevance of the proposed methods assumes - and this is what I informally refer to as 'The Big Assumption' - that expert-specified mathematical models of psychological phenomena can be meaningful enough and capture enough inherent complexity to be useful for inference and/or to inform experimental design.[4]

In the event that one were to end up concluding that the answer to these questions is 'no - there is no way to represent a particular complex psychological phenomenon in a mathematically tractable form', what would this mean for researchers? For a start, this conclusion is unlikely to apply to *all* psychological/behavioural phenomena, even if it might apply to a greater or lesser extent to different subdomains - for a start, we have certainly made useful progress in psychology as a whole - and also, not all questions of psychology are equally difficult to investigate (regardless of whether they are or are not equally complex in their nature). The negative response also does not preclude the integration of adequately complex (but nearly uninterpretable) data-driven models from the domain of machine learning and artificial intelligence. Indeed, the author is currently working on such projects in parallel with the work for this thesis (M. Vowels, 2020), which involve the use of existing computer vision techniques such as OpenFace (Baltrusaitis et al., 2018) and OpenPose (Z. Cao et al., 2018), as well as those which are designed by my colleagues and myself (see, for example: M. Vowels, N. Camgoz, and Bowden, 2021; M. J. Vowels, N. C. Camgoz, and Bowden, 2021). Unfortunately, such modern and highly parameterised approaches do, in general, require access to large amounts of very rich data. For instance, investigating couple conflict might, in the ideal case, involve the installation of sensors (microphones and cameras) in couples' homes, so that 24 hour surveillance is available to capture their spontaneous interactions for subsequent multi-modal (video and audio) analysis. This clearly poses ethical and privacy concerns, even if one is able to afford to run the associated data collection methodology with any appreciable number of participants. Unfortunately, the traditional/existing paradigm may not represent a viable option either, and yield meaningless results: using self-report questionnaires to create a

[4]Even when we are able to design seemingly precise interventions for experiments, it is practically challenging to disentangle the role of a particular therapist from the specifics modality being tested. Furthermore, notwithstanding the unique interactions which occur between therapist, patient, and modality, the endeavour to understand what it is precisely about the intervention/modality which has the desired effect is also difficult to disentangle. Indeed, interventions in psychology have been described as 'fat-handed', making it extremely difficult to reason effectively, and/or causally about underlying mechanisms (Eronen, 2020; Eronen and Bringmann, 2021).

retrospective time-aggregate of someone's previous conflict with their partner creates a myriad of measurement issues, and quite possibly prevents us from answering the questions we care about - in this example, questions which possibly concern the relationships between complex dynamic interactions in body pose, linguistic, para- and non-verbal behaviours.

Furthermore, the level of abstraction required to yield useful output is also to be determined as part of the modeling process. Consider, hypothetically, a research project investigating the role of non-verbal hand, arm, body and face movements in communication. It might be reasonable to suggest that there exist substantial individual differences and heterogeneity, such that the manner of gesturing for one person, and/or the way a person interprets the gestures of another, are significantly different across individuals. These individual differences might also not be explainable with demographic, cultural, or a finite set of person-specific factors (such as unique childhood experiences, etc.). Additionally, the number of possible gestures, as represented by an appropriate lexicon, might be assumed to be very high, and interact in complex ways with the verbal language one uses to communicate. If one is interested in identifying which gestures or gesture combinations impact the outcomes of conversations, it is necessary to dramatically constrain the conditions of experimentation *and* to collect an extremely large volume of data in order to control for the enormous degree of situational and contextual variation. Only then could one infer, for example, gesture combination 241 for conversation topic 13, with arousal and valence category 5, for people with cultural background type 38 and relationship type 5, has impact y in the portion $t = 4$ to $t = 25$ of an engagement. Once one begins to take averages over different topics of conversation, people of different backgrounds, etc., one begins to lose specificity for which roles one particular gesture plays in different situations. Forming some 'global' recommendation that gesture combination 241 is good to integrate into general communication on average then represents a somewhat blunt and unhelpful recommendation. Indeed, such non-specific recommendations for types of interaction might result in the opposite to their intended effect - being perceived instead as prescribed, unnatural, forced, and/or awkward.

The logic of this example may apply more or less in reality, but the underlying message is broadly relevant to psychological and social phenomena, and the challenge it poses might

more generally fall under the umbrella of heterogeneity (Bryan, Tipton, and Yeager, 2021). The difficulties associated with social phenomena and the inherent degree of heterogeneity perhaps also renders unsurprising the move away from hard-designed/coded language models, towards the use of highly over-specified models with billions of parameters. Without this level of complexity, the language models are extremely limited and fail to generate realistic text. Even though simple models are appealing for their straightforward interpretability, it is worth acknowledging the point of view that, as humans, we are biased towards explanations of phenomena that happen to be meaningful to us, even if no such simple explanations necessarily exist. Indeed, if humans are considered to only be able to handle 7 ± 2 cognitive entities at any one time (Rudin, 2019; Miller, 1956), why should any phenomenon, which happens by its very nature to involve the interaction of *more than* 7 ± 2 components, be assumed to be human interpretable by default (and modeled as if it were)? If we, as humans, are unable to explain an outcome of a highly complex but accurately predictive model in an intuitive way, does it make the model inappropriate? Once again, the willingness to abandon the anthropocentric constraint on the human-comprehensibility of our modeling techniques is partly what has lead to the dramatic increase in effective language modeling.

In combination with these issues of complexity, the running of underpowered studies is noted to be rife in psychology and social science (as alluded to at the start of this Introduction). Accordingly, if we run an underpowered study (much less than 80%, for instance) and we find a significant result, there is a high chance that this result is a false positive (*i.e.*, the conditional probability of it being a false positive *given* that we have a significant result is much higher that which we would like to expect given an alpha level of 0.05). Small samples also imply less expensive studies, and the reward (in terms of publications) possibly justifies the undertaking, at least from a career perspective. As there is no way to know what the true effect size is ahead of time, particularly if we are taking a global aggregate over a wide set of conditions for complex phenomena (such as the impact of gestures on conversational outcomes), we might be tempted to overestimate the minimum effect size of interest (Lakens, 2022) when undertaking our power analyses. This then also leads us to underestimate the number of participants required for our study. Any attempt we make to include moderators into traditional linear methods to deal with heterogeneity then dooms us to being locked into underpowered regimes (moderators often

requiring, as they do, substantially larger sample sizes to achieve the same power as the main effect of interest).

Issues of statistical power notwithstanding, more fundamentally we must still grapple with the meaningfulness of the resulting quantity we wish to estimate, and whether it represents too much of an aggregate over too many conditions of variation. Even if confronting the inherent complexity of a phenomenon has the potential to create a research impasse, I would instead argue that it encourages us to adopt an appropriate level of skepticism and take extra care when drawing conclusions from models which are likely to be significantly underspecified. It goes without saying that in the scientific endeavour we ought to be careful, and to always 'do our best' to design realistic and helpful models of reality. However, research in psychology and social science in particular concerns the well-being of humans, and from an ethical standpoint, a heavy responsibility falls on the researchers to avoid the proliferation of damaging, misleading, or false information that might unavoidably derive from underpowered studies which use unreasonably basic models to represent almost intractably complex phenomena. Unfortunately, the engagement of researchers with good research practices does not regularly align with the incentives (such as the pressure to publish) built into the academic machine (DeDeo, 2020; van Dalen, 2021).

1.0.3 Proposed Solutions and Thesis Structure

In this Chapter I outlined three principal challenges faced by psychologists and social scientists: Functional misspecification, structural misspecification, and complexity. These three problems overlap somewhat, and together form a strong motivation to identify a way forward.

Chapter 2 - Misspecification

One remedy for the structural misspecification issue involves the engagement of psychologists with causal methods, such as the causal Directed Acyclic Graph framework (Pearl, 2012) or the potential outcomes framework (D. B. Rubin, 2005). The relative scarcity of such methods in psychology and social science is quite an oversight, and has interesting putative historical

causes (Pearl, 2012). As Reynolds (2021) explain: “Nonetheless, these and other undergraduate texts give students very little information about our modern understanding of causality. These ‘traditional’ views can likely be traced to three giants in the field of statistics: Sir Francis Galton, Karl Pearson, and Sir Ronald Fisher, with Pearson potentially having the greatest impact.” Regardless of the historical origins of the situation, the failure of researchers to take these approaches seriously has led to untestable theories (Scheel, 2022), a conflation of causality and correlation, a lack of assumptions testing, and a lack of transparency (Grosz, Rohrer, and Thoemmes, 2020; Rohrer, 2018).

On the other hand, a remedy for functional misspecification involves the engagement of psychologists with machine learning and both non- and semi-parametric statistical methods. These approaches can allow psychologists to avoid imposing unreasonable constraints to fit unknown functional relationships between variables, without sacrificing statistical inference (via, for example, the null-hypothesis significance testing framework). Unfortunately, at least some of these techniques (both causal and machine learning) are non-trivial to understand, implement, and adopt, particularly for researchers in a field which, in general, is known for its limited technical background/training (Boker and Wenger, 2007) and patchy statistical education (Cassidy et al., 2019). This is, of course, a great shame, psychology representing, as it does, the study of something which is of the upmost importance to us - ourselves. It is also something which deserves the application of methods with a level of flexibility and complexity which can match the level of complexity of the phenomenon they are intended to model. Humans are not simple by design, and arguably deserve better representation than with straight lines and structurally reductionist models.

In Chapter 2, I present the following work:

Vowels, M.J., 2021. Misspecification and Unreliable Interpretations in Psychology and Social Science. *Psychological Methods*. DOI: 10.1037/met0000429.

In the Chapter, I demonstrate the nature of misspecification problems and discuss how they manifest in typical psychology and social science research. I argue that most models used in psychology and social science are limited in their functional form and misspecified in terms of causal structure. The result is that subsequent interpretations conflate predictive and causal

language and are also unreliable. I make four recommendations for researchers in these fields to update and improve their research practice by (1) giving more consideration to the use of flexible and varied predictive modeling and model explainability techniques, such as those from the domains of machine learning and information theory; (2) to seek collaboration with experts from the fields of statistics and machine learning; (3) to be transparent about whether they are adopting a predictive or causal approach; and (4), to distill and simplify their research questions and hypotheses in order to increase the chances that these questions and hypotheses can be practically addressed and tested.

Chapters 3 and 4 - Applications of Machine Learning and Causal Analytical Methods

These two chapters provide application examples of two important recommendations made in Chapter 2. Firstly, that researchers engage with machine learning techniques: Chapter 3 provides an example of machine learning and machine learning explainability techniques applied to predict perceived partner support from relational and individual variables. It is based on the following publication:

Vowels, L.M., **Vowels, M.J.**, Carnelley, K.B., Kumashiro, M., 2022. A machine learning approach to predicting perceived partner support from relational and individual variables. *Social Psychological and Personality Science*, DOI: 10.1177/19485506221114982.

Then, in Chapter 4, I provide a second application example for causal discovery, machine learning, and causal inference (specifically targeted learning) all together for the task of identifying causal links between attachment styles and mental health, for data collected during the COVID-19 pandemic:

Vowels, L.M., **Vowels, M.J.**, Carnelley, K.B., Millings, A., Miller, J.G., Under Review. Toward a Causal Link between Attachment Styles and Mental Health during the COVID-19 Pandemic.

As such, these two chapters provide evidence that the recommendations I made can be fruitfully applied in practice.

Chapter 5 - Model Complexity

Further considering the ideas and recommendations presented in Chapter 2 regarding misspecification, I develop the relevance of a consideration for structure to see how we can reduce the complexity of statistical models without biasing the effect sizes we wish to estimate. Indeed, given the discussion above about 'The Big Assumption', any reduction in complexity is valuable in making the resulting statistical estimation problem tractable. As such, in Chapter 5, I present the following work, which presents techniques for reducing the complexity of a structural model:

Vowels, M.J., 2023. Prespecification of Structure for the Optimization of Data Collection and Analysis. *Collabra: Psychology*.

In the Chapter, I argue that graphical representations of our theories provide us with an opportunity to encode our domain knowledge about a particular phenomenon of interest, and make our assumptions more explicit. I introduce unfamiliar readers to the rules of Directed Acyclic Graphs, and explain how to use these rules to understand the consequent statistical structure in the data. Furthermore, I show that, by using these rules (in particular, the concept of conditional independencies), we can significantly shrink the required causal structural model without affecting the validity of the associated estimates, thereby reducing the required sample size and enabling us to redirect resources and funds towards the collection of variables which are critical to answering the questions we care about.

Chapter 6 - Outrunning Causality

Whilst I make recommendations for the use of machine learning in Chapter 2, in Chapter 6 I also demonstrate how these machine learning models are far from immune to the structural misspecification issue, and that the structural and functional considerations are tied together. With the increased uptake and application of new methods from the domain of machine learning, it is not uncommon to also see such models being misunderstood and misused. As such, in Chapter 6, I demonstrate that even if researchers wish to use machine learning to explore their

data, predictive methods strongly interact with the underlying structure in such a way that the exploration can nonetheless yield misleading results:

Vowels, M.J., Under Review. Trying to Outrun Causality with Machine Learning: Limitations of Model Explainability Techniques for Exploratory Research.

In the Chapter, I question the utility of measures of predictive importance and explainability techniques to psychologists wishing to explore the data to guide their research. Indeed, how useful is it for the development of a theory to know that variable X is useful for predicting variable Y in arbitrary algorithm f, if the estimation of usefulness is specifically tied to the algorithm and the choice of the other predictors? I conclude that one cannot 'outrun causality in machine learning', and that despite of the powerful function approximation capabilities of machine learning algorithms, they cannot be used to reliably explore the data even for relevant predictive variables, let alone causal variables.

Chapter 7 - The Typical Human

Above, I discussed what I refer to as The Big Assumption - whether a psychological phenomenon can, in principal, be represented by a tractable, researcher-specified mathematical model (such that the model is also useful and sufficiently accurate). On a less philosophical level, I try to understand how traditional statistical methods behave as we start to accommodate the complexity of psychological phenomena, and demonstrate that even datasets with 4 Gaussian variables start exhibiting unintuitive behaviours which we should be aware of. Chapter 7 presents the following work:

Vowels, M.J., Under Review, Typical Yet Unlikely: Using Information Theoretic Approaches to Identify Outliers which Lie Close to the Mean.

Whilst this work differs from the others in that it does not directly consider issues of structure or functional form, it takes a complementary perspective in terms of the implications of high-dimensionality and complexity. I discuss how various manifestations of the arithmetic mean (which, as I discuss, may itself represent an overly simplistic model) have been used both productively and unproductively as a blunt way to characterize samples and populations.

Through an exploration of multi-dimensional space, I show that the mean, far from representing normality, actually represents abnormality, in so far as encountering a datapoint close to the mean in datasets comprising more than a handful of dimensions becomes incredibly unlikely, even with a large number of datapoints.

The approach I propose to ameliorate the associated problems are also inspired by information theory, which is a domain I recommended researchers to engage in in Chapter 2. In contrast with the arithmetic average, the information theoretic quantity known as 'typicality' provides a way to establish normality (or rather, whether a datapoint is typical or atypical), which is particularly useful in high-dimensional regimes. Given that researchers in psychology and social science frequently deal with multivariate datasets, and that the peculiarities associated with multi-dimensional spaces start occurring in relatively low dimensions (as few as four), it is important that researchers have some awareness of the concepts presented in this paper. To conclude, I finish with a demonstration for how the typical measure can be adapted to outlier detection, and provide an evaluation to verify its performance in comparison with a popular alternative.

Chapter 8 - Conclusion

In Chapter 8 we provide a discussion of possible avenues for future work, discuss some of the limitations of the proposals made, particularly in relation to The Big Assumption, provide a discussion about how the field can adapt and make positive changes, and finish with a summary conclusion.

1.0.4 Summary

This thesis provides an exploration of various important problems and challenges facing researchers in psychology and social science. In the last three years I have had the opportunity to apply some of the proposals made in this thesis to 'real-world' psychological applications (in addition to the two included in Chapters 3 and 4) - in the Declaration Section, I provide a list of such additional works (five accepted for publication, three under review at the time of writing).

The statistical, causal, and machine learning approaches I discuss in this work have been adapted to a wide variety of problems relating to obesity, COVID, and mental health, partner support, sexual desire, and others. These projects help motivate and justify the real-world applicability of the proposals made herein.

Finally, whilst the four technical works presented in this thesis are not intended to provide finite solutions to the problems described above (particularly in light of the deeper question concerning The Big Assumption), they are strongly motivated by them. Furthermore, in spite of this introduction's otherwise gloomy tone, I am positive that if researchers acknowledge the problems, research methodology and analysis in psychology and social science can begin to move in a positive direction. Indeed, the nature of at least some of the problems (for example, the limited use of non-linear and causal models) would seem to encourage an optimistic interpretation of the situation: That there presently exists a tremendous opportunity to innovate and modernise the current approach to research, simply by assimilating recent advances and developments in other domains such as engineering, machine learning, and statistics.

CHAPTER 2

Misspecification and Unreliable Interpretations in Psychology and Social Science

> *[The] lack of truth in current practice, supported by statements such as "All models are wrong but some are useful," allows a user to make arbitrary choices even though these choices result in different answers to the same estimation problem. In fact, this lack of truth in current practice presents a fundamental drive behind the epidemic of false positives and lack of power to detect true positives our field is suffering from."*
>
> M. J. van der Laan and Starmans (2014)

Notwithstanding minor edits, this chapter is equivalent to the following publication:

Vowels, M.J., 2021. Misspecification and Unreliable Interpretations in Psychology and Social Science. *Psychological Methods*. DOI: 10.1037/met0000429.

Abstract: Numerous causes have been attributed to the replication crisis in psychology and the social sciences, many of which concern problematic analytic and statistical practices. In this work we focus on three issues that we believe deserve more attention. Namely, the use of models with limited functional form, the use of misspecified causal models (misspecified

either due to limited functional form, or incorrect structure), and unreliable interpretations of results. We demonstrate a number of consequences relating to these issues via simulation, and provide recommendations for researchers to improve their research practice. While the issues raised in this work have been identified previously, we believe it is extremely important to encourage psychologists and social scientists to engage with the debate surround areas of possible analytical and statistical improvements.

2.1 Introduction

Meta-researchers have increasingly drawn attention to the replicability crisis affecting psychology and social science (Oberauer and Lewandowsky, 2019; Botella and Duran, 2019; Aarts et al., 2015; Stevens, 2017; Marsman et al., 2017; Shrout and Rodgers, 2018; Yarkoni, 2019). A key element of the crisis relates to common and fundamentally problematic analytic and statistical practices, some of which we believe deserve more attention. In our view, these problematic practices have the potential to seriously affect the reliability and interpretation of research and therefore to hinder scientific progress.

These problematic practices relate to observational research and modeling in psychology and social science, and may be broadly categorized as issues with (1) the use of statistical/predictive models with limited functional form; (2) the misspecification of causal models; and (3) unreliable and interpretations of predictive or causal models. All of these issues affect a researcher's ability to accurately model some aspect of the joint distribution of the data, for the purpose of predicting an outcome, estimating a causal effect, and drawing scientific conclusions. The first issue relates to the ubiquitous use of linear models, and a failure to consider more powerful, possibly data-adaptive techniques for both predictive and causal modeling. The second relates to the use of misspecified implicit (e.g. multiple linear regression) or explicit (e.g., structural equation) causal models which do not sufficiently reflect the true structure in the data. The final issue relates both to how predictive models are often (mis)interpreted as causal models, and vice versa, and also to how these interpretations are likely to be unreliable given the models' underlying limitations and assumptions.

We address the three issues in turn through both didactic illustration and simulation, and make a number of recommendations for improving research practice. While these issues relating to research practice have been previously discussed, we believe it is extremely important to continue to encourage and stimulate consideration and engagement with the debate surrounding areas of possible analytical improvement. Furthermore, in spite of researchers having already made important recommendations for improving practice (e.g., Lakens, Hilgard, and Staaks, 2016; Scheel et al., in press; Gigerenzer, 2018; Jostmann, Lakens, and Schubert, 2016; Lakens and Evers, 2014; Orben and Lakens, 2020) we see relatively little change in the research communities of psychology and social science (Claesen et al., 2019; Scheel et al., in press).

For convenience, we have included some key definitions of relevant terminology, which is followed by a review of the literature. The paper is then split into four main parts. In Part 1, we describe how the typical models used in psychology are limited by their functional form and discuss the implications of this issue and possible ways to address it. Part 2 is concerned with misspecification in causal modeling, and how the typical models used in psychology and social science do not adequately reflect the true structure of the data. We discuss how this impacts interpretability, how a consideration for causal structure is essential when designing a model, and identify some challenges associated with undertaking causal modeling. Part 3 introduces the notion of explainability as an alternative to interpretation, as a means of deriving insight from predictive models. We discuss interpretation, considering the relevant points on limited functional form and misspecification covered in Parts 1 and 2, and discuss how interpretations in psychology and social science tend to be a conflation of causal and predictive interpretations. Finally, Part 4 brings together the principal points discussed in previous parts, and sets out four recommendations for improving practice.

2.1.1 Definitions/Explanations

In this section we define and explain, for the purposes of this paper, six terms: ‘approach’, ‘model’, ‘predictive’, ‘causal’, ‘functional form’, and ‘misspecification’, and summarize these definitions in Table 2.1. The term ‘approach’ relates to the broad intention of the researcher when investigating a phenomenon of interest, and informs research methodology, data collection

procedure, analysis (including the model), interpretation, etc. In this paper we consider both predictive and causal approaches. The term 'model' relates to the mathematical relationship between variables associated with a phenomenon (i.e., variables in the joint distribution), *as reflected in the algorithm or technique used for analysis*. The type of *model* may be predictive or causal, or a hybrid of the two, although the type of model will generally be strongly informed by the approach.

'Predictive' approaches have been described as "the study of the association between variables or the identification of the variables which contribute to the prediction of another variable" (Blanca, Alarcon, and Bono, 2018). Prediction may help us to answer questions such as 'when?', 'which?', and 'how much?'. In contrast to a causal approach (defined below), prediction enables us to identify an association between two or more variables and to thereby estimate or classify an outcome or category, but it won't necessarily tell us *what if* the predicted phenomenon does or does not occur, *why* the predicted phenomenon may occur (or not), or *how* to intervene. As such, prediction is unlikely to generate understanding as it does not directly inform us about the underlying causal mechanisms. Prediction involves the specification, fitting, or learning of a function to enable one to forecast or predict outcomes for new datapoints.

Table 2.1: Basic working definitions.

Approach	Relating closely to the hypothesis/research question, it describes the broad intention behind research methodology, analysis, and interpretation.
Model	Part of the approach, it is the mathematical relationship between variables, *as reflected in the algorithm or technique used for analysis*. It may be predictive or causal, or a hybrid.
Predictive	The "study of the association between variables or the identification of the variables which contribute to the prediction of another variable" (Blanca, Alarcon, and Bono, 2018). The word "association" here alludes to the fact that the associations or relationships between variables are not necessarily causal. As such, prediction may help us to answer questions such as 'when?', 'which?', and 'how much?'.
Causal	The study of cause-effect relationships between variables, which facilitates understanding and answers questions such as such as 'why?', 'how?', and 'what if?' (Pearl, 2009)
Functional Form	The nature of the mathematical function describing the relationship between variables.
Misspecification	When a model does not sufficiently reflect the causal structure of the data, or is not flexible enough to estimate the underlying functions relating the variables, it is structurally and/or functionally misspecified, respectively.

'Causal' approaches help us to answer causal questions such as 'why?', 'how?', and 'what if?' (Pearl, 2009). Causal questions may be answered using causal modeling techniques (such as structural equation modeling), for observational as well as randomized and experimental data (Pearl, 2009). Causal modeling techniques generally entail a specification of what is known as the *data generating process*. On the basis that one of the principal aims of psychology and social science, as well as science more generally, is to *develop understanding* (Gelman, 2014; Pearl, 2009; M. J. van der Laan and S. Rose, 2011; Grosz, Rohrer, and Thoemmes, 2020), the causal approach provides the means for researchers to achieve this aim.[1]

It is possible to combine considerations for causal structure deriving from domain knowledge into predictive models, and thereby construct a hybrid model, without making a predictive approach a causal approach. This additional domain specific information is also known as inductive bias (K. P. Murphy, 2012). For instance, a language model might be designed to account for the ordering of words in a sentence, in addition to the words themselves, on the basis that we know *a priori* that this ordering can affect the meaning (Rabiner and Schafer, 1978). However, it is important to note that such a model would still be predictive in spite of the integration of such structural inductive bias. Indeed, given the complexity of language it would be practically impossible to pre-specify a full and 'correct' causal graph. Nonetheless, the more that predictive models incorporate domain knowledge or causal inductive bias, the more chance they have of reflecting the real-world and subsequently of being interpreted to yield causal understanding. However, unless the causal effect(s) of interest are *identifiable* (see Part 2), the model will fulfil a predictive role more than a causal role. This is because structural misspecification (i.e., a model structure that does not account for all real-world causal relationships) is not problematic for prediction in the same way as it is problematic for causal inference. As such, unless a specific causal effect is *identifiable*, and the model is designed to yield such causal information, we would classify hybrid models as forming part of a predictive, rather than causal, approach.

We use the term 'functional form' to describe the mathematical relationship between variables

[1] Both predictive and causal models may be parametric, semi-parametric, or non-parametric, Bayesian, or frequentist, and may or may not incorporate significance testing (Shmueli, 2010; Bishop, 2006; K. P. Murphy, 2012; M. J. van der Laan and S. Rose, 2011; J. Peters, Janzing, and Scholkopf, 2017; Pearl, 2009).

in a model. For instance, a linear regression has a linear functional form, whereas a neural network (I. Goodfellow, Bengio, and Courville, 2016) has a highly flexible, non-linear, data-adaptive functional form. For any data generating process, there may exist an optimal functional form with which to model it, and identifying this functional form is one of the goals of predictive modeling. If the functional form of the model is insufficient, then the model is biased. Conversely, if the functional form of the model is overly flexible, care must be taken to avoid excessively high variance and to avoid 'overfitting'. Finally, the term *misspecification* describes the scenario in which the true causal structure, and/or the functional form of the relationships between variables in the data generating process, are not sufficiently reflected in the model. The issue of misspecification due to incorrect structure may therefore be compounded by issues of limited functional form.

2.1.2 Background

Over the last ten years, meta-researchers have drawn increasing attention to a purported crisis in the human sciences (particularly psychology) known as the replication crisis. The crisis has been discussed at length by many different meta-researchers (e.g., Oberauer and Lewandowsky, 2019; Botella and Duran, 2019; Aarts et al., 2015; Stevens, 2017; Marsman et al., 2017; Shrout and Rodgers, 2018; Yarkoni, 2019) who argue that research in the human sciences fails to replicate. For example, only six out of 53 landmark cancer studies were found to replicate (Begley and Ellis, 2012), and between one third and one half of 100 psychology studies in top-ranking journals could be replicated (Aarts et al., 2015; Marsman et al., 2017).

One of the positive outcomes of the widespread awareness of the replicability crisis is the fact that attention has been drawn to many questionable, suboptimal, or problematic aspects associated with the research procedure in general. Indeed, it is only by recognition of these issues, and engagement in relevant constructive debate, that research practice can be improved. A wide range of contributing factors to this crisis have been highlighted and discussed, and include: A lack of understanding about and misuse of p-values and statistical tests (Cassidy et al., 2019; Gigerenzer, 2018; Gigerenzer, 2004; Colquhoun, 2014; Colquhoun, 2017; Colquhoun, 2019; McShane et al., 2019); overly generous claims and warped interpretations (Yarkoni, 2019;

Spellman, 2015; Scheel et al., in press); issues relating to the testing of theory (Oberauer and Lewandowsky, 2019; Muthukrishna and Henrich, 2019a); immature theories (Scheel et al., in press); misunderstandings about statistical power and low sample sizes (Sassenberg and Ditrich, 2019; Baker et al., 2020; Correll et al., 2020); measurement problems (Flake and Fried, 2020); a lack of meta-analyses (Schmidt and Oh, 2016); a lack of assumptions testing (Ernst and Albers, 2017); pressure to publish (Shrout and Rodgers, 2018); double-dipping and overfitting (Kassraian-Fard et al., 2016; Kriegeskorte et al., 2009; Mayo, 2013; Yarkoni and Westfall, 2017); a failure to consider the consequences of aggregation and non-ergodicity (Fisher, Medaglia, and Jeronimus, 2018; O. Peters and Werner, 2017); academia and research being a strategy game with unscientific incentives (Gigerenzer, 2018; DeDeo, 2020); a reluctance of journals to publish replications (G. Martin and Clarke, 2017; Gernsbacher, 2019); issues with the peer review process (Heesen and Bright, 2020); reporting errors (Nuijten et al., 2016); a lack of research practice standardization (Tong, 2019); the conflation of predictive and causal approaches and interpretations (Grosz, Rohrer, and Thoemmes, 2020; Yarkoni and Westfall, 2017; Shmueli, 2010); and general scientific misconduct (Stricker and Günther, 2019).

More specifically, meta-researchers have highlighted how psychologists and social scientists tend to mix causal and predictive language (Grosz, Rohrer, and Thoemmes, 2020). For instance, Grosz, Rohrer, and Thoemmes (2020) explain how "some parts of the articles read as if the entire endeavor were noncausal; yet other parts make sense only in the context of trying to answer a causal research question". A typical example of this can be found in recent work looking at the associations between residential green space and child behavior and intelligence (Bijnens et al., 2020). In a summary bullet point, the researchers stated that their results "indicate that residential green space is especially beneficial for intellectual and behavioral development", which is a causal interpretation denoting that the green space itself affects development. This was immediately followed by a consecutive bullet point, which stated that "low residential green space in urban children is *associated* with a "shift" towards a higher incidence of low IQ...",[2] which is a predictive, or associational interpretation. This conflation of causal and predictive terminology is confusing and misleading to readers because it either suggests that the research was causal (when it wasn't), or that, regardless of the type of approach, the results have

[2]Italics our own.

causal implications (which they don't necessarily). Indeed, in a popular daily newspaper review of the article, the headline reads: "Growing up near green space makes city children more intelligent and better-behaved" (Rudgard, 2020), which is unambiguously causal and clearly has the potential to greatly concern parents not living in areas near green space.[3] Ambiguous claims and pseudo-causal interpretations therefore have the potential to be amplified by the media, who apprise the public of scientific progress, resulting in misunderstandings and confusion.

Similarly, meta-researchers have drawn attention to how it is common for psychology and social science researchers to use associational/predictive techniques to test otherwise causal hypotheses (Shmueli, 2010; Yarkoni and Westfall, 2017; C. Glymour, 1998; M. Hernan, 2018a; Grosz, Rohrer, and Thoemmes, 2020). Shmueli (2010) explains how "the type of statistical models used for testing causal hypotheses in the social sciences are almost always association-based [i.e., predictive] models." One can only surmise the possible causes behind this tendency for conflation, but it may relate to the controversial history of causal inference in observational social science and psychology. The conflation may stem from the conflict between recognizing the importance of asking causal questions, without wanting to be seen to be actually using causal methods with observational data. Indeed, the literature on causality in psychology and social science has been described as "one of the oddest literatures in all of academia" (Dowd, 2011), and researchers in these fields are notoriously reluctant to adopt appropriate modeling techniques (Grosz, Rohrer, and Thoemmes, 2020; M. Hernan, 2018a). Others have mocked the reluctance to undertake causal inference in psychology and the social sciences by referring to causality as "the C-word" (M. Hernan, 2018a; M. Hernan, 2018b), and others refer to its use as "taboo" (Grosz, Rohrer, and Thoemmes, 2020). Indeed, Grosz, Rohrer, and Thoemmes (2020) explain how causal modeling is only undertaken "implicitly, opaquely, and without an articulation of the underlying assumptions". The result has been a tendency to use predictive language such as 'associations', 'links, 'correlations', 'relationships', and to avoid causal language such as 'causes', 'impacts', 'effects' despite designing their models and experiments on the basis of deeply considered theories about the causal structure of the phenomenon of interest (Shmueli, 2010).

[3] For other examples see Grosz, Rohrer, and Thoemmes (2020), and for additional discussion see Shmueli (2010) and Yarkoni and Westfall (2017).

In addition to a general reluctance to adopt clearly articulated causal approaches, one might also argue that the various manifestations of conflation indicate a lack of understanding about the differences between predictive and causal modeling (Yarkoni and Westfall, 2017; Shmueli, 2010; Grosz, Rohrer, and Thoemmes, 2020). Indeed, in the example above concerning residential green space, the conflation of predictive and causal language is more likely to be due to a possible lack of understanding about the distinction and limitations of predictive and causal approaches, rather than a taboo around causality. After all, it seems that such a taboo would result in an absence of causal language altogether, rather than a conflation. There is a relatively well established modeling technique known as Structural Equation Modeling (SEM) (Kline, 2005; Blanca, Alarcon, and Bono, 2018). The point to note about the use of SEM in psychology and social science is that, while SEM is a technique which explicitly encodes causal structure, the way the technique is often presented and interpreted, obfuscates its causal nature (Grosz, Rohrer, and Thoemmes, 2020). This leads to an awkward conflation of causal modeling with predictive interpretations, resulting in ambiguity and a lack of clarity regarding intentions and assumptions. It may be that researchers are unaware that their SEMs are explicitly causal and fail to sufficiently understand how the results from the analysis are underpinned by a number of restrictive (and often untestable) assumptions.

There is also evidence of a possible lack of understanding relating to the use of predictive models in psychology and social science. Yarkoni and Westfall (2017) provide a number of examples of where researchers seem to have clearly identified that they are adopting a predictive approach but use suboptimal and misguided predictive modeling practice. A wide range of powerful predictive modeling techniques exist, including neural networks (I. Goodfellow, Bengio, and Courville, 2016), random forests (Breiman, 2001a), gradient boosting machines (T. Chen and Guestrin, 2016) etc., many of which derive from developments in machine learning. In spite of the abundance of available options, researchers in psychology and social science most often employ simple linear models when undertaking predictive /associational research (Yarkoni and Westfall, 2017; Blanca, Alarcon, and Bono, 2018). The assumption of linear functional form is often restrictive and has been previously noted to be problematic (M. J. van der Laan and S. Rose, 2011; Asuero, Sayago, and A. Gonzalez, 2006; Onwuegbuzie and Daniel, 1999; Achen, 1977; King, 1986; Meehl, 1990; Taleb, 2019) and frequently ignored (Ernst and Albers,

2017). Furthermore, some researchers seem to be unaware of certain basic principles relating to predictive (as well as causal) research, such as those relating to overfitting (Yarkoni and Westfall, 2017; Bishop, 2006; Heyman and Slep, 2001) and 'double-dipping' (Kassraian-Fard et al., 2016; Kriegeskorte et al., 2009; Mayo, 2013). Overfitting and double-dipping refer to modeling (mis)practices which increase the fit of a model to the specific data sample being used, and which negatively impact the validity and generalizability of results. Indeed, *any* modeling decision that affects the parameters of the model based on information from the same data sample with which the model is validated results in overfitting, biased effect sizes, and the inflation of p-values and other performance metrics (Bishop, 2006; Yarkoni and Westfall, 2017; Heyman and Slep, 2001). Regardless of whether a researcher is undertaking a predictive or causal approach, overfitting inflates the apparent success of the mapping function at the expense of generalizability to new samples, and has been argued to be a major contributor to the current replicability crisis (Shrout and Rodgers, 2018; Gelman and Loken, 2013).

Given the prior commentary, it can be seen that we are not the first to draw attention to problematic analyses and a potential lack of analytical understanding in the fields of psychology and social science (Claesen et al., 2019; Scheel et al., in press). Indeed, a recent article titled 'Declines in religiosity predict increases in violent crime - but not among countries with relatively high average IQ' was retracted from the Journal of Psychological Science on the basis of methodological weaknesses and political sensitivity. The Editor in Chief at the time, Steve Lindsay apologized on multiple grounds, and stated that "In terms of science, Clark et al. may not be worse than some other articles published in Psych Science during my editorship..." (Lindsay, 2020). This may suggest that methodological weakness, as described in terms of "blurred distinctions between psychological constructs versus measures and speculations/extrapolations far removed from the data" is somewhat par for the course in the "young science" (Lindsay, 2020) of psychology.

2.2 Part 1: Limited Functional Form - Modeling Relationships Between Variables

In this part we address certain issues that may arise when using modeling techniques that have limited functional form. The term *functional form* relates to the mathematical form used to represent the relationship between variables. When we refer to the functional form of a model as being limited we mean that the model does not have the flexibility to sufficiently reflect the complexity of the relationship between variables, possibly resulting in poor predictive ability and biased results. Identifying or deriving an adequately flexible functional form with which to model the relationship between variables, in circumstances where causal relationships are not of concern, is somewhat synonymous with the task of prediction. As such, the majority of this section will be written with consideration of its relevance to predictive modeling, where the goal is to learn a function that optimally maps predictor variables to outcome variables. However, a consideration for functional form is just as important for causal modeling, for which we are tasked with modeling *both* the functional relationships between variables *as well as* the causal structure of the data generating process. For purposes of prediction alone, it suffices to be solely concerned with finding the optimal mapping function to achieve some desired level of predictive performance. We expect models that reflect the structure of reality to also be good predictors, but this is not necessarily the case the other way around; good predictive functions do not necessarily reflect the structure of reality.

We begin by introducing some of the technical formalism behind predictive modeling, and briefly list some of its wide ranging applications. Following this, we discuss the limitations of undertaking prediction using the two most common and basic methods used in psychology and social science: Correlation and linear regression models. We demonstrate how these methods, in the basic form adopted in psychology and social science, are fundamentally limited in their ability to account for non-linearities present in the data. This motivates a need for more flexible, powerful, potentially data-adaptive predictive methods. Previous research has highlighted that the use of such techniques is rare in psychology and social science, where it is much more usual to use models with restrictive linear functional form (Yarkoni and Westfall, 2017; Blanca,

Alarcon, and Bono, 2018). Linear functions may be useful to consider for their computational efficiency and for their tendency to naturally *under-fit* the data, thereby improving generalization particularly when the quantity of data is limited. However, these factors are not sufficient to fully explain the rarity of non-linear, powerful, and/or data adaptive techniques in psychology and social science, and we posit that a possible lack of awareness of these alternative methods is more likely.

2.2.1 Applications and Basic Formalism

The topic of identifying the optimal functional form with which to represent the relationship between variables is vast and well covered by many authors, particularly those in the field of machine learning in the context of prediction (Bishop, 2006; Duda, Hart, and Stork, 2001; K. P. Murphy, 2012). Prediction has been described as "the study of the association between variables or the identification of the variables which contribute to the prediction of another variable" (Blanca, Alarcon, and Bono, 2018) and therefore relates closely to the more general task of identifying the optimal function that maps between sets of variables. The applications for predictive models are wide ranging, and include personalized medicine (Rahbar et al., 2020), data science competitions (Tauchert, Buxmann, and Lambinus, 2020), time series forecasting (Makridakis, Spiliotis, and Assimakopoulos, 2020), facial and object recognition (Krizhevsky, Sutskever, and Hinton, 2012; Jonsson et al., 2000), and many others. Such techniques are therefore extremely valuable and influential in shaping our modern world.

The basic formalism for predictive modeling is as follows: Researchers may be confronted with a dataset comprising samples from a population $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y}$. [4] In words, we have a set of samples of predictors or random variables[5], which take on values in the set $\mathcal{X}$ and which are related to some outcome variables[6] which take on values in the set $\mathcal{Y}$. If the outcome is

[4] We adopt the following notation: upper-case bold symbols (e.g, $\mathbf{X}$) indicate matrices, lower-case bold symbols (e.g, $\mathbf{x}$) represent vectors, and lower-case symbols (e.g, x) indicate scalars. In general, we will use vector or matrix notation, rather than scalars, to increase generality. Subscripts $\{i, k\}$ (e.g, $\mathbf{x}_{ik}$) indicate datapoint $i = \{0, 1, ...(N-1)\}$ for variable or feature $k = \{0, 1, ...(K-1)\}$, where N is the number of datapoints (i.e., sample size), and K is the total number of variables or features.

[5] These are sometimes called 'independent variables', but due to the fact that they are usually non-independent, we avoid this potentially unhelpful terminology.

[6] These are sometimes called 'dependent variables', but due to the fact that many dependencies exist we also avoid this terminology.

binary or categorical, the task of prediction becomes equivalent to one of classification. The goal of prediction usually involves finding a mapping function $f : \mathcal{X} \rightarrow \mathcal{Y}$. We will use the terms *predictive function* and *predictive model* to refer to the mapping function used to make predictions.

2.2.2 The Common Assumption of Linear Functional Form

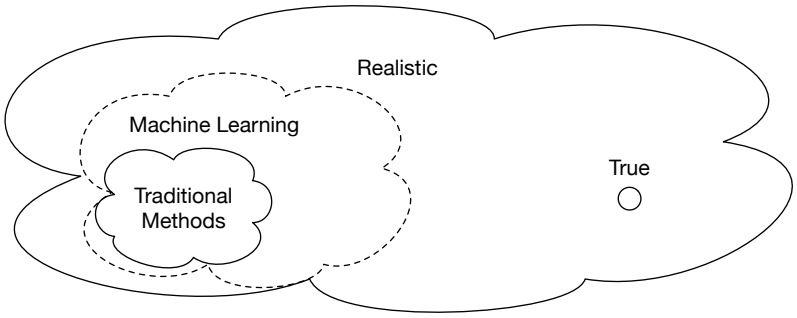
Variations on simple measures of correlation and linear models (including linear SEMs) were found to be the most frequently used modeling techniques in psychology research in recent years (Blanca, Alarcon, and Bono, 2018; Bolger, Zee, et al., 2019).[7]

The principal assumption associated with these models is that the true relationships between the variables are sufficiently represented as linear. Such models therefore have a limited functional form that can only represent linear relationships. In other words, they describe relationships between predictor and outcome variables that can be summarized in terms of a weighted sum. Of course, in reality the true relationship between variables may be highly complex and nonlinear. Indeed, assuming our dataset is sampled from a 'true' population distribution, there exists a 'true' functional form describing the functional relationships between the variables. Figure 2.1 illustrates how traditional methods (including linear regression) have the most limited capacity (owing to strong restrictions on the functional form) to model complex real-world phenomena (Coyle et al., 2020; M. J. van der Laan and S. Rose, 2018; M. J. van der Laan and Starmans, 2014).

A discussion about the limitations of linear predictive/causal models and correlation is not new (M. J. van der Laan and S. Rose, 2011; Asuero, Sayago, and A. Gonzalez, 2006; Onwuegbuzie and Daniel, 1999; Achen, 1977; King, 1986; Meehl, 1990; Taleb, 2019). However, in spite of this prior commentary there is evidence that researchers in psychology and social science may still be reluctant to adjust their methodological practice accordingly (e.g.,Ernst and Albers, 2017; Yarkoni and Westfall, 2017).

[7] It might be argued that any arbitrary function can be represented as some linear sum of features, and that therefore all models are fundamentally linear. However, using such a broadly encompassing definition term 'linear model' makes discussion pedantic. As such, we use the term to describe the typical linear regression model where the outcome is modeled as a linear sum of raw variables or low-order functions of these variables (such as exponents: $\mathbf{x}^1, \mathbf{x}^2$; and interactions: $\mathbf{x}_1\mathbf{x}_2$ etc.).

Figure 2.1: Approximating Realistic Data Distributions



Note. Traditional techniques such as linear regression may be severely limited in their capacity to model highly complex, non-linear data. Machine Learning methods may help to expand the coverage of realistic data distributions, but the true distribution may still lie outside. Combining flexible function approximation techniques from machine learning, with an incorporation of domain knowledge and model structure, can help us get as close as possible to modelling the true data distribution (M. J. van der Laan and S. Rose, 2011).

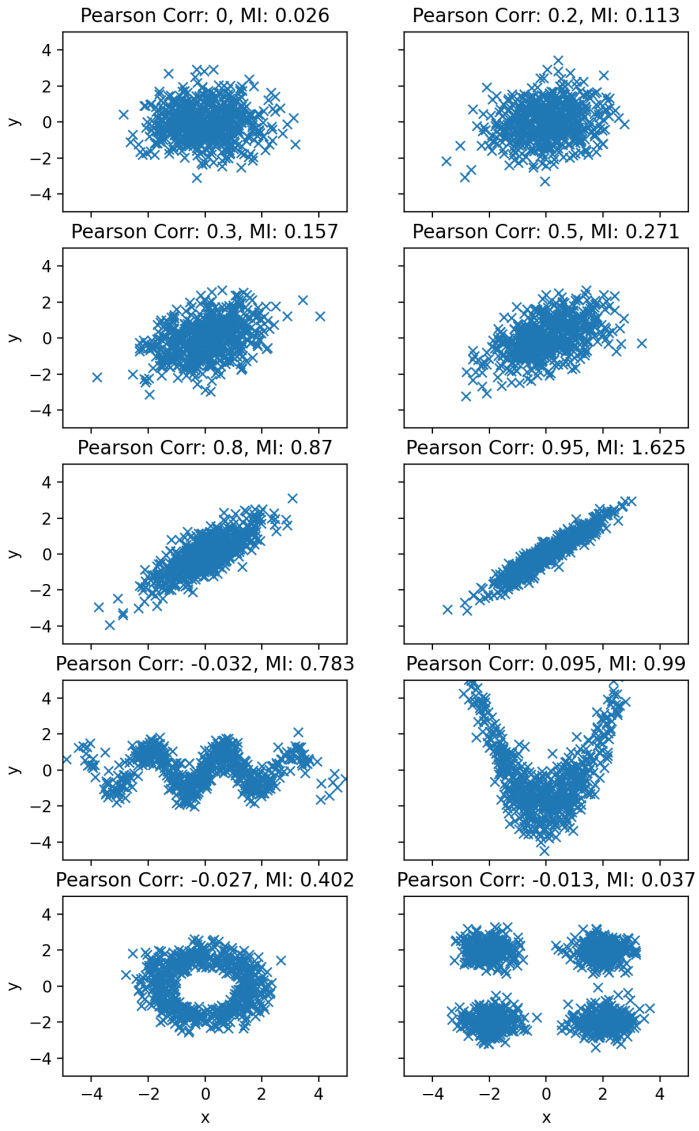
Correlation

Correlation is generally used to measure the association or statistical dependence between variables (i.e., to identify variables which may be good predictors). It ranges between $[-1, 1]$ and can be used as a basic predictive model. For example, when one variable is high, a correlated variable is also likely to be high. However, as one of the most common ways to measure dependence, there are two important aspects relating to correlation to bear in mind, particularly when interpreting or drawing conclusions about measures of correlation.

In the bivariate case, the coefficients from a standardized linear regression correspond with the correlation between the predictor variables and the outcome. Figure 2.2 shows a number of bivariate distributions along with their correlation coefficient. The first thing to note from the upper six plots is that correlation itself is a non-linear metric for dependence. Lower values of the Pearson Correlation Coefficient (PCC) are associated with a disproportionately lower dependence than higher values (and this is also reflected visually in the plots). The second thing to note from the lower four plots is that the PCC catastrophically fails to capture non-linear dependence.

The first issue is important for researchers to understand when drawing conclusions about relative levels of correlation. For example, the difference between $PCC = 0.1$ and $PCC = 0.2$

Figure 2.2: Pearson Correlation and Shannon Mutual Information.



Note. Simulations demonstrating the relationships between the Pearson measure of correlation, and the Mutual Information metric for measuring statistical dependence. The upper six plots depict linear bivariate relationships, whereas the lower four plots are non-linear.

is less dramatic than, say the difference between $PCC = 0.8$ and $PCC = 0.9$, in spite of the former describing a much higher proportionate increase. The second issue relates to an assumption of linearity: If the relationship between the two variables is linear, then correlation provides a measure of linear dependence; if the relationship is non-linear, then correlation may provide meaningless measures of dependence. In cases where the relationship is non-linear, researchers will need to either linearize the relationship (e.g., by creating a new variable that accounts for this non-linearity), or consider using an alternative measure of dependence. One such alternative to correlation is Shannon Mutual Information (M.I.), which gives us a measure for how much information one variable contains about another (Cover and Thomas, 2006; Kraskov, Stogbauer, and Grassberger, 2004; G. V. Steeg and Galstyan, 2012; G. Steeg and Galstyan, 2013; Gao, G. Steeg, and Galstyan, 2015; Kinney and Atwal, 2014). The estimates for M.I. are also shown in Figure 2.2, and it can be seen that M.I. not only handles non-linear relationships between variables, but also increases linearly with the degree of dependence of the variables. Note that M.I. ranges between $[0, H]$ where H is the entropy of either distribution when the two distributions are identical (i.e., $I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) = H(\mathbf{y})$ when $\mathbf{x} = \mathbf{y}$).[8] M.I. cannot be negative, and as such it is not able to indicate the 'direction' of the association in the way that correlation can. However, this is an acceptable limitation given that many non-linear relationships are non-monotonic (i.e. they are not always either increasing or decreasing) and in these cases a notion of positive or negative direction is unhelpful.

Linear Models/Regression

Linear regression is another very common modeling technique used for both predictive and causal modeling. In the case of a typical linear multiple regression in psychology or social science (which constitute a relatively small sub-class in the class of Generalized Linear Models), the predictive mapping function f consists of a weighted sum of basis functions of the features or variables $\mathbf{X}$. These basis functions are usually exponents of the variables/features (e.g., $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2 ...$). In most cases, the input features constitute the raw (or, at most, normalized and/or transformed) data collected from Likert scales, demographics, or coded observations.

[8] Readers are pointed to Cover and Thomas (2006) for an introduction to information theoretic concepts such as entropy and mutual information.

Sometimes, combinations of features are included which represent interactions (e.g. $\mathbf{x}_{ik} \times \mathbf{x}_{ik'}$) for $k \neq k'$. The regression function is usually fit using a predictive heuristic such as Ordinary Least Squares (OLS). OLS finds the solution to the regression such that the values of the function parameters minimize the average squared difference between predictions and observations $\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \frac{1}{N} \sum_i^N (\hat{y}_i - y_i)^2$. Here, $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\theta}}$ represents a vector of dimension K of weighting coefficients/parameters estimated from the sample. These parameters derive from a family of possible parameters $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$, which in turn define a space of possible linear functions $f_{\hat{\theta}} \in \mathcal{F}$. OLS therefore identifies the parameters $\boldsymbol{\theta}$ that minimize the mean squared error. The total predictive function may be represented as: $f_{\hat{\theta}}(\mathbf{X}) = \mathbf{y} = \mathbf{X}\hat{\boldsymbol{\theta}}$. Various link functions may be used to adapt the function to other outcome distributions (e.g., the logistic link for Bernoulli distributed outcomes).

There is one principal assumption for linear regression which is important for achieving both successful causal *and* predictive modeling. Namely, that the outcome can be well approximated using a weighted linear sum of the input variables. Indeed, the linearity imposes a strong functional constraint that restricts the function's flexibility and is, therefore, an assumption about *functional form* (M. J. van der Laan and S. Rose, 2011). Linear methods are unlikely to match the functional form of realistic data distributions, and to get closer to the true functional form, researchers should consider using more flexible predictive methods.

2.2.3 Improving on the Functional Form of Linear Models

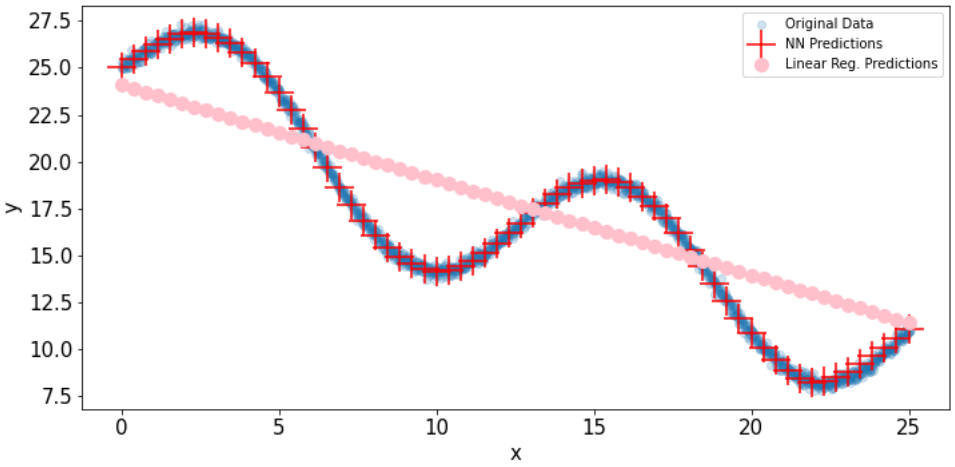
In order to improve the predictive or associational performance of a predictive function, researchers may need to explore either *feature engineering* approaches, or other functional approximation techniques such as those commonly used in machine learning. Introducing hierarchical structure within linear functions can improve the fit (Yarkoni, 2019; Gelman and Hill, 2007; Bolger, Zee, et al., 2019), but even hierarchical linear models are constrained according to linear functional associations.

Feature engineering involves the substitution of raw input variables with functions of these variables called *features*. Depending on the functional form used to derive these features, the features themselves may then be linearly related to the outcome, facilitating better overall

functional approximation. For instance, researchers may include more exotic basis functions such (e.g., sinusoidal functions; M. J. Vowels, K. Mark, et al., 2018, or kernels; Scholkopf, 2019), or simply combine features to form new ones (e.g., interaction features which are composed by multiplying two variables together). Feature engineering may thereby help to account for the non-linearities of the data in the features themselves, but in doing so, each feature may need to be carefully chosen or designed. For example, in Figure 2.2, the plot in the fourth row on the right has a simple basis function which is $\mathbf{x}^2$. While the raw values of $\mathbf{x}$ could not be used to model the outcome as part of a linear sum, the squared values could be used to essentially linearize the predictor in question. However, in real-world applications (i.e., research scenarios with real data) we will not know the functional form *a priori* and it may be difficult to ascertain. For instance, the function may not be an exact quadratic function $\mathbf{x}^2$, but some other, arbitrarily complex function. The feature engineering process may or may not be guided by knowledge about the domain of interest. For example, in the case of a time series with known seasonal variation (e.g., financial data exhibiting fluctuation due to the business cycle) the use of sinusoidal basis functions may be well justified and aid prediction and generalization (Hamilton, 1994; M. J. Vowels, K. Mark, et al., 2018).

Besides generalized linear models with feature engineering, there exist many alternative and much more powerful function approximation techniques, such as those common in machine learning. These techniques are able to *learn* functional relationships from the data themselves and can be used instead of, or in combination with, feature engineering. For instance, random forests (Breiman, 2001a) comprise a group of decision trees that are capable of learning highly non-linear relationships and interactions between variables, without these interactions needing to be pre-specified. The mapping learned by the forest adapts to the data in order to minimize a performance objective (e.g., mean squared error). One of the advantages of random forests is that they employ bootstrapping and thereby mitigate problems with learned functions overfitting the data. Neural networks are an alternative approach to function approximation which are also data-adaptive and are highly parameterized (sometimes with billions of parameters) (I. Goodfellow, Bengio, and Courville, 2016). They learn by iteratively updating their parameters according to an error signal until some criterion for convergence is met. An example of predictions from a simple neural network compared with those of a linear regressor on a bivariate problem is

Figure 2.3: Neural network versus linear regression function predictions.



Note. Demonstrates how linear functional forms cannot capture the non-linear relationships. In contrast, non-linear, data-adaptive techniques such as neural networks, can.

shown in Figure 2.3. It can be seen the neural network has fit the data almost perfectly, whilst the linear regression approximates the mean slope of the line, ignoring the cycling fluctuation.

2.2.4 Overfitting and Double-Dipping

As described previously overfitting and double-dipping refer to the consequences of various modeling practices which increase the fit of a model to a specific data sample, but which negatively impact the validity and generalizability of results. An awareness of overfitting becomes increasingly crucial when attempting to model non-linear functional relationships between variables. These topics have been extensively covered elsewhere, particularly in the machine learning literature (where overfitting is sometimes associated with what is known as the bias-variance trade-off) (Belkin et al., 2019; Bishop, 2006; R. R. Murphy, 2000; Yarkoni and Westfall, 2017; Mayo, 2013). Prior research has highlighted how modeling practices that result in overfitting are common in psychology and social science, as well as a number of other fields, and have been noted for their possible contribution to the replicability crisis (Shrout and Rodgers, 2018; Gelman and Loken, 2013; Yarkoni and Westfall, 2017). Even the common forward and backward method for variable inclusion constitutes data-driven overfitting practices which have the potential to significantly impact model generalizability and interpretability, and

yet these practices are routinely included as part of standard statistical education and practice in psychology (e.g., see Field, 2009). We mention such (mis)practice again here because, when using powerful function approximation techniques, a consideration for overfitting is even more important. There are numerous techniques for mitigating issues with overfitting, including regularization, cross-validation, train-test splits etc. and it is important that researchers in psychology and social science familiarize themselves with these fundamental concepts, especially when accounting for complex, non-linear associations between variables.

2.2.5 Summary

In Part 1, we presented how models with limited functional form may be unable to represent the complex relationships between variables. The typical analyses used in psychology and social science include simple measures of correlation, and various manifestations of linear regression. While such modeling techniques are limited in their predictive capacity, there are many algorithms used in the field of machine learning which can learn an appropriately flexible functional form from the data themselves. When using more powerful techniques, it is especially important to validate models on an out-of-sample test set (e.g., by using a cross-validation method, or train/test splitting) in order to avoid overfitting. However, it is worth noting that overfitting (and the related problem of double-dipping) is also possible with simple linear models, and prior meta-research suggests that researchers may be unaware of these issues. Finally, the rarity of modeling techniques with powerful, data-adaptive functional form represents a possible missed opportunity in psychology and social science, and we encourage researchers to consider the functional form of their models, and familiarize themselves with the associated pitfalls and limitations (e.g., overfitting), in order that they can get closer to modeling the true relationships underpinning the phenomenon under study.

2.3 Part 2: Causal Model Misspecification

As described in the Introduction, prior research has highlighted a reluctance to adopt explicit causal approaches (Grosz, Rohrer, and Thoemmes, 2020; M. Hernan, 2018a). Causal techniques

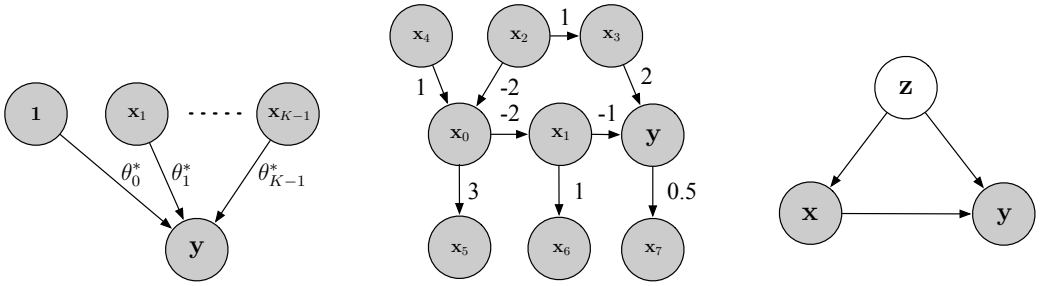
provide the means to answer fundamental questions that help us to develop an understanding of the world (Pearl, 2009; M. J. van der Laan and S. Rose, 2011). To the best of our knowledge, we are not aware of a well-established theory in psychology or social science which does not incorporate at least some level of consideration for cause and effect, and, if there is one, we would question its utility in so far as it can help us understand the world. Models which sufficiently align with the structure of reality may facilitate causal inference, even with observational (as opposed to experimental) data (C. Glymour, 2001; Pearl, 2009; Pearl, M. Glymour, and Jewell, 2016; Grosz, Rohrer, and Thoemmes, 2020) and have wide ranging applications including advertisement (Bottou et al., 2013), policy making (Kreif and DiazOrdaz, 2019), the evaluation of evidence within legal frameworks (Pearl, 2009; Siegerink et al., 2016), and the development of medical treatments (Petersen et al., 2017; M. J. van der Laan and S. Rose, 2011). There are a number of challenges associated with adopting a causal approach.

Misspecification represents one of the principal challenges associated with causal inference, and arises when the true causal structure and/or the functional form of the relationships between variables in the data generating process are not sufficiently reflected in a causal model. Misspecification results in biased effect size estimates which are not meaningfully interpretable. In this Part, we primarily focus on misspecification stemming from problems associated with structure and to do so, we consider misspecfication in restricted linear settings. As we will show, even in this restricted setting, it is extremely important that the model sufficiently accounts for the true structure of the data in order that the resulting model is interpretable. We stress that this section is not intended as a technical guide to undertaking causal inference in general (for more information on causal inference see e.g., Pearl, 2009; Petersen et al., 2017; Pearl, M. Glymour, and Jewell, 2016; C. Glymour, 2001; Angrist and Krueger, 2001; D. B. Rubin, 2005; Gelman and Hill, 2007).

2.3.1 Recovering Causal Effects

Given the frequency with which psychologists and social scientists adopt linear regression methods to test causal theories (Shmueli, 2010; Blanca, Alarcon, and Bono, 2018), it is extremely important that researchers understand the structural bias associated with the use of

Figure 2.4: Simple Directed Acyclic Graphs



(a) (b) (c)

Note. Example causal Directed Acyclic Graphs (c-DAGs). Example (a) depicts the case where all 'predictor' or causal variables are exogenous (i.e., they have no causal parents and are independent of each other). This corresponds with the causal structure of a simple multiple regression, where the dependent outcome $\mathbf{y}$ is a linear sum of the $\mathbf{x}$ variables. The empirical causal effect of each variable is equivalent to the multiple regression coefficient estimates. Example (b) is adapted from J. Peters, Janzing, and Scholkopf, 2017. Example (c) depicts a graph with an unobserved confounding variable $\mathbf{z}$.

such models. In this section, we demonstrate how typical linear regression models used in psychology and social science impose a strong implicit causal/structural form which is unlikely to reflect the true causal structure of the data, even when the functional form is linear, and are therefore misspecified. We show that, through a consideration of the causal structure of the phenomenon under study, one can nonetheless use linear regression to recover causal effects under a number of restrictive assumptions.

Multiple Regression Without Misspecification

In this section we demonstrate the strong, implicit structural form associated with multiple regression. We begin by demonstrating that multiple linear regression (in its basic form) is not misspecified with respect to the true data generating process when all predictors are exogenous (see structure in Figure 2.4(a)). In such a scenario, the resulting model is interpretable.

If the true data generating process could be described as a weighted sum of a set of input variables, then our goal of prediction within the Ordinary Least Squares multiple linear regression framework (as described in the previous section) would also be adequate for causal modeling, causal parameter estimation, or causal inference. Such a model might be depicted graphically as in Figure 2.4(a). In this scenario, there would exist parameters $\boldsymbol{\theta}^*$ (also known as effect sizes)

which represent the true causal parameters, and our OLS-derived parameters would represent empirical/sample estimations thereof.

The graphs in Figure 2.4 are known as causal Directed Acyclic Graphs (c-DAGs), and they represents a generalization of the graphical representation often used in Structural Equation Modelling (SEM) (Pearl, 2009; Koller and Friedman, 2009; Rohrer, 2018). The arrows indicate causal directional relationships between variables, parameterized by θ, and the grey nodes indicate observed variables. The *acyclicity* pertains to the restriction that there can be no closed loops (i.e., feedback) in the graph. Graph terminology (e.g., 'parent', 'ancestor', 'descendant', 'child') is useful in describing the top-level relationships between variables. For example, a node with an incoming arrow is a child of its parent variable, and further upstream or downstream variables are ancestors or descendants respectively.

In general, the arrows in a c-DAG indicate causal dependencies, and there is no implied functional form that prescribes how the variables are combined at a node (i.e., there could be highly non-linear, adaptive functions with interactions). Furthermore, the nodes represent variables which may or may not be univariate or parametric. In other words, a node labelled $\mathbf{x}$ does not restrict the dimensionality or (non-)parameterization of $\mathbf{x}$ itself. For instance, a node $\mathbf{x}$ could comprise multiple predictors which do not conform to a parameterized distribution. Hence, c-DAGs encode the fundamental essence of the causal structure, without imposing potentially irrelevant restrictions. We have included some extra information in the c-DAG of Figure 2.4(a) for the sake of demonstration. This particular c-DAG represents the intercept parameter of a multiple linear regression as a vector of ones multiplied by the parameter θ_0^*. The structural equations for this graph may be represented in Equation 2.1:

$$\begin{aligned} \mathbf{x}_{k=0} &:= \mathbf{1} \\ \mathbf{x}_k &:= \mathbf{U}_k(0,1) \text{ for } k = 1, ..., (K-1) \\ \mathbf{y} &:= \sum_{k=0}^{K-1} \theta_k^* \mathbf{x}_k + \mathbf{U}_y(0,1) \text{ for } k = 0, ..., (K-1) \end{aligned} \tag{2.1}$$

Let us assume that $\mathbf{U}_k$ and $\mathbf{U}_y$ are N-dimensional vectors of identically and independently

distributed (i.i.d.) normally distributed random noise. The ':=' symbol (endearingly referred to as the walrus operator in the Python programming world) denotes *assignment* rather than equality. This distinction is useful in reflecting the structural/causal direction of the arrows in the c-DAG. For example, the outcome $\mathbf{y}$ is a function of its inputs, and the equation should not be rearranged to imply that the inputs are a function of the outcome (the arrows point in one direction). These equations encode the fact that all the input variables are exogenous (i.e. completely independent of each other and determined only by i.i.d. noise) and that the outcome is a weighted linear combination of these variables. In this setting we might understandably refer to the input variables as the independent variables, and the outcome as the dependent variable. As mentioned, these equations correspond with a simple multiple linear regression and can be solved to find $\boldsymbol{\theta}$ using OLS. We demonstrate this by undertaking a simulation for $K = 4$ with $\theta_0^* = 3.3$, $\theta_1^* = 0.1$, $\theta_2^* = 0.3$ and $\theta_3^* = 0.5$. We set $N = 5000$ so that we do not have to be concerned about the stochastic variability associated with small samples, and the results are shown in Table 2.2.

```
import statsmodels.api as sm
import numpy as np
N = 5000  # N = sample size
# simulate data
x1 = np.random.randn(N,1)
x2 = np.random.randn(N,1)
x3 = np.random.randn(N,1)
X = np.concatenate((x1, x2, x3), 1)  # combine predictors into array
y = 3.3 + 0.1*x1 + 0.3*x2 + 0.5*x3 + 0.3 * np.random.randn(N,1)
X = sm.add_constant(X, prepend=True)  # add intercept term as x0
mod = sm.OLS(y, X)  # initialize multiple regression model
res = mod.fit()  # fit the regression model
```

From this demonstration it can be seen that the OLS regression successfully recovered $\hat{\boldsymbol{\theta}}$ close to $\boldsymbol{\theta}^*$. In this case, the data generating process directly matched the model we used to estimate the parameters and was therefore *not* misspecified. When there is no misspecification, the estimated

Table 2.2: Estimated parameters for DAG in Figure 2.4(a).

	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
$\mathbf{y}$	3.31	0.11	0.31	0.50

parameters may be interpreted as *causal* parameters that tell us about the phenomenon (in this case, a simple, simulated phenomenon). Indeed, the parameters here can be interpreted as 'one unit increase in $\mathbf{x}_1$ yields a θ_1 increase in $\mathbf{y}$', as is common practice in psychology and social science.

The interpretability of the model was only possible because the structure of the data matched the structure of a multiple linear regression, equivalent to Figure 2.4(a), where all 'predictors' are exogenous. However, this is an unrealistic scenario, and in most real-world cases, the predictors will not be exogenous. In the next section we demonstrate what happens when we apply the multiple regression model to scenarios when the causal structure is more realistic.

Multiple Regression - Misspecified for Realistic Structure

In the previous section we showed how a simple multiple regression can be used to recover meaningful, causal parameter estimates, so long as the true causal structure of the data corresponds with the implicit causal structure implied by the multiple regression. However, the implicit causal structure of a linear regression is extremely restrictive and, when modeling real-world data, it is likely to be misspecified. In this section we demonstrate what happens when such misspecification occurs.

Let us see what happens when we follow the same procedure to try to estimate some parameters for another simple data generating process which follows the example in Figure 2.4(b). We assume the following data generating structural equations (adapted from J. Peters, Janzing, and Scholkopf, 2017):

$$\begin{aligned} \mathbf{x}_4 := \mathbf{U}_4, \quad \mathbf{x}_2 := 0.8\mathbf{U}_2, \quad \mathbf{x}_0 := \mathbf{x}_4 - 2\mathbf{x}_2 + 0.2\mathbf{U}_0, \quad \mathbf{x}_1 := -2\mathbf{x}_0 + 0.5\mathbf{U}_1, \\ \mathbf{x}_3 := \mathbf{x}_2 + 0.1\mathbf{U}_3, \quad \mathbf{x}_5 := 3\mathbf{x}_0 + 0.8\mathbf{U}_5, \quad \mathbf{x}_6 := \mathbf{x}_1 + 0.5\mathbf{U}_6, \\ \mathbf{y} := 2\mathbf{x}_3 - \mathbf{x}_1 + 0.2\mathbf{U}_y, \quad \mathbf{x}_7 := 0.5\mathbf{y} + 0.1\mathbf{U}_7 \end{aligned} \tag{2.2}$$

For these equations we have simplified the notation to make things clearer: $\mathbf{U}_k \sim \mathcal{N}(0, 1)$. The structural process is still linear and the additive noise is Gaussian, so we do not yet need to worry about utilizing flexible function approximation techniques (such as those discussed in Part 1).

It is worth studying these equations to understand their implications. Note that, for instance, $\mathbf{x}_3$ is only determined by $\mathbf{x}_2$, as well as its own exogeneous noise $\mathbf{U}_3$. This means that, if we perform surgery on these equations by, for example, setting $\mathbf{x}_3$ to a different value or distribution, we have cut off its dependence to its parent. Such graph surgery enables us to explore a range of causal queries such as interventions and counterfactuals, and is formalized by Pearl's *do*-calculus (Pearl, 2009).

Given the simple linear form in Equation 2.2 for Figure 2.4(b), it is possible to traverse the paths in the c-DAG and to combine the effects multiplicatively. Such a process should be familiar to those who have studied path diagrams and SEM (Kline, 2005). For instance, the effect of $\mathbf{x}_0$ on $\mathbf{y}$ is the multiplication of the effect of $\mathbf{x}_0 \to \mathbf{x}_1$ with the effect of $\mathbf{x}_1 \to \mathbf{y}$. Together, we have the mediated path: $\mathbf{x}_0 \to \mathbf{x}_1 \to \mathbf{y}$. According to Equation 2.2 and Figure 2.4, the effect of $\mathbf{x}_0$ on $\mathbf{y}$ therefore corresponds with $-2 \times -1 = 2$. In this case, $\mathbf{x}_1$ is *mediating* the effect of $\mathbf{x}_0$ on $\mathbf{y}$. Readers may already be aware of the issues relating to the inclusion of mediators in a regression analysis (see e.g., Cinelli, Forney, and Pearl, 2022; Rohrer, 2018; Pearl, 2009), and this is trivially demonstrated by comparing the regressions of $\mathbf{y}$ onto $\mathbf{x}_0$ whilst (a) adjusting for $\mathbf{x}_1$ and (b) and not adjusting for $\mathbf{x}_1$. Here, adjusting for a variable is equivalent to *controlling* for it, but the adjustment terminology is more appropriate for structural scenarios (Pearl, 2009). First let us simulate the data as follows:

```
N = 5000
```

```
x4 = np.random.randn(N,1)
x2 = 0.8 * np.random.randn(N,1)
x3 = x2 + 0.1 * np.random.randn(N,1)
x0 = x4 - 2*x2 + 0.2 * np.random.randn(N,1)
x5 = 3*x0 + 0.8 * np.random.randn(N,1)
x1 = -2*x0 + 0.5 * np.random.randn(N,1)
x6 = x1 + 0.5 * np.random.randn(N,1)
y = 2*x3 - x1 + 0.2 * np.random.randn(N,1)
x7 = 0.5*y + 0.1 * np.random.randn(N,1)
```

These Python variables reflect those in Equation 2.2 above. The bivariate correlations and p-values for each of these variables are shown in Table 2.3.

Table 2.3: Bivariate Pearson correlations and p-values for the DAG in Figure 2.4(b).

$r(p)$	$\mathbf{x}_0$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$
$\mathbf{y}$	.92(.00)	-.92(.00)	-.58(.00)	-.56(.00)	.76(.00)	.91(.00)	-.93(.00)	1.00(.00)

The results in Table 2.3 demonstrate a strong and statistically significant bivariate correlation between each predictor and the outcome. Now, when using only $\mathbf{x}_0$ as a sole predictor in a simple linear regression model, we estimate the effect of $\mathbf{x}_0$ on $\mathbf{y}$ to be $\hat{\theta}_0 = 1.28$, where the $\hat{}$ notation indicates it is an empirical estimate. Recall that the true effect of $\mathbf{x}_0$ on $\mathbf{y}$ is 2. In spite of the large sample size, the output estimate is highly biased and does not seem to correspond with any of the parameters in the original simulation. Indeed, regardless of how large the sample size is, this coefficient estimate will converge to a value that is far from the true estimand. This is because the structure of the data generating process was not considered: We simply applied a linear regression to the data without accounting for the fact that the implicit structure of a linear regression does not match the structure in the data. In this situation, the regression might still have some limited utility as a purely *predictive* function, but its parameters should not be interpreted as anything relevant to the causal structure of the phenomenon of interest because it is *misspecified*.

When confronted with the dilemma of multiple observed variables, typical practice in psychology

and social science might involve using the forward or backward method for variable inclusion (Field, 2009). Besides the problems associated with such practice and overfitting (as described in Part 1 above), such practice is likely to result in misspecification. Another approach might be to simply include all variables in the model. Indeed, all the $\mathbf{x}_k$ variables are highly and statistically significantly correlated with the outcome $\mathbf{y}$, so if we were not already aware of the implicit causal structure of linear regression, this might seem like a sensible thing to do. When we include all variables in the model, this results in $\hat{\theta} = -0.01$. Recall again that the true effect of $\mathbf{x}_0$ on $\mathbf{y}$ is 2. The estimate of -0.01 is highly biased. This is because including all the variables in the model imposes the structure shown in Figure 2.4(a), where all variables are exogenous.

Including $\mathbf{x}_0$ and the mediating variable $\mathbf{x}_1$ confirms that including mediating variables is problematic: The regression including both $\mathbf{x}_0$ and $\mathbf{x}_1$ yields $\hat{\theta} = -.94$. As expected, the effect of $\mathbf{x}_0$ on the outcome is highly biased, and of the opposite sign (i.e., negative rather than positive) to the true causal effect. It should now be clear that the use of what might be called naive multiple regression cannot yield meaningfully interpretable parameters unless the model corresponds with Figure 2.4(a), and this is highly unlikely. Indeed, it is arguable as to whether the interpretation of this parameter (and even its direction) is of any scientific value at all. Utilizing hierarchical or Bayesian approaches will not help so long as the structure of the model is misspecified.

Addressing Misspecification Using Causal Inference Techniques

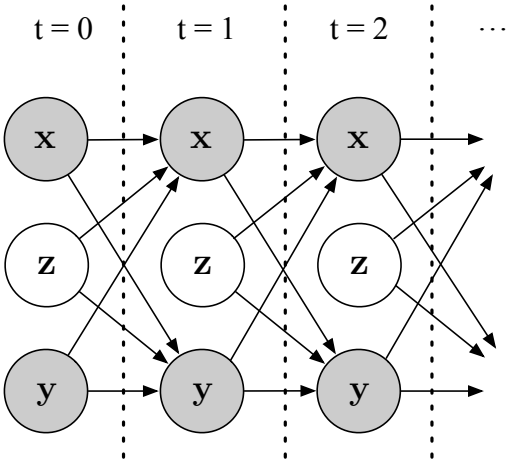
We have seen that using naive multiple regression is inadequate when trying to estimate a causal effect from data with a non-trivial structure, even when the underlying functional form of the relationships is linear. Whether or not the structure is of relatively low complexity, the resulting coefficient estimates can be wildly biased. This illustrates that, regardless of whether the functional form matches the true functional form of the data (and in the simulations above, it did), it is impossible to recover meaningful effect size estimations with a misspecified model. In order to recover an unbiased estimate of the true effect, we need to understand techniques from the field of causal inference.

Structural Equation Modelling (SEM) was reported to be one of the most common methods used in psychology and social science (Blanca, Alarcon, and Bono, 2018), and enables unbiased estimation of the parameters, so long as the structure of the SEM model matches or at least subsumes the structure of the data generating process, and so long as a number of restrictive assumptions are met (J. Peters, Janzing, and Scholkopf, 2017). These assumptions apply to causal inference in general. The subsumption point relates to the fact that researchers, when faced with uncertainty about the structure of the data generating process, should choose to expand their model class rather than restrict it. In other words, researchers should, in general, choose to include an extra arrow in their SEM rather than remove one. The choice to expand the model allows for the possibility of an effect in the data, whereas a removal of a causal link enforces an absence of dependency and thereby represents a strong model restriction that needs to be well justified before its imposition.

In practice, we rarely have access to the true model when we create an SEM (D'Amour, 2019; Y. Wang and Blei, 2019; Tenenbaum and Griffiths, 2002). Indeed, as the SEM grows in complexity and/or its causal constraints, the chance of it becoming misspecified increases. If certain assumptions are made, and we reduce our goal to the estimation of a specific and restricted set of effects (e.g., just the effect of $\mathbf{x}_0$ on $\mathbf{y}$), it may be sufficient to leverage domain knowledge and causal inference techniques to acquire a reliable estimate without having to correctly specify the full graph. Such techniques have been extensively covered elsewhere (J. Peters, Janzing, and Scholkopf, 2017; Pearl, 2009; Imbens and D. Rubin, 2015; Pearl, M. Glymour, and Jewell, 2016; Angrist and Krueger, 2001) and include the use of instrumental variables, propensity score matching, and regression discontinuity designs (Blossfeld, 2009), but we briefly cover one particular technique known as *backdoor adjustment* below (Pearl, 2009).

Backdoor adjustment involves identifying what are known as *backdoor paths*. An example of a backdoor path between $\mathbf{x}_0$ and $\mathbf{y}$ in Figure 2.4(b) is $\mathbf{x}_0 \leftarrow \mathbf{x}_2 \rightarrow \mathbf{x}_3 \rightarrow \mathbf{y}$. $\mathbf{x}_2$ and $\mathbf{x}_3$ are therefore part of what is known as the backdoor adjustment set; a set of variables which, if adjusted for, block the backdoor path. We can adjust for all the backdoor variables, or the minimal set sufficient to block the path (in our case, either $\mathbf{x}_2$ or $\mathbf{x}_3$ will do). Including $\mathbf{x}_0$ and $\mathbf{x}_3$ yields $\hat{\theta} = 2.00$.

Figure 2.5: Example Directed Acyclic Graph for Time Series



Note. c-DAG for a time series setting, highlighting the complexity associated with identifying a particular causal effect, especially when there may be unobserved confounding (J. Peters, Janzing, and Scholkopf, 2017).

We have now recovered an unbiased estimate of the effect of $\mathbf{x}_0$ on $\mathbf{y}$ (which was approximately equal to two), and we only needed to regress $\mathbf{y}$ onto two variables, despite our world knowledge dictating that at least eight were involved in the data generating processes as a whole (indeed, all variables in this simulation are highly and significantly correlated with the outcome). If we are also interested in the mediation through $\mathbf{x}_1$ then we can undertake separate regressions to break the problem down. The estimated parameters are then meaningfully interpretable insofar as the correspond with the parameters in the true data generating process. In other words, if $\theta = 2$, then every unit increase in $\mathbf{x}_0$ results in two units increase in $\mathbf{y}$.

Does Time Help?

Researchers may believe that the inductive bias imposed with the directionality of time is helpful in identifying the causal effect and correctly specifying a causal model. Indeed, the fact that time cannot flow backwards constrains the possible directions of our arrows in our c-DAG, and therefore reduces the complexity of a time series model. However, in spite of the fact that a time series model may be the only way to answer a certain causal question, such time series problems may be far more complex than cross-sectional models, owing to the introduction of the additional time dimension. Therefore, certain causal questions may only be answerable

by considering time, but the causal effect of interest may be considerably harder to identify as a result. Figure 2.5 depicts a simple scenario with two variables, $\mathbf{x}$ and $\mathbf{y}$, and a hidden confounder $\mathbf{z}$. Each variable influences its own future as well as the future of the other variable. In the presence of the unobserved confounder the causal effect between $\mathbf{x}$ and $\mathbf{y}$ (however this might be defined) is *unidentifiable*. The complexity of this graph could grow further still if we include causal arrows between $\mathbf{x}$ and $\mathbf{y}$ (and potentially $\mathbf{z}$) for the same time point (i.e., $\mathbf{x}$ and time one influences $\mathbf{y}$ at time one), or if we add any additional (un)observed variables. In spite of the restriction that the arrows cannot flow backwards, this structure therefore has the potential to be immensely troublesome from the point of view of identifiability. Indeed, the use of causal inference with time series phenomena is a very current and ongoing research topic in the fields of causal inference and machine learning (J. Peters, Janzing, and Scholkopf, 2017; Krishnan, Shalit, and Sontag, 2017; Lohmann et al., 2012). Interested readers are pointed to an accessible introduction of the topic, and its use in psychology, by Gische, West, and Voelkle (2020).

2.3.2 Challenges, Assumptions, and Limitations of Causal Modeling

It is worth emphasizing that, with only naive multiple linear regression models, we were unable to acquire a meaningful effect size estimate for non-trivial data generating process. This is because multiple linear regression imposes its own implicit structural/causal form which is likely to be misspecified when used in real-world applications. Indeed, we used a relatively simplistic synthetic simulation to demonstrate that multiple linear regression yields meaningless estimates, but in real-world applications the graph may actually be significantly more complex which makes it extremely challenging to correctly specify the structure of the c-DAG, and therefore to use techniques such as backdoor adjustment. This is because, without a sufficient understanding of the causal structure, we would be unable to identify the necessary backdoor adjustment variables.

More generally, it is extremely difficult to obtain reliable effect size estimates from observational data concerning complex real-world social phenomena using these techniques. Indeed, the infamous ‘crud’ factor, which describes the fact that “everything [in social science] correlates to some extent with everything else” makes causal inference in social science and psychology

particularly challenging (Meehl, 1990; Orben and Lakens, 2020).[9] One challenge is finding suitable backdoor adjustment variables, identifying other causal variables such as colliders, mediators, instrumental variables, proxy variables etc. so that the causal effect of interest is actually *identifiable* using the observed data (for techniques, see e.g., Cinelli, Forney, and Pearl, 2022; D. B. Rubin, 2005; Imbens and D. Rubin, 2015; Angrist and Krueger, 2001; Pearl, 2009; Y. Wang and Blei, 2019; D'Amour, 2019). Another challenge relates to the fact that social scientists are often concerned with the study of complex social systems with dynamic interdependencies. Such systems may not exhibit readily identifibale cause and effect pairs (Blossfeld, 2009).

In the same way that we chose to identify a *single* causal effect using the backdoor adjustment method, it may be beneficial for researchers to attempt to simplify their causal research questions. For example, in contrast with the typical use of SEM in psychology and social science (where the researcher attempts to derive multiple effect estimates simultaneously), targeted learning adopts the philosophy by 'targeting' a specific causal effect of interest, and orienting the analysis around its estimation using machine learning to reduce misspecification (M. J. van der Laan and S. Rose, 2011). The 'no free lunch theorem' familiar to machine learners applies here: causal inference yields the most information, but it is not easy (Wolpert and Macready, 1997). Attempting to undertake inference across multivariate, complex, linear SEM graphs is therefore extremely ambitious in light of its limited functional form and likely misspecification, and is highly unlikely to yield meaningful estimates. That said, exploratory work can still be highly valuable (Shrout and Rodgers, 2018). Part of the development process for SEMs (or, more generally, the underlying theory about the phenomenon) could involve causal directionality tests and validation via causal discovery techniques from machine learning (J. Peters, Janzing, and Scholkopf, 2017; Scholkopf, 2019). Such techniques, at least in restricted circumstances, may be able to test the directionality of the causal effects (Goudet et al., 2019; J. M. Mooij et al., 2010), identify backdoor adjustment set variables (Gultchin et al., 2020), estimate the magnitude of causal effects using flexible function approximation techniques (Yoon, J. Jordan, and van der Schaar, 2018; Shi, Blei, and Veitch, 2019), or infer hidden confounders from proxy variables

[9] The crud factor also results in an abundance of meaningless statistical significance, owing to the fact that null-effects are practically non-existent in social phenomena (Meehl, 1990).

using variational inference (Louizos, Shalit, et al., 2017; M. J. Vowels, N. Camgoz, and Bowden, 2021). We recommend both Targeted Learning (M. J. van der Laan and S. Rose, 2011) as well as deep latent variable neural network models (Louizos, Shalit, et al., 2017; M. J. Vowels, N. Camgoz, and Bowden, 2021) as possible approaches to the significant problem of causal effect size estimation, although many others exist (Gultchin et al., 2020; Shalit, Johansson, and Sontag, 2017; Shi, Blei, and Veitch, 2019; W. Zhang, L. Liu, and J. Li, 2021; Yao et al., 2018).

Even once a researcher believes that they have accounted for these difficulties and have simplified their research question or hypothesis, their consequent estimations then rest on the assumption known as *ignorability*; that there are no further latent/unobserved factors that must be somehow accounted for. Figure 2.4(c) depicts the presence of an unobserved confounder $\mathbf{z}$. Particularly in cases where researchers are dealing with observational (as opposed to experimental) data, the assumption of ignorability may be strong, untestable, and unrealistic. Other assumptions may also be relevant, depending on the causal question being asked, such as the stable unit treatment value assumption and the positivity assumption for estimating treatment effects. It is important researchers familiarize themselves with all relevant assumptions and limitations before undertaking causal inference, and make them explicit in their work (e.g., when they use SEM) (Grosz, Rohrer, and Thoemmes, 2020).

Finally, the simulations here assumed linear and additive structural equations of the form: $\mathbf{x}_1 := \theta_0 \mathbf{x}_0 + \mathbf{U}_1$. However, and as discussed earlier, c-DAGs are general and do not restrict the functional forms relating the variables. Indeed, in real-world scenarios the assumption of linearity may impair the capacity of the model to estimate unbiased coefficients, in much the same way as it limited predictive models (Coyle et al., 2020; M. J. van der Laan and S. Rose, 2011; M. J. van der Laan and Starmans, 2014; M. J. van der Laan and S. Rose, 2018). The difficulties of effect estimation are therefore compounded by the difficulties associated with identifying an appropriate functional form for the dependencies between variables (i.e., identifying what Blossfeld, 2009, calls “effect shapes”). Unless the structure of the model *and* its functional form sufficiently match those of the true data generating process, *and* we have an identifiable causal effect, the model may be misspecified and uninterpretable.

Causal Modeling in Practice

The three most common methods used in psychology are ANOVA, multiple linear regression (including hierarchical linear regression), and Structural Equation Modeling (SEM) (Blanca, Alarcon, and Bono, 2018). The first two are forms of linear model which encode strong *implicit* structural biases about the nature of the causal generating process (i.e., they encode the assumptions of exogenous independent input variables). The third method encodes *explicit* inductive bias relating to the causal generating process (Grosz, Rohrer, and Thoemmes, 2020). All three methods, tending to be linear, restrict the functional form associating the variables. The linearity and structural biases (whether implicit or explicit) yield misspecified models which are unlikely to match the true data generating process and pivot on untestable and unrealistic assumptions (such as strong ignorability). Misspecfication and the strong ignorability assumption are not of great concern if the goal is prediction: We may not care whether a mapping function reflects the data generating process, only that it provides good predictive performance.[10]

Furthermore, all three methods are frequently fit, evaluated, and manipulated according to predictive strategies (e.g., variable inclusion processes, structural changes) and the structure in the graph is not properly tested or validated (Scheel et al., in press; Kline, 2005; Ropovik, 2015). This is problematic for three reasons: First, linear models are not optimal for modeling complex real-world dependencies between variables; second, these models are rarely (if ever) tested on an out-of-sample dataset, meaning that any inference performed using these models is likely to have limited generalizability; third, the structure (and therefore, the practitioner's theory) is almost invariably accepted as valid *a priori* (Ropovik, 2015), despite misspecification being highly likely (M. J. van der Laan and S. Rose, 2011; VanderWeele, 2020).

Summary, and a Note on RCTs

It is important that researchers recognize the significant difficulties associated with estimating meaningful causal effects with observational data. We described how difficult it is to obtain

[10] However, predictive models may generalize better if they are robust to shifts in these unobserved confounders (Suter, Miladinovic, and Scholkopf, 2019).

reliable causal effect size estimates, and we have also demonstrated how a failure to consider the causal structure may yield biased, meaningless effect sizes, regardless of whether the researcher adopts a predictive or causal approach. We provided one example of a causal inference technique known as backdoor adjustment, as a way to identify the causal effect of interest. Doing so enabled us to simplify the analytical problem from one of estimating all path coefficients in a complex graph, to one of estimating a specific effect by identifying variables from an adjustment set. In practice, identifying these backdoor variables represents a significant challenge, because it requires sufficient causal knowledge. In addition to these difficulties, causal inference rests on a number of strong assumptions, perhaps the strongest of all being that of ignorability: That there are no unobserved confounders. Finally, researchers must also consider the functional form used to represent the causal dependencies between the variables. As such, problems with identifiability, ignorability, misspecification due to incorrect structure, and misspecification due to limited functional form have the potential to compound each other.

Given the complexity associated with avoiding misspecification, on top of considering functional form, readers may come to the conclusion that causal inference should be reserved for Randomized Controlled Trial (RCT) and experimental contexts. Actually, we do not think the situation is this clear-cut. The common view is that RCTs represent the "gold standard" of research. However, a growing literature highlights the limitations of RCTs, and how observational studies may, at least in certain circumstances, represent a promising alternative, particularly in terms of lower cost, reduced ethical implications, and larger sample size (Frieden, 2017; Deaton and Cartwright, 2018; Bothwell, Greene, and Podolsky, 2016; Jones and Podolsky, 2015). Furthermore, in a social science context, randomized experiments may be practically infeasible and potentially unethical (Blossfeld, 2009). To clarify, we do not wish to engage in a debate about the merits and pitfalls associated with undertaking causal inference on experimental versus observational data, but we do note that the perception of RCTs as representing a gold standard is potentially limiting and scientifically unhelpful.

2.4 Part 3: Unreliable Interpretations

In this part, we introduce explainability and interpretability, and describe how misspecified models with limited functional form may be neither explainable, nor interpretable. When the complexity of a model is increased to mitigate the issue of limited functional form it may be explainable in spite of possible misspecification due to incorrect structure. We discuss a range of problems relating to conflated and unreliable interpretations in psychology and social science. In our view, the conflation arises not just as a result of the alleged taboo against causal inference (Grosz, Rohrer, and Thoemmes, 2020), but also due to an apparent lack of understanding concerning the limitations associated with the interpretability of misspecified models with limited functional form and/or incorrect causal structure.

2.4.1 Explainability and Interpretability

Scrutinizing the parameters of a model in a predictive sense is referred to as *explaining*, in that we are explaining the behavior of the model, rather than *interpreting* the model's parameters in relation to some external real-world causal phenomenon (Rudin, 2019). We therefore distinguish *interpretability* from *explainability*. In this paper we use the term interpretation to describe the process of using a model to understand something about the structure in the data or phenomenon, and is therefore of particular relevance to causal approaches. As we will show, linear models are not immune to problems affecting interpretability both for reasons of limited functional form as well as misspecification (see Parts 1 and 2). Explainability, on the other hand, refers to the capacity to explain why a model makes a certain prediction or classification, based on its functional form or algorithmic rules (Rudin, 2019), and is therefore a term particularly relevant to predictive approaches. As the complexity of a model's functional form increases, it becomes increasingly difficult to either interpret or explain a model (Rudin, 2019).

2.4.2 The (Un)Interpretability of Linear Models

Linear models are deceptively simple to *explain* because their model coefficients seem to provide a direct means to understand why the model made a certain prediction. It is common to either

explain or interpret the parameters of a linear model as follows: For a one unit increase in $\mathbf{x}_k$, the model produces a θ_k increase in the outcome, assuming all other variables are fixed. If the model is not misspecified (i.e., it has adequate functional form and causal structure), then this parameter may be interpreted in a causal sense as well as in a predictive/explainable sense. In other words, the parameter not only tells us something about how the model's output changes with respect to a change in its input, but also something about the external phenomenon being modeled. However, if the model is misspecified due to incorrect structure, then the parameter may only be used to explain the behavior of the model, and will not correspond meaningfully with some external causal quantity.

Perhaps surprisingly, if the model is misspecified *both* in terms of its functional form and its structure, then the model may be neither interpretable nor explainable. In this scenario, complex cancellation effects may render the coefficients of linear models meaningless (Lundberg, G. Erion, et al., 2020; Breiman, 2001b; Haufe et al., 2014). Just because a predictive model (e.g., multiple linear regression) indicates that variable $\mathbf{x}_1$ has statistically significant association with an outcome, does not imply that it is meaningful to interpret this coefficient either in terms of a specific quantified value, or in terms of an ordinal level of variable importance. The problems are caused both by the function's inability to account for non-linear relationships and by the mismatch of the function's implicit structural (i.e., causal) form with the true form of the data. We demonstrated the latter issue in Part 2. For the former, we generate a synthetic example, closely following that of Lundberg, G. Erion, et al. (2020).[11] Essentially, the relationship between the outcome and two particular features in a semi-synthetic dataset is modified to include an increasing amount of non-linearity following the relationships in Equation 2.3.

$$\mathbf{y} = \sigma((1-q)(0.388\mathbf{x}_1 - 0.325) + q(1.714\mathbf{x}_1^2 - 1) + 1.265\mathbf{x}_2 + 0.0233) \tag{2.3}$$

Here, σ is the logistic link function, q is the degree of non-linearity, which is varied between zero (describing a linear relationship) and one (describing a model with a quadratic relationship), $\mathbf{y}$ is the outcome, and $\mathbf{x}_1$ and $\mathbf{x}_2$ are the two predictor variables. The choice of the factors (e.g.,

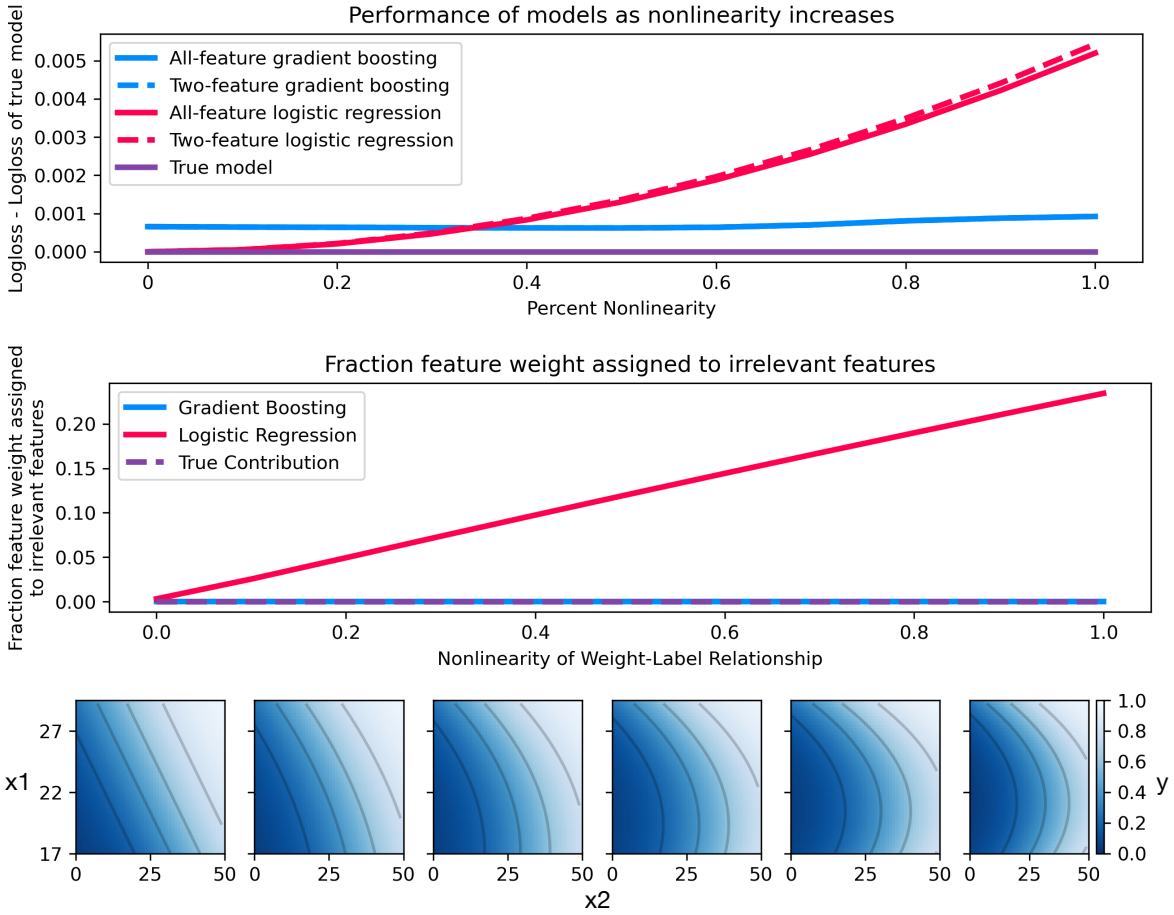
[11] Full code for the original example can be found here: `https://github.com/suinleelab/treeexplainer-study/`.

0.388) and intercepts (e.g., -0.325) are arbitrary, and derive from the classic NHANES I dataset (Launer, 1994; Fang and Alderman, 2000) from which the predictors and outcome are drawn. The relationship between the predictors and the outcome as q is increased from zero to one is shown in the lowest plot of Figure 2.6. Two models were fit to these synthetic data: a linear logistic regressor, and a machine learning algorithm known as *XGBoost* (T. Chen and Guestrin, 2016). The upper plot in Figure 2.6 shows how the logistic regressor's error increases as the non-linearity increases. In contrast, the XGBoost model's prediction error remains low. Notably, when q is close to zero (i.e., the percent non-linearity is low), the linear model outperforms the XGBoost model, and has the potential to directly match the data generating process. The middle plot shows how the contribution of irrelevant features to the outcome changes as the non-linearity increases. For the XGBoost model, any irrelevant features are ignored regardless of the degree of non-linearity, and their weights remain at zero (which is in line with the true model). On the other hand, the linear model assigns weight (i.e., the coefficients of the model change) to irrelevant features as the non-linearity increases. This is highly problematic for explainability and interpretability - it results in irrelevant features being indicated to be of predictive importance even when they are not.

2.4.3 The (Un)Interpretability of Models with Complex Functional Form - Camels in the Countryside

In Part 1 we suggested that researchers explore machine learning methods which facilitate the modeling of complex, non-linear relationships between variables. These techniques are applicable to predictive as well as causal approaches. In spite of their flexible functional form, powerful predictive approaches are explainable but not necessarily interpretable. We now describe a famous example which highlights how using powerful function approximation circumvents limitations in functional form does not yield interpretable models. This is one of the principal limitations of purely predictive approaches and closely relates to misspecification (see Part 2). The example involves the classification of images of cows and camels, where images of cows frequently feature countryside backgrounds and images of camels tend to feature sandy or desert regions (Arjovsky et al., 2020). A predictive function will not respect the orthogonality

Figure 2.6: The Uninterpretability of Linear Models in the Presence of Non-Linearity



Note. Demonstrates how the predictive performance of a logistic regressor drops as non-linearity increases, whereas the XGBoost (T. Chen and Guestrin, 2016) model does not (top); shows how irrelevant feature attribution increases with non-linearity for the linear regressor, but for XGBoost it does not (middle); the relationship between variables in the dataset for these experiments becomes increasingly non-linear. These experiments were close adaptations of those by Lundberg, G. Erion, et al. (2020).

and semantics of the animal or background, and the background provides a convenient cue, albeit one which is irrelevant and *confounding*, with which to classify the animal. Hence, a cow in a desert may be wrongly classified as a camel, and a camel with a countryside background may be wrongly classified as a cow. This issue may never become problematic in practice, so long as the function is not exposed to a new distribution of images, where the joint distribution of backgrounds and animals changes. This highlights how predictive models, owing to their misspecification, are sensitive to what is known as covariate or distributional shift. Given a change in the number of photographs of cows in desert regions, or camels in the countryside, the performance of the classifier may suffer considerably.

This example concerning issues relating to classification of high-dimensional image data may appear somewhat unrelated to the typical data that psychologists are concerned with, but actually the problem of confounding is just as important in the low-dimensional setting (Cinelli, Forney, and Pearl, 2022; Rohrer, 2018). Indeed, predictive models are usually fit by minimizing an error criterion (e.g., mean squared error or binary cross entropy), and there is therefore nothing to restrict these models from leveraging any or all statistical correlations present in the data. The use of predictive model explainability techniques (discussed in more detail below) can be used to help identify whether the model might be leveraging factors which have the potential to be confounding, and can provide considerable insight. Unfortunately, if the confounders are latent/unobserved, then it may be very difficult to identify and avoid such problems. Consequentially, predictive models are rarely interpretable.

2.4.4 Limited Functional Form and Misspecification Results in Conflated and Unreliable Interpretations

The examples above highlighted that when the functional form of a model is limited in its capacity to model the relationships between variables, the model coefficients become meaningless and the model is unexplainable. A further problem arises when the model is misspecified for structural reasons. The issues associated with limited functional form and causal misspecification therefore compound to yield model coefficients that are (doubly) uninterpretable. Treating them otherwise would be to interpret these coefficients as being causally meaningful, and this is

an example of conflated and unreliable interpretation. If the functional form of the model were correct (i.e., both the model as well as the relationships between variables were linear), then a linear model would be explainable, but not interpretable. This is because the outcome predicted by the model would indeed be changing according to a β_k change in the input variable $\mathbf{x}_k$, but owing to misspecification, this β_k would still not correspond with any causal quantity. As such, it is only when linear models are neither misspecified due to limited functional form (compared with the true relationship in the data) nor structurally misspecified, that they are interpretable.

2.4.5 Explainability Techniques

The ability to interrogate and explain our predictive models is important, particularly given that the deployment of such models for automated decision making processes have the potential to seriously impact individuals' lives (Hardt, Price, and Srebro, 2016; Kilbertus et al., 2017; Locatello et al., 2019; Y. T. Cao and Daume III, 2019; H. Liu et al., 2019; Howard and Borenstein, 2018; A. Rose, 2010; Louizos, Swersky, et al., 2017; Moyer et al., 2018; Buolamwini and Gebru, 2018). Indeed, the European Union has recently decreed that the use of machine learning algorithms (which includes the use of predictive functions) be undertaken in such a way that any individual affected by an automated decision has the right to an explanation regarding that decision (Aas, Jullum, and Loland, 2019; European Union, 2016). In the previous section we described the camels in the countryside problem, whereby powerful predictive models with flexible functional form do not respect causal structure in the data. However, complex models (often called *black box* models) are more difficult to explain than linear models, and we therefore need explainability techniques to do the explaining for us.

Model explainability is a burgeoning area of machine learning, in which commendable strides have been made in recent years (e.g., Alaa and van der Schaar, 2019; Wachter, Mittelstadt, and Russell, 2018; Lundberg, G. Erion, et al., 2020). The techniques facilitate a form of *meta-modeling*, whereby a simpler, human-interpretable and thereby explainable model is used to represent the more complex, underlying model (Rudin, 2019). One popular explainability technique derives from a game theoretic approach to quantifying the contribution of multiple players in a collaborative game; namely, Shapley values (Shapley, 1953). Recently, Shapley

values have been adapted to yield meaningful explanations of models that correspond well with human intuition (Lundberg and S.-I. Lee, 2017; Lundberg, G.G. Erion, and S.-I. Lee, 2017; Lundberg, G. Erion, et al., 2020; Sundararajan and Najmi, 2020; H. Chen et al., 2020). Indeed, these methods were used with XGBoost in the experiments demonstrating the problems with linear model interpretability above (Figure 2.6). The family of Shapley methods provide breakdowns which indicate how much each input variable or feature contributes to a model's prediction for any individual datapoint. Such individualized prediction and explainability is particularly important for (e.g.) individualized treatment assignments, and thereby mitigates concerns regarding the use of aggregation in psychology and social science (Bolger, Zee, et al., 2019; Fisher, Medaglia, and Jeronimus, 2018). The methods can be used equally for complex functions (such as neural networks) as well as for simple linear functions. By combining powerful function approximation with explainability techniques, we may be able to achieve accurate forecasts and outcome predictions, while maintaining the capacity to understand what our model is actually doing when it makes a prediction.

From a research standpoint, explainability techniques allow researchers to understand, in a purely associational sense, which variables and interactions between variables are important when making a prediction.[12] For example, if one identifies that a variable, previously considered to be important, contributes negligible predictive value then one might investigate whether this variable does or does not fit into a particular theoretical framework. We would therefore argue that researchers should consider a combination of predictive methods with explainability tools as a useful means to contribute new knowledge, particularly during the early and/or exploratory stages of investigation. It is, however, worth emphasizing that just because a predictive model finds a particular feature (ir-)relevant to making a prediction, does not mean that this association is meaningful outside of the function/model (as with camels in the countryside). Furthermore, an explainability technique represents a form of model in its own right, and the process of modeling a model brings its own difficulties (see e.g., Rudin, 2019; Kumar et al., 2020). Indeed, if the explanation model is good at explaining the data in a simple, human-readable form, then the explanation model provides evidence that a simpler, more explainable model was possible to

[12] We avoid the term 'correlational' on the basis of our earlier discussion - correlations do not describe dependence well when the functional form is non-linear.

begin with. These difficulties notwithstanding, the explainability techniques provide a valuable means to leverage predictive model for exploratory research.

2.4.6 Summary

In Part 3, we have described how either limited functional form, or model misspecification, or both, result in uninterpretable models. In such cases, any attempt to interpret the models in spite of these limitations results in conflation and unreliability. The interpretations are conflated because a misspecified model cannot be interpreted causally, and they are unreliable because predictive models can only be explained. This distinction is important because, if a misspecification has occurred (perhaps because we intentionally adopted a predictive/non-causal approach), one can restrict the purview of scientific conclusions to the specific mathematics of the algorithm used for prediction. In other words, powerful function approximation techniques may be able to accurately predict outcomes and have the flexibility to match the functional form of the true data distribution, but they do not necessarily respect or reflect the *causal* structure in the data generating process. Does this mean that predictive techniques cannot generate understanding? Not entirely. There are many scenarios, particularly during the exploratory stages of a research project, for which researchers may not yet have a strong, empirically supported inductive bias or theory about the data generating process. Rather than testing specific theoretical hypotheses during these early stages, it may be pertinent to ask more general research questions. The goal may then be to amass varied evidence (e.g., by using predictive models) to gradually uncover a basis for the development of an increasingly refined theory (Gelman, 2014; Shrout and Rodgers, 2018; Oberauer and Lewandowsky, 2019; Tong, 2019). Of course, researchers should be transparent about whether this is their goal, and carefully consider how they interpret predictive models. Model explainability techniques may be useful in building up an intuition about 'what is important' in the phenomenon of interest. However, these techniques are not without their own limitations, and we urge researchers to engage broadly with experts in the practice of these techniques to ensure that (a) their approaches are optimal for their research, and (b) that their interpretations (or explanations) are tempered according to the limitations of their models.

2.5 Part 4: Discussion and Recommendations

2.5.1 Modeling in Practice

Flexible predictive modeling approaches appear to be used rarely in psychology and social science, indicating a missed opportunity in these fields.[13] Predictive modeling may be extremely useful, particularly as part of the research exploration stage (Yarkoni and Westfall, 2017). When combined with model explainability techniques (such as those deriving from Shapley values), predictive methods provide a powerful way to interrogate associations present in the data. So long as practitioners recognize the limitations and are transparent about their approach and any associated assumptions, conclusions can still be drawn from predictive models, provided that they are not presented as causal conclusions.

We would argue that, in general, undertaking meaningful causal inference is extremely challenging, and significantly more so than fitting predictive functions to data. Indeed, the former should subsume the latter as part of a causal pipeline for (a) mitigating issues with limited functional form by using (e.g.) data-adaptive function approximation to model the *functional* relationships between variables, and (b) mitigating issues with model misspecification by carefully considering *causal* relationships between variables. As described earlier, researchers have noted the ambiguity in the use of implicit causal (rather than predictive) language even in studies which otherwise appear to be predictive (Grosz, Rohrer, and Thoemmes, 2020; M. Hernan, 2018a). It has been suggested that this reluctance to be explicitly causal stems from a strange history of discouragement for its use in observational studies (Grosz, Rohrer, and Thoemmes, 2020; Dowd, 2011).

In terms of understanding, our view is that, in general, researchers in psychology and social science lack some competence in the practice of prediction and causal inference. If researchers were more competent at prediction, they would avoid interpreting linear model parameters using implicit causal language (Grosz, Rohrer, and Thoemmes, 2020), avoid using naive linear models to test causal hypotheses derived from causal theories, and instead be using varied and flexible

[13] For an example of researchers in psychology using machine learning techniques see Joel, Eastwick, Allison, et al. (2020).

function approximation techniques, model explainability tools, and train/test data splitting and/or cross-validation techniques (Yarkoni and Westfall, 2017). On the other hand, if researchers were more competent in causal inference, they would be less ambitious about specifying and interpreting large (causal) SEM graphs, more restrained when it comes to interpreting the coefficients of misspecified models, more transparent about assumptions when defining explanatory models (Grosz, Rohrer, and Thoemmes, 2020), use more explicitly causal language and terminology (Grosz, Rohrer, and Thoemmes, 2020), more clearly distill and identify the specifics of their causal questions or hypotheses, and be less likely to worsen the bias and generalizability of their inferences by adopting *ad hoc*, data driven variable model manipulation techniques during the analysis stage. Finally, if researchers had a clearer understanding about the differences between predictive and causal approaches, then we would also see more delineation between the two. Typical practice therefore involves a combination of unreliable interpretations regarding models with limited functional form and causal misspecification.

2.5.2 Recommendations

1. We recommend that psychologists and social scientists give more consideration to predictive approaches, particularly during the exploratory stages of a research project.

The inherent complexity and non-linearity of the typical phenomena of interest to psychologists and social scientists may make the goal of causal inference arbitrarily complex (Meehl, 1990). This may partly explain why researchers in psychology and social science are generally discouraged from drawing causal conclusions from observational data, despite them doing so implicitly anyway (Grosz, Rohrer, and Thoemmes, 2020; Dowd, 2011). Indeed, the use of SEM could be taken as evidence of an explicit intention to undertake causal research, as the very structure of the model is an imposition of the researcher's view on the data generating process. The use of an explicit causal graph with opaque predictive interpretations represents a further example of the conflation of predictive and causal approaches. In cases where the models themselves are misspecified both in terms of linear functional form and untestable structural assumptions, the interpretation of such models becomes unreliable.

When researchers wish to model the relationships between variables, either as part of a causal

model, or for purposes of prediction, then it may be extremely advantageous for them to consider techniques common in machine learning, particularly in combination with model explainability techniques. Indeed, Yarkoni and Westfall (2017) have previously made a similar recommendation. Powerful function approximation techniques including feature engineering or data-adaptive techniques such as neural networks or random forests, can be used to leverage as many associations present in the data sample as possible. In the case of predictive modeling, a consideration for the causal structure of the data is possible but not necessary. Incorporating causal inductive bias may aid in generalization, but it is not strictly necessary to achieve good predictive performance. Unfortunately, the use of techniques with potentially data-adaptive, flexible functional form is extremely rare in psychology and social science, where the use of models with restrictive linear functional form is ubiquitous (Yarkoni and Westfall, 2017; Blanca, Alarcon, and Bono, 2018).

2. We recommend that psychologists and social scientists seek collaboration with statisticians and machine learning engineers/researchers, whose principal focus is to understand, practice, and develop function approximation and causal inference techniques. Given that there exist entire fields dedicated to the study of relevant modeling approaches (e.g., statistics, machine learning, causal inference), independently of the empirical human sciences, it is perhaps unrealistic to expect an expert in, say, psychology or social science, to have equal expertise in the practice of predictive and explanatory modeling, particularly when the mathematical knowledge required to understand these techniques is both significant and rare in these fields (Boker and Wenger, 2007). Furthermore, new methods are continually developed and updated in the fields of statistics and machine learning. As well as encouraging researchers to make themselves more familiar with the topics of predictive and causal modeling, we also recommend they seek collaboration with experts in the practice of their chosen analytical approach. Note that this recommendation has been made by researchers previously in various contexts (e.g., (Lakens, Hilgard, and Staaks, 2016)).

3. We recommend researchers be transparent about whether they are adopting a predictive or causal approach and to qualify their interpretations. We have discussed how unreliable interpretations may stem from issues of limited functional form and causal misspecification, and

how these issues may be common in the fields of psychology and social science. We encourage researchers to ask themselves what an interpretation of an effect size or parameter derived using a naive (i.e., misspecified) model actually means: Is it actually an explanation for how much the output of the *model* changes with respect to a change in the input; or is it being interpreted causally (e.g., this childhood intervention increased well-being by θ-amount)? In either case, researchers need to be transparent and clearly articulate whether they are adopting a predictive or causal approach. Each approach is associated with assumptions and limitations which need to be clearly stated in order to contextualize any explanations or interpretations which are made. Predictive model explainability tools have their own limitations and may actually contradict the results from undertaking causal inference: While the inclusion of a mediator in a regression can completely block a causal path reducing the estimated effect to zero, a strong effect might be indicated by an explanation of a predictive model. Similarly to Grosz, Rohrer, and Thoemmes (2020), we therefore recommend that researchers clearly state their approach as well as its associated assumptions and limitations, and moderate their explanations, interpretations, and conclusions accordingly.

4. We recommend that researchers distill their research questions and hypotheses. It may be pertinent for researchers to attempt to distill and simplify causal questions so that they are both minimal and sufficient. For example, in our discussion of causal inference, we chose to identify a single causal effect, and for this it was sufficient to identify the minimal backdoor adjustment set necessary to render this causal effect identifiable. As such, a full graph did not need to be specified, even though it may need to be considered in order to find the backdoor adjustment variables. M. J. van der Laan and S. Rose (2011) recommend a similar "targeted" approach. More generally, by distilling our research questions and hypotheses, we may be able to increase the chance that our modeling attempts are successful, and that we have realistic expectations of the level of understanding that can be achieved with some acceptable level of confidence. This recommendation therefore overlaps with the recommendation for transparency in so far as distilling a research question or hypothesis will make it easier to be transparent.

2.6 Conclusion

The replicability crisis has drawn attention to numerous weaknesses in typical psychology and social science research practice. However, in our view, issues relating to limited functional form, model misspecification, and unreliable interpretations have not been sufficiently addressed in prior meta-research. Indeed, while it is difficult, if not impossible, to quantitatively apportion the crisis according to its myriad causes, in our view the issues covered in this work represent significant contributing factors.

In this paper, we demonstrated the nature of these problems and how they manifest in typical psychology and social science research. The typical models used in psychology and social science are limited in their functional form and misspecified in terms of causal structure. The result is that subsequent interpretations conflate predictive and causal language and are also unreliable. We make four recommendations for researchers in these fields to update and improve their research practice by (1) giving more consideration to the use of flexible and varied predictive modeling and model explainability techniques; (2) to seek collaboration with experts from the fields of statistics and machine learning; (3) to be transparent about whether they are adopting a predictive or causal approach; and (4), to distill and simplify their research questions and hypotheses in order to increase the chances that these questions and hypotheses can be practically addressed and tested. While we have focused on the fields of psychology and social science, we believe the highlighted issues are relevant to all empirical human sciences fields. There is little doubt in our minds that the lack of understanding about progress, assumptions, limitations, and pitfalls associated with predictive and explanatory modeling has contributed to the replicability crisis, and we implore researchers to address these shortfalls, lest they hinder scientific progress. Every research question and hypothesis may present its own unique challenges, and it is only through an awareness and understanding of varied statistical methods for predictive and causal modeling, that researchers will have the tools with which to appropriately answer and test them.

CHAPTER 3

Application of Machine Learning and Explainability Techniques

In Chapter 2, I recommended that researchers engage with machine learning and machine learning explainability methods for exploring data. Firstly, note that I will discuss the limitations and risks associated with this approach in more detail in Chapter 6. Secondly, this Chapter contains an example of such an application for the prediction of perceived partner support from relational and individual variables, and is drawn from the following publication:

Vowels, L.M., **Vowels, M.J.**, Carnelley, K.B., Kumashiro, M., 2022. A machine learning approach to predicting perceived partner support from relational and individual variables. *Social Psychological and Personality Science*, DOI: 10.1177/19485506221114982.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

Abstract: Perceiving one's partner as supportive is considered essential for relationships, but we know little about which factors are central to predicting perceived partner support. Traditional statistical techniques are ill-equipped to compare a large number of potential predictor variables and cannot answer this question. The current research used machine learning analysis (random forest with Shapley values) to identify the most salient self-report predictors of perceived partner support cross-sectionally and six months later. We analyzed data from five dyadic datasets (N = 550 couples) enabling us to have greater confidence in the findings and ensure

generalizability. Our novel results advance the literature by showing that relationship variables and attachment avoidance are central to perceived partner support while partner similarity, other individual differences, individual well-being, and demographics explain little variance in perceiving partners as supportive. The findings are crucial in constraining and further developing our theories on perceived partner support.

3.1 Introduction

Perceiving one's partner as supportive is considered an essential element in romantic relationships, but we lack knowledge about which factors are central to predicting such perceptions. Several relationship theories (e.g., attachment theory, self-determination theory, interdependence theory) have underscored the centrality of partner support in promoting well-functioning relationships. Existing research has examined several potential factors that are considered important for perceived partner support, but it has not compared the relative importance of these different factors, in part because traditional statistical analyses are not well-equipped to examine a large number of potential predictors at once. The purpose of the present study was to leverage the power of machine learning to compare which theoretically relevant relational and individual variables—from the perspectives of both the support receiver and the support provider—predict the most variance in perceived partner support.

3.1.1 Established Relational Predictors of Perceived Partner Support

According to attachment and interdependence theories, actors should perceive partners as more supportive when the relationship is characterized by high satisfaction, empathy, commitment, trust, and willingness to sacrifice, and low conflict (B. Feeney and Collins, 2015; Kelley and Thibaut, 1978; Mikulincer and Shaver, 2009; C.E. Rusbult and Van Lange, 2003; Ryan and Deci, 2000). This is because partners in these relationships can count on each other to provide support and are thus more open to support when needed or may be more willing to take risks (C.E. Rusbult and Van Lange, 2003). This in turn leads the recipients to perceive their partners as supportive. Furthermore, the transactive goal dynamics theory suggests that high goal

correspondence allows partners to better coordinate their efforts to achieve their goals and thus are likely to be more supportive (Fitzsimons and E.J. Finkel, 2018). Finally, self-expansion theory (A. Aron, E. Aron, et al., 1991) suggests that inclusion of other in the self enables greater shared intimacy, in turn leading partners to share resources and to perceive each other as more supportive. Based on these theories, we would expect relationship variables (see Table 3.1 for the full list of variables) to be important for perceiving partners as supportive, but it is not clear whether there are specific relational variables that contribute to perceptions of support more than others.

3.1.2 Established Individual Predictors of Perceived Partner Support

Interestingly, few theories on partner support have explicitly discussed which individual differences variables are the most likely to explain why some partners perceive and are perceived as more supportive than others (see attachment theory for an exception; Mikulincer and Shaver, 2009). Attachment theory suggests that avoidantly attached individuals perceive partners as less supportive because they doubt partners' availability (A. Martin, Paetzold, and Rholes, 2010) whereas anxiously attached individuals doubt their worthiness of being supported but feel others as capable of providing support, which has resulted in mixed findings for attachment-anxiety (B. Feeney, 2004; B. Feeney and Thrush, 2010; Jakubiak and B. Feeney, 2016; A. Martin, Paetzold, and Rholes, 2010). According to attachment theory, individuals who trust themselves are also more likely to trust other's capacity to be supportive when needed (Collins and B. Feeney, 2004) and thus are more likely to perceive their partners as supportive. Thus, we expect that individuals higher in promotion orientation (i.e., regulatory focus on dreams and aspirations; Righetti and Kumashiro, 2012, self-control; Zuo et al., 2020, and self-efficacy; L. Vowels, Carnelley, and Francois-Walcott, 2021; L. Vowels and Carnelley, 2022) feel a greater sense of autonomy (self-determination theory; Ryan and Deci, 2000) and trust in their ability to achieve their goals (attachment theory; Collins and B. Feeney, 2004). Furthermore, we expect that individuals high in self-esteem (Harris and Orth, 2020) or self-respect (Kumashiro, E.J. Finkel, and C.E. Rusbult, 2002) would perceive their partners as more supportive as they may self-select into healthier relationships or be able to elicit higher quality support from their partners.

Additionally, while better physical (Reblin and B.N. Uchino, 2008) and emotional well-being (Canevello and Crocker, 2010; Drigotas, 2002) have often been considered as outcomes of perceived partner support, it is also likely that individuals with higher well-being are more easily supported. For example, depression makes people more pessimistic and view everything in a negative light (Anzalidi and Shifren, 2019). Thus, we expect that people who have higher well-being are more optimistic in their perceptions of partner behaviors and act in ways that tend to elicit positive behaviors from their partners.

Finally, demographic variables such as relationship length, age, and gender have previously been associated with perceived partner support, with mixed results (Bühler et al., 2019; Jakubiak, B. Feeney, and Ferrer, 2020; Verhofstadt, Buysse, and Ickes, 2007). Several researchers have hypothesized that support for goals is likely more important in early stages of the relationship with the importance of support declining over time (Bühler et al., 2019; Jakubiak, B. Feeney, and Ferrer, 2020; Verhofstadt, Buysse, and Ickes, 2007) while other researchers have found that longer relationship length predicted higher perceived partner support (Lantagne and Furman, 2017). Furthermore, because women are traditionally socialized to be more caring, partners may find women more supportive. Indeed, previous research has found women to be perceived as more supportive, but both men and women felt equally supported by their spouse (Verhofstadt, Buysse, and Ickes, 2007). There is no prior literature on education, ethnicity, or children on perceived partner support, but we have provided a rationale for their inclusion in Table S1 in the supplementary (accessible here: `https://osf.io/bjvhu` and provided at the end of this chapter).

3.1.3 Machine Learning

Previous research has relied exclusively on traditional linear models (Breiman, 2001b; Lundberg, G. Erion, et al., 2020; Luque-Fernandez et al., 2018; Orben and Przybylski, 2019; J. Peters, Janzing, and Scholkopf, 2017). Machine learning algorithms have several key advantages over these models: they can learn highly non-linear relationships between variables, handle a large number of predictors at once, and estimate complex interactions between different variables. As such, they are not susceptible to problems of multicollinearity or limited functional form

(e.g., expecting an association to be linear while the real relationship is cubic) misspecification (M. J. Vowels, 2021). Because of this, using machine learning provides a much more flexible and powerful approach to predicting an outcome. Machine learning algorithms are traditionally fed as many predictors as possible to maximize prediction. It then learns which variables are important for predicting the outcome. In the present study, we use a random forest algorithm (Breiman, 2001a), which is a form of explainable decision tree that can handle highly non-linear relationships and complex interactions without overfitting to the data.

Machine learning models, including random forests, have traditionally been "black box models" where the researcher is unable to understand what the algorithm has used for predicting the outcome. However, recent developments in machine learning have provided tools that allow interpretation of the results through explanations of machine learning models (Lundberg and S.-I. Lee, 2017; Lundberg, G.G. Erion, and S.-I. Lee, 2017). This work is particularly interesting because it enables researchers to combine the use of powerful machine learning algorithms and state-of-the-art tools for model explainability that can provide accurate predictions as well as increase our understanding of which factors are the most important in predicting the model outcome. The latter is of particular importance because one of the principal aims of psychology is to develop a deeper understanding of human behavior (Grosz, Rohrer, and Thoemmes, 2020). In the present study, we take advantage of this new development in machine learning by using Shapley values (Lundberg and S.-I. Lee, 2017; Lundberg, G.G. Erion, and S.-I. Lee, 2017) to estimate the effect size and direction of the effect of each variable predicting perceived partner support.

3.1.4 The Current Research

Our aim was to examine which relational and individual factors are the most predictive of perceived partner support. We examined two types of perceived partner support (B. Feeney and Collins, 2015): perceived partner responsiveness (i.e., being available and responsive to the partner's needs, and understanding and validating one's overall self; Reis, M. Clark, and Holmes, 2004) and perceived affirmation of the ideal self (i.e., perceiving and behaving in a manner consistent with the partner's ideal self; Drigotas et al., 1999). The former is a broader construct

and is considered one possible central organizing theme for the diverse phenomena relationship scientists study (Reis, 2007), whereas the latter is more specific and focused explicitly on partner's role in helping individuals become closer to their ideal self (Drigotas et al., 1999). As such, although both are frequently used to examine partner support in romantic relationships, they may be predicted by different factors due to affirmation being more specifically about the ideal self.

The predictor variable selection for the present study was guided by existing theoretical frameworks to test the explanatory power of different relational and individual variables (see Table 3.1 for the variables, expected direction of the effect, and state of the current evidence). The selection was somewhat limited by the availability of variables across the datasets. Furthermore, because there are (at least) two people in romantic relationships, it is important to understand whether one person's outcome is only determined by their own variables (actor effects) or whether their partner's reports also predict the actor's outcomes (partner effects). Our hope is to add to the current understanding of the factors that are the most and least likely to predict perceived support. We used data from five dyadic datasets that had a large number of common predictor variables and addressed the following research questions: 1) How much variance in the overall outcomes can we explain? 2) Are relational or individual variables more important for predicting partner support? 3) Do partner effects explain additional variance in outcomes above actor effects? And 4) Can we predict support over time?

3.2 Method

3.2.1 Participants and Procedure

The preregistration and materials for the project can be found on the Open Science Framework `https://osf.io/v368c/` and provided at the end of this chapter.[1]

Five dyadic datasets (E. Finkel, 2020a; E. Finkel, 2020b; C.E. Rusbult, Kumashiro, Coolsen, et al., 2019; C.E. Rusbult, Kumashiro, E.J. Finkel, et al., 2019) were combined in this project

[1]We also preregistered analyses for self-efficacy but due to the journal word limit have not included them in the main manuscript. The results can be found in the Supplementary Material (accessible here: `https://osf.io/bjvhu` and provided at the end of this chapter). We also added longitudinal analyses to the manuscript.

to create a large dataset of couples. These datasets were chosen because they included a large number of predictor variables that were the same across the samples. We are aware of no other datasets with such high overlap in the variables. All datasets included cross-sectional self-reported data collected from both dyad members in romantic relationships. Two of the datasets included only dating couples ($n_1 = 74$, $n_4 = 92$), one dataset included newly committed couples (e.g., engaged, married, moving in together; $n_3 = 178$), and two datasets included married couples ($n_2 = 120$, $n_5 = 77$). The final sample consisted of 550 couples (1,100 individuals). Dataset 3 was also used to predict support six months later and included 161 couples.

On average, participants were 28.32 years old (SD = 10.90, range = 18-79) and had been in a relationship for 5.59 years (SD = 8.13, range = 0.08 – 61.50). Most of the participants were white ($n = 876$, 80%) with a minority being African American ($n = 83$, 8%), Hispanic ($n = 35$, 3%), or Asian ($n = 72$, 7%). The sample was primarily well-educated: 196 (18%) participants had a graduate degree (M.S./PhD), 466 (42%) a Bachelor's degree, 379 (34%) at least some college, and 60 (5%) had no college courses. The couples were either married ($n = 266$, 48%), cohabiting ($n = 127$, 23%), or dating and not living with each other ($n = 220$, 40%), and most of the couples did not have any children ($n = 462$, 84%). All data were collected in the United States.

Measures

The outcome variable, perceived partner support was measured using the 18-item responsiveness scale (Reis, M. Clark, and Holmes, 2004) in four datasets and the partner affirmation scale (Drigotas et al., 1999) in three datasets. The rest of the variables from each dataset were included in the final dataset as predictors if the variable appeared in at least three of the five datasets. These variables were divided into actor's and partner's individual (17) and relational (11) predictors (summarized in Table 3.1; see supplemental material for the description of the scales used, accessible here: `https://osf.io/bjvhu` and provided at the end of this chapter).

Table 3.1: The List of Included Variables with a Theoretical Rationale for Inclusion.

Variable	Expected Direction	Relevant Studies	Prior Evidence	Important Predictor*
Relational Variables				
Core relationship variables				
trust	Positive	B. Feeney and Collins (2015)	Yes	Yes
commitment	Positive	Kelley and Thibaut (1978)		Yes
empathy toward partner	Positive	Mikulincer and Shaver (2009)		Yes
conflict	Negative	C.E. Rusbult and Van Lange (2003)		Yes
satisfaction	Positive	Ryan and Deci (2000)		Only baseline
willingness to sacrifice	Unclear			No
Partner similarity				
goal correspondence	Positive	Gere and Schimmack (2013) L. Vowels, Carnelley, and Francois-Walcott (2021)	Yes	No
actual inclusion of the other in self	Positive	A. Aron, E. Aron, et al. (1991) A. Aron and B. Fraley (1999)	None	No
Individual Variables				
Attachment theory[a]				
attachment avoidance	Negative	A. Martin, Paetzold, and Rholes (2010) B. Feeney and Thrush (2010) B. Feeney (2004)	Yes	Yes
attachment anxiety	Negative	Jakubiak and B. Feeney (2016)	Mixed	Only long. affirmation
Individual differences				
self-control	Positive	Zuo et al. (2020); only relationship satisfaction)	None	No
regulatory focus (promotion)	Positive	Righetti, C. Rusbult, and Finkenauer (2010)	Yes	No
regulatory focus (prevention)	No association	Righetti, C. Rusbult, and Finkenauer (2010)	Yes	No
self-efficacy	Positive	L. Vowels, Carnelley, and Francois-Walcott (2021)		No
self-esteem	Positive	Harris and Orth (2020)	Yes	No
self-respect	Positive	Kumashiro, E.J. Finkel, and C.E. Rusbult (2002); pro-relationship beh. only	None	No
Individual well-being				
physical health	Positive	Reblin and B.N. Uchino (2008)	Yes (as an outcome)	Not consistently
life satisfaction	Positive	Drigotas (2002)	Yes (as an outcome)	Affirmation + responsiveness long.
depression	Negative	Canevello and Crocker (2010)	Yes	No
Demographic variables				
relationship status	Unclear	Bühler et al. (2019)	Mixed	No
relationship length	Unclear	Jakubiak, B. Feeney, and Ferrer (2020)		No
gender	Unclear	Verhofstadt, Buysse, and Ickes (2007)	Mixed	No
age	Unclear	Bühler et al. (2019) Jakubiak, B. Feeney, and Ferrer (2020)	Mixed	No
ethnicity	Unclear	None	None	No
education	Unclear	None	None	No
children	Unclear	None	None	No

Note. For further details on the theoretical justifications, please see Table S1 in the supplemental file (accessible here: `https://osf.io/bjvhu` and provided at the end of this chapter). Some of the variables were not present in all analyses due to them not being included in all datasets. All variables were present at least in one analysis for each outcome.

[a]. Because attachment styles are the only individual differences variables that have been linked to perceived partner support theoretically, we chose to include them in a separate category.

* Summary of the findings across the analyses: predictors were considered important if they explained at least 5% of the variance in the model performance.

'long.' = longitudinal, 'beh.' = behaviours.

Table 3.2: The Overall Prediction Results for Each Outcome Variable for Individual and Relational Variables and Models with Actor Effects Only and with Actor and Partner Effects.

	couples	**% Variance**	**MSE**	R^2	**Individual**	**Relational**
Outcome	n	M (SE)	M (SE)	M (SE)	$\%_a$ / $\%_p$	$\%_a$ / $\%_p$
Responsiveness						
Model 1	473	50.4 (0.03)	0.48 (0.03)	.50 (0.03)	42.9	57.1
+ Partner		50.1 (0.02)	0.48 (0.03)	.50 (0.02)	32.3 / 13.4	51.1 / 3.2
Model 2*	382	55.3 (0.02)	0.47 (0.04)	.54 (0.02)	35.7	64.3
+ Partner*		54.8 (0.02)	0.48 (0.03)	.54 (0.02)	26.6 / 11.9	57.4 / 4.0
Model 3	353	48.2 (0.03)	0.38 (0.03)	.47 (0.03)	30.8	69.2
+ Partner		48.1 (0.02)	0.35 (0.03)	.47 (0.03)	22.9 / 11.6	60.0 / 5.5
Longitudinal	161	27.6 (0.06)	0.34 (0.02)	.25 (0.06)	49.4	50.6
+ Partner		26.7 (0.05)	0.34 (0.03)	.24 (0.06)	27.1 / 13.4	33.3 / 0.7
Affirmation						
Model 1*	356	34.5 (0.04)	1.16 (0.06)	.34 (0.05)	48.2	51.8
+ Partner*		35.4 (0.05)	1.13 (0.07)	.36 (0.04)	31.3 / 22.3	40.8 / 4.9
Longitudinal	161	18.2 (0.07)	1.26 (0.13)	.15 (0.07)	51.1	48.9
+ Partner		16.3 (0.06)	1.29 (0.13)	.13 (0.07)	34.7 / 16.2	37.7 / 11.4

Note. MSE = mean squared error, M = mean, SE = standard error. $\%_a$ refers to the percentage of variance explained by actor variables, $\%_a$ refers to the percentage of variance explained by partner variables. The first model for each outcome variable included as many samples as possible and subsequent models included as many variables as possible. The full list of excluded variables and samples can be found on the OSF project page as well as at the end of this chapter.
* Results presented in figures.

Table 3.3: The Impact of All Variables of the Most Predictive Models for Responsiveness and Affirmation.

Responsiveness				**Affirmation**			
Cross-sectional		**Longitudinal**		**Cross-sectional**		**Longitudinal**	
Variable	*Importances*	*Variable*	*Importances*	*Variable*	*Importances*	*Variable*	*Importances*
relationship satisfaction	**0.26**	attachment avoidance	**0.17**	trust	**0.22**	conflict	**0.17**
empathy toward partner	**0.11**	trust	**0.15**	life satisfaction	**0.15**	attachment anxiety	**0.1**
physical health	**0.1**	conflict	**0.14**	relationship satisfaction	**0.14**	attachment avoidance	**0.09**
conflict	**0.09**	empathy toward partner	**0.07**	commitment	**0.06**	trust	**0.08**
attachment avoidance	**0.08**	life satisfaction	**0.05**	self-efficacy	**0.06**	life satisfaction	**0.05**
trust	**0.05**	commitment	0.04	attachment avoidance	**0.05**	commitment	**0.05**
age	0.03	self-esteem	0.04	depression	0.03	empathy toward partner	**0.05**
promotion orientation	0.03	relationship length	0.03	empathy toward partner	0.03	relationship satisfaction	0.04
commitment	0.03	promotion orientation	0.03	physical health	0.03	promotion	0.04
self-esteem	0.03	goal correspondence	0.03	impression mngmnt	0.02	relationship length	0.03
relationship length	0.02	self-control	0.03	attachment anxiety	0.02	self-esteem	0.03
IOS	0.02	competence	0.02	age	0.02	prevention orientation	0.02
goal correspondence	0.02	prevention orientation	0.02	relationship length	0.02	goal correspondence	0.02
self-control	0.02	self-respect	0.02	self-esteem	0.02	sacrifice	0.02
self-respect	0.01	relationship satisfaction	0.02	self-control	0.02	age	0.02
attachment anxiety	0.01	physical health	0.01	self-respect	0.01	health	0.02
subjective well-being	0.01	age	0.01	married	0.01	depression	0.02
social desirability	0.01	attachment anxiety	0.01	some college	0.01	autonomy	0.02
prevention orientation	0.01	sacrifice	0.01	social desirability	0.01	relatedness	0.02
self-efficacy	0.01	impression mngmnt	0.01	Bachelors	0.01	IOS	0.02
children	0.01	IOS	0.01	IOS	0.01	self-control	0.02
sacrifice	0.01	social desirability	0.01	dating	0.01	competence	0.01
depression	0.01	self-efficacy	0.01	gender	0	impression mngmnt	0.01
impression mngmnt	0.01	depression	0.01	children	0	self-respect	0.01
gender	0.01	relatedness	0.01	graduate	0	self-efficacy	0.01
married	0.01	married	0.01	Hispanic	0	social desirability	0.01
graduate	0.01	cohabiting	0.01	cohabiting	0	married	0.01
Black	0	autonomy	0.01	Black	0	gender	0.01
White	0	Bachelors	0	White	0	graduate	0
Bachelors	0	dating	0	no college	0	Hispanic	0
cohabiting	0	White	0	Asian	0	Bachelors	0
some college	0	Hispanic	0			cohabiting	0
dating	0	gender	0			dating	0
no college	0	some college	0			some college	0
Hispanic	0	no college	0			White	0
Asian	0	graduate	0			no college	0
		children	0			Asian	0
		Black	0			Black	0
		Asian	0			children	0

Note. Model importances have been normalized to represent percentage change on the model making the effect sizes more interpretable. Variables that had at least 0.05 impact on the model are bolded. IOS = inclusion of other in the self; impression mngmnt = impression management; autonomy, relatedness, and competence are the needs based on self-determination theory.

Data Analysis

Details of the data preparation and analyses can be found in the supplemental material (accessible here: `https://osf.io/bjvhu` and provided at the end of this chapter). The results were analyzed using Python 3.7. Each dataset was analyzed using a random forest regressor (Breiman, 2001a). A random forest is a type of decision tree that trains on bootstrapped sub-samples of the data to avoid overfitting. We used the default "scikit learn" random forest regressor with ten-fold cross-validation (Pedregosa et al., 2011). The metrics for test data model performance used were the mean-squared error (which is the averaged squared difference between the prediction and the observed value), the R^2, and the variance explained. The full last model trained was saved and explained using the "SHapley Additive exPlanations" package (SHAP) (Lundberg and S.-I. Lee, 2017; Lundberg, G.G. Erion, and S.-I. Lee, 2017; Lundberg, G. Erion, et al., 2020). The results are provided as feature importances, which describe how important the variable is for the model outcome and how much it changes the outcome.

The analyses were conducted separately by first including as many participants as possible in each analysis and then by including as many variables as possible. This resulted in a total of four analyses (three for perceived partner responsiveness, one for affirmation) which were conducted twice: once including only actor effects and once including both actor and partner effects. The included variables and the results for all analyses can be found on the OSF project page as well as at the end of this chapter. Random forests in their current form are unable to explicitly model hierarchies in the data and it is possible that hierarchical data can inflate the predictive performance. However, given we were primarily interested in the relative performance of different predictors, which is not affected by hierarchical data, this is less of an issue in the current study.

3.3 Results

3.3.1 Total Variance Explained (Research Questions 1-3)

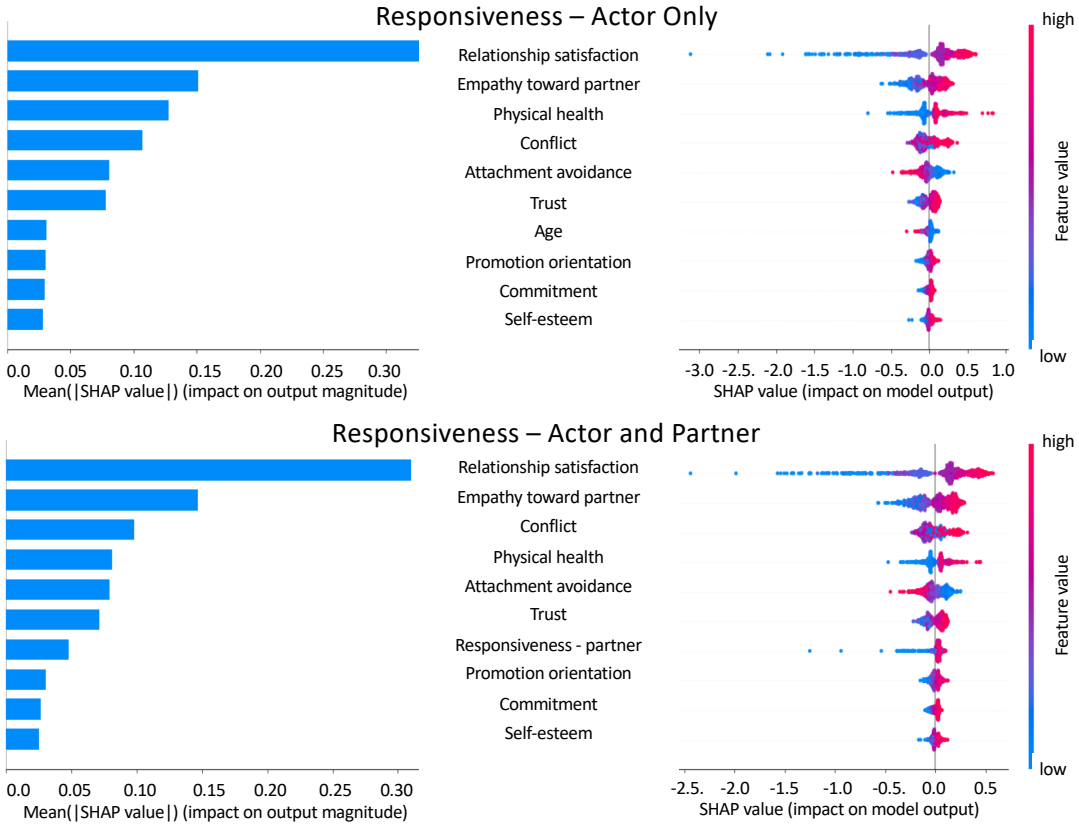
Table 3.2 presents the overall prediction results for each outcome variable for each model for relational and individual variables as well as for models including actor effects only and for models including both actor and partner effects. In the actor only models, we were able to explain the most variance in perceived responsiveness overall (48.2% – 55.3%) with relational variables generally predicting the largest percentage of the variance (57.1% - 69.2%). Individual variables predicted a total of between 30.8% – 42.9% of the variance. Partner effects did not improve the predictive power of the models; if anything, partner effects contributed noise to the data and made the prediction less accurate. However, in the models with partner effects included, partners' individual variables predicted between 11.6% and 13.4% of the variance. In contrast, partners' relational variables predicted very little variance (3.2% - 5.5%).

For perceived affirmation, the model with actor effects was able to predict 34.5% of the variance with relational and individual variables predicting similar amounts of variance (48.2% and 51.8%, respectively). In the models with both actor and partner effects, actors' relational variables predicted the most variance (40.8%) followed by actors' individual variables (31.3%). Partners' individual variables contributed 22.3% of the variance, whereas partners' relational variables contributed very little (4.9%).

3.3.2 Most Predictive Variables (Research Question 4)

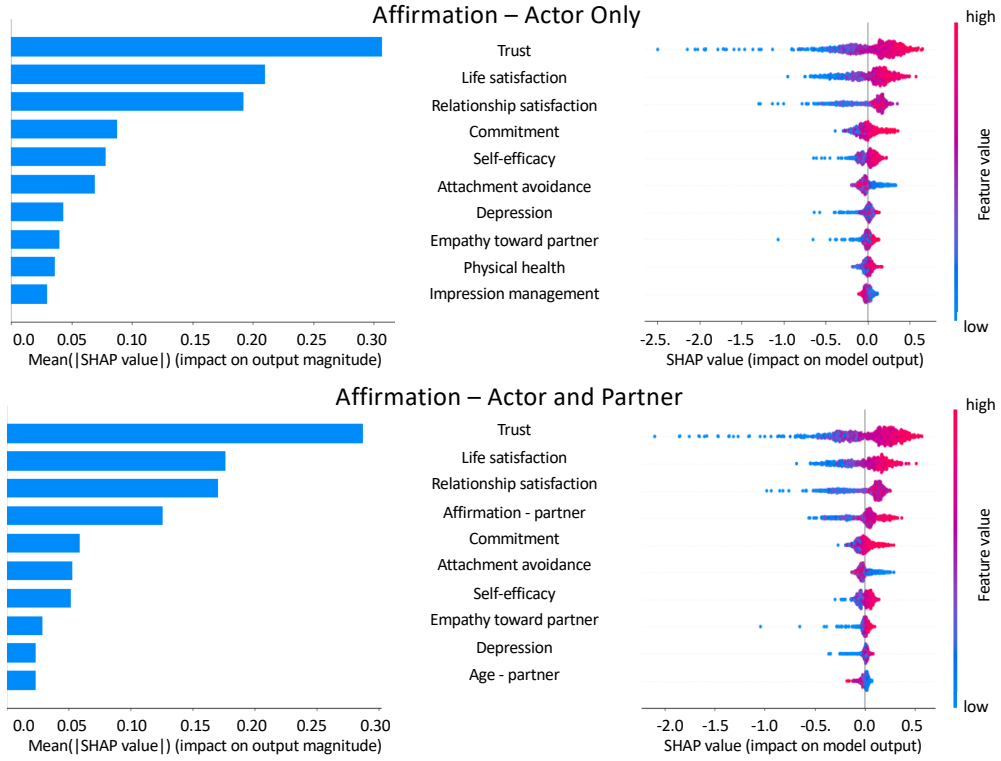
In most of the models, the predictive importance of the variables decreased after only a small number of predictors. The rest of the predictors contributed only a small amount of variance into the model individually. We used 5% as a cutoff for percentage change in the model. We present the top-10 variables for each outcome in the figures and all predictors in Table 3.3 for the percentage model change for the main actor models. In the figures, the left side provides the average effect of each variable on the model outcome. The right side of the figure provides the estimates for each individual participant. Red indicates a higher value of the predictor

Figure 3.1: The Top-10 Most Important Predictors for Responsiveness for Models with Actor Effects and Actor and Partner Effects.



Note. The figure presents the results from the most predictive model.

Figure 3.2: The Top-10 Most Important Predictors for Affirmation for Models with Actor Effects and Actor and Partner Effects.



Note. The figure presents the results from the most predictive model.

variable and blue indicates a lower value. For example, red is equal to 1 and blue is equal to 0 for binary variables. The Shapley values are additive and can be interpreted similarly to an average effect from a linear model. For example, one unit increase in relationship satisfaction predicted a corresponding average increase of 0.33 units in perceived responsiveness. The individual effects show that low relationship satisfaction predicted up to a -3.0-unit change in perceived responsiveness compared to average relationship satisfaction whereas a high relationship satisfaction score predicted up to 0.5-unit increase in perceived responsiveness compared to average relationship satisfaction. In Table 3.3, the impact is rescaled to be between 0 and 1 for ease of interpreting and comparing effect sizes.

Perceived partner support was measured using two variables: perceived responsiveness and affirmation. There were four relational (relationship satisfaction, empathy toward partner, trust, and commitment) predictors that were consistently predictive of higher levels of perceived responsiveness (see Figure 3.1) and affirmation (see Figure 3.2). Experiencing higher conflict in the relationship in general predicted lower perceived responsiveness and affirmation. Willingness to sacrifice and inclusion of other in the self, on the other hand, explained very little variance in the outcomes.

Out of individual (attachment, individual differences, individual well-being, and demographics) variables, only higher actor attachment avoidance predicted lower perceived partner responsiveness and affirmation across analyses. Better physical health also predicted higher perceived responsiveness whereas greater life satisfaction and depression predicted higher perceived affirmation. There were several variables that explained very little variance in the outcome including all demographic variables and individual differences variables (other than attachment). There were no partner variables that predicted perceived responsiveness and affirmation consistently across analyses.

3.3.3 Exploratory Longitudinal Analyses

Finally, to examine whether the variables at baseline would be able to predict support six months later, we used Sample 3 ($n = 322$ [161 couples]) to estimate the longitudinal associations

between the predictor variables and the outcome.[2] Overall, we were able to predict 27.6% of variance in responsiveness and 18.2% of variance in affirmation with only actor effects. Models with actor and partner effects were somewhat less predictive (26.7% for responsiveness and 16.3% for affirmation). Relational and individual variables were equally predictive of support (see Table 3.2 for the full model results and Table 3.3 for normalized impact on the model). The only consistently important predictors across analyses were trust, commitment, attachment avoidance, and life satisfaction. Trust and commitment consistently predicted higher responsiveness and affirmation six months later, but relationship satisfaction did not. Both higher attachment anxiety and higher attachment avoidance predicted a decrease in perceived affirmation six months later but only attachment avoidance predicted responsiveness. Participants who reported higher life satisfaction at baseline reported higher perceived affirmation and responsiveness six months later.

3.4 Discussion

The purpose of the present study was to add to the growing body of literature on perceived partner support by using explainable machine learning to understand which variables reliably predict perceived partner support and which variables do not. It was the first study to compare a large number of variables providing novel insights into who perceives their partners as supportive and in which types of relationships. It is important to understand what researchers, practitioners, and policymakers should, and should not, focus on when designing interventions to improve support, whether for quitting smoking, achieving career goals, or beating cancer. Overall, we were able to predict a large amount of variance in both outcomes at baseline and six months later but not predict any change over time. Joel, Eastwick, Allison, et al. (2020) also found that variables included in existing datasets were unable to account for changes in relationship satisfaction and commitment over time. Thus, it appears that while we can predict outcomes as

[2]Because all five datasets had different lengths of follow-ups, it was not possible to examine longitudinal associations in a combined dataset. We selected the largest dataset that used a full measure of both responsiveness and affirmation at follow-up. Furthermore, because controlling for variables in a machine learning model introduces bias in the predictive accuracy of the model but does not affect the relative importance of the other variables, we did not include baseline support in the models in line with Joel, Eastwick, Allison, et al. (2020). We also estimated models where we predicted the change from baseline to follow-up. The R^2 for these models were negative suggesting we were unable to predict change.

a field, we are unable to predict changes over time. Because perceived partner support has been robustly associated with better individual and relationship well-being, it is useful to understand variables that predict perception of support. However, we should also be able to predict changes in our outcomes. Predicting actual change will likely become an important challenge for the future of relationship research.

3.4.1 Summary of the Most and Least Important Predictors and Implications for Theory

There were two types of variables that reliably predicted perceived support both at baseline and six months later: general relationship variables and attachment styles. The finding that general relationship variables is important for perceived partner support is unsurprising and in line with major relationship theories suggesting that happier relationships are important for perceived partner support (B. Feeney and Collins, 2015; Kelley and Thibaut, 1978; Mikulincer and Shaver, 2009; C.E. Rusbult and Van Lange, 2003; Ryan and Deci, 2000). Specifically, higher trust, commitment, and empathy toward partner, and lower conflict predicted an increase in perceived partner support. However, there were some relationship variables that varied across analyses and were less robustly associated with perceived partner support. Interestingly, relationship satisfaction was only predictive at baseline but not longitudinally suggesting that perhaps when taking away shared method variance, overall relationship satisfaction is not that important for perceived support, at least when compared against other relationship variables. Willingness to sacrifice was the only relationship variable that did not predict perceived partner support. Sacrifice is often seen as a mixed blessing in relationships (Day and Impett, 2018; Impett and Gordon, 2010) and we showed that people who are willing to sacrifice do not perceive their partners as more supportive and are not perceived as more supportive.

Of individual variables, actors' attachment avoidance was the only consistent individual predictor of partner support: highly avoidant people perceived their partners as less responsive and affirming. This finding is theoretically consistent given that individuals high in attachment avoidance are theorized to have a negative model of others and do not trust others' capacity to be there when needed (Bartholomew, 1990). Previous research has also found avoidance

to be associated with perceiving partners as less supportive (Collins and B. Feeney, 2004; Florian, Mikulincer, and Bucholtz, 1995; A. Martin, Paetzold, and Rholes, 2010). Interestingly, attachment avoidance was also more predictive of perceived partner support longitudinally than relationship related variables highlighting its centrality for perceived partner support. High attachment anxiety only predicted lower affirmation six months later. Results for attachment anxiety are often mixed because while anxious individuals seek reassurance and support excessively, they doubt whether they are worthy of receiving the support (Collins and B. Feeney, 2004; A. Martin, Paetzold, and Rholes, 2010). The finding may be explained by attachment anxious individuals being more focused on relationship maintenance than individual goal pursuit (Mikulincer and Shaver, 2007). As such they may perceive their partners also as less supportive.

Furthermore, there were five categories of variables that did not reliably predict perceived support: partner similarity, individual differences, individual and relational demographics, individual well-being, and all partner variables. Understanding which variables are not that influential in predicting perceived partner support is important so that researchers do not spend unnecessary time and resources on examining these variables and can instead focus on variables that are central to perceived partner support. There are several variables (e.g., inclusion of other in the self, gender, goal correspondence, regulatory focus, self-esteem, and self-efficacy) within these broader categories that would be expected to theoretically predict perceived partner support but when compared against more central predictors, are not that important. Finally, in line with previous research (Joel, Eastwick, and E.J. Finkel, 2017; Joel, Eastwick, Allison, et al., 2020), we found that while partner-reports explained a small amount of variance across outcomes, they did not explain any variance over and above actor-reports, did not predict much variance in the outcome, and even made the prediction worse longitudinally. Together, these findings can help constrain relationship theories to focus more on variables that are central to perceived partner support.

3.5 Supplementary Material

3.5.1 Discussion of Key Limitations

In this work, a number of key limitations are worth bearing in mind. Firstly, the combination of datasets represents a statistically 'bold' undertaking, especially when different variables are measured using different measures (*e.g.*, the outcome variable). The combination of samples with differences in demographics does not necessarily lead to either more diverse or more representative samples. Furthermore, and as described in the main text, not all variables were available in all samples, and we therefore struck a balance (a variable was included if it appeared in at least three of the five datasets). In our view, these compromises are necessary if one is to be able to explore the data in the best possible way. It would not, for instance, be a good idea to run the same analyses on only one of the datasets (the sample size becomes impractically small). To help justify the sample combination, we would reiterate the exploratory nature of this study - we do not provide (causal) effect size estimates, nor do we undertake null hypothesis significance testing, nor to we lend interpretative causal weight to the SHAP impact estimates provided as part of the analysis.

3.5.2 Details of Predictor Variables

Self-efficacy toward long-term goals was measured using a single item from the self-control scale (M = 5.86, SD = 1.65; "I am able to work effectively toward long-term goals"; (Tangney, Baumeister, and Boone, 2004) in all datasets. Self-control (Tangney, Baumeister, and Boone, 2004) indicates the extent to which one is able to control their emotions and desires and was measured using 12 items in Samples 1-2 and 10 items in Samples 3-5 (e.g., "I'm lazy") from the same scale. Self-esteem (Rosenberg, 1965) was measured using a 10-item Likert scale (e.g., "At times I think I'm not good at all"). Self-respect was measured using the 10-item self-respect scale (Kumashiro, E.J. Finkel, and C.E. Rusbult, 2002) in Samples 1-3 and 5 and with a single item from the scale in Sample 4 (e.g., "I have a lot of respect for myself"). Attachment style was measured using the Experience in Close Relationships (Brennan, C.

Clark, and Shaver, 1998) scale in Studies 1 and 2 and the Experience in Close Relationships – revised (R. Fraley, Waller, and Brennan, 2000) scale in Studied 3-5. Both include 36 items with two subscales: attachment anxiety (18 items; e.g., "I worry about being abandoned" and attachment avoidance (18 items; e.g., "I prefer not to show my partner how I feel deep down"). Participants' regulatory focus (i.e., whether one is concerned with promotion of dreams and goals or prevention of negative outcomes) was measured using the 11-item Regulatory Focus Questionnaire (Higgins et al., 2001) in Samples 1 and 2. The scale included six items for promotion (e.g., "I often do well at different things that I try") and five items for prevention orientation (e.g., "Growing up, I typically obeyed rules and regulations that were established by my parents"). In Samples 3-5, regulatory focus was measured using the 18-item General Regulatory Focus Measure (Lockwood, C. Jordan, and Kunda, 2002). The scale has nine items for promotion (e.g., "I frequently imagine how I will achieve my dreams and aspirations.") and nine items for prevention (e.g., "In general, I am focused on preventing negative events in my life."). Socially desirable responding was measured using the two-component model (Paulhus, 1984) including social desirability (ten items; e.g., "I have not always been honest with myself") and impression management (ten items; e.g., I'm a completely rational person").

There were also a number of physical and psychological well-being related variables that were included in the study. Subjective well-being was measured using the Satisfaction with Life scale (Diener et al., 0049), which includes five items (e.g., "In most ways, my life is close to ideal"). Symptoms of depression were measured using the 8-item depressive symptoms subscale of the Personal and Relationships Profile (e.g., "I feel sad quite often"; Straus et al., 1999) in Samples 1 and 2 and with a 13-item depression subscale from the Psychological Adjustment to Illness Scale (e.g., "Feeling blue"; Derogatis and Lopez, 1983) in Samples 3-5. Physical health was measured using a single item ("In general would you say your health is?") rated from poor to excellent in Samples 1 and 2. In Samples 3-5, physical health was measured using a 33-item Cohen-Hoberman Inventory of Physical Health (Allen, Wetherell, and M. Smith, 2017) which includes a checklist of symptoms such as back pain, weight change, and poor appetite.

There were a total of 11 relational variables in the datasets. Relationship status (dating, cohabiting, married) and children (yes, no) were dummy coded. Relationship length was measured in

years. Trust was measured using the 3-item (e.g., "How much do you trust your partner?") trust subscale of the Perceived Relationship Quality Components Inventory (Fletcher et al., 2011) in Samples 1 and 2. In Samples 3-5, trust was measured using a 12-item (e.g., "I can rely on my partner to keep the promises he/she makes to me.") Trust in Close Relationships scale (Rempel, Holmes, and Zanna, 1985). Relationship commitment was measured using the 7-item (e.g., "I want our relationship to last forever") commitment subscale of The Investment Model Scale (C.E. Rusbult, Martz, and Agnew, 1998). Relationship satisfaction was measured using the 5-item (e.g., "I feel satisfied with our relationship") subscale from the Investment Model Scale (C.E. Rusbult, Martz, and Agnew, 1998) in Samples 1, 2, and 5. Satisfaction was not explicitly measured in Samples 3 and 4 so we used a single item from the Dyadic Adjustment Scale: "The dots on the following line represent different degrees of happiness in your relationship. The middle point "happy" represents the degree of happiness of most relationships. Please circle the dot that best describes the degree of happiness – all things considered – of your relationship" rated from extremely unhappy to perfect.[3] The degree to which partners' experienced their identities to be interlinked was measured using the Inclusion of the Other in the Self (A. Aron, E. Aron, et al., 1991) measure in which participants were asked to select from Venn diagrams with increasing levels of overlap according to how much they felt the other was included in the self. Empathy toward partner was measured using 8-item (e.g., "I feel terribly sorry when things aren't going well for my partner".) partner subscale of the Interpersonal Reactivity Index (Davis, 1980).

Three further measures were also included but were only available in a subset of samples. General relationship conflict was measured in Samples 1-3 and 5 using the conflict subscale of the Personal and Relationships Profile (Straus et al., 1999). The scale asks about disagreement on various topics such as money, friends, or sex (e.g., "My partner and I disagree about when to have sex"). Willingness to sacrifice (Van Lange et al., 1997) was measured in Samples 1-3 and involved participants first rating four of their most important activities and then they were asked about each activity: "Imagine that it was not possible for you to engage in Activity and

[3]Due to the differences in satisfaction measures across the samples, sample moderated the association between satisfaction and the different outcomes but the nature of the effect remained the same. This was the only variable that sample moderated despite some differences in other measures across the samples. However, because of the importance of relationship satisfaction in predicting support and because random forest can fit complex non-linear interactions (Breiman, 2001a) we retained this variable in the model.

maintain your relationship (impossible for reasons that are not your fault). To what extent would you consider giving up the activity?" A measure of goal compatibility (i.e., how problematic one partner's goals were for the other partner) was available in Samples 1-4 and was measured using five items in Samples 1 and 2 (e.g., "Sometimes I feel like my goals are incompatible with my partner's goals") and nine items in Samples 3 and 4 (e.g., "My partner does not completely approve of my goals"). The majority of the scales used across the samples were well-established and have good reliability and validity. The final list of variables differed somewhat from the preregistration upon discovering that some of the variables were not available across at least three datasets at baseline and therefore it was not possible to include them.

Table 3.4: The List of Included Variables with a More Detailed Theoretical Rationale for Inclusion.

Variable	**Rationale**
Relationship variables	
relationship status relationship length trust commitment satisfaction empathy toward partner willingness to sacrifice conflict goal compatibility/correspondence actual inclusion of the other in self	Support is likely to be more important in early stages of the relationship with the importance of support declining over time (Bühler et al., 2019; Jakubiak, B. Feeney, and Ferrer, 2020). Based on most relationship theories, we expect that overall better quality relationships (high trust, commitment, satisfaction, empathy, willingness to sacrifice, and lower conflict) will predict higher support (B. Feeney and Collins, 2015; Kelley and Thibaut, 1978; Mikulincer and Shaver, 2009; C.E. Rusbult and Van Lange, 2003; Ryan and Deci, 2000). Interdependence theory suggests that higher goal conflict results in less support. Previous research has shown that when goal conflict is high, partners are less likely to provide support toward each other's opportunities and less likely to perceive their partners as supportive (Gere and Schimmack, 2013; L. Vowels, Carnelley, and Francois-Walcott, 2021).
Individual variables	
gender	The research on gender differences in support is mixed with some studies finding that women are more supportive than men and others finding no differences (Verhofstadt, Buysse, and Ickes, 2007).

age	Bühler et al. (2019) found that support became more important for relationship satisfaction as people aged whereas Jakubiak, B. Feeney, and Ferrer (2020) found that the association between support and relationship satisfaction was stronger in the younger sample compared to the older sample. Therefore, it is unclear whether and, if so, how age might be associated with support.
ethnicity	No prior studies have examined this, but it is possible that support toward goals is more important in certain ethnic groups than others.
education	No prior studies but it is possible that the importance of support toward life's opportunities differs across education levels. For example, people who are less educated may be more focused on making ends meet rather than pursuing life's opportunities compared to people who are more highly educated.
children	No prior studies but the presence of children may mean that partners have to divide their support between partner and children and therefore be less supportive.
self-control	Self-control has previously been associated with higher relationship satisfaction (Zuo et al., 2020). It is possible that individuals higher (vs. lower) in self-control also make better support providers and also perceive their partners as more supportive but this has not been examined in previous literature.
self-esteem	Self-esteem has been positively associated with support (Harris and Orth, 2020). We expect that self-esteem will be positively associated with support.
self-respect	Self-respect has been associated with pro-relationship behaviors (Kumashiro, E.J. Finkel, and C.E. Rusbult, 2002). We expect that self-respect will be positively associated with support.
attachment anxiety	Attachment theory (Bowlby, 1969) suggests that individuals higher in attachment anxiety doubt their worthiness and seek excessive support and reassurance. Results are mixed on whether anxious individuals experience their partners as less supportive (A. Martin, Paetzold, and Rholes, 2010) as they generally see themselves as not worthy of support but also seek excessive reassurance.

attachment avoidance	Attachment theory suggests that avoidant individuals rely on themselves and do not trust others and thus experience their partners as less supportive, which has been supported in the literature (B. Feeney, 2004)
physical health	Support has previously been linked to physical health in that support predicts better physical health (Reblin and B.N. Uchino, 2008). However, it is also possible that individuals who have poorer physical health perceive their partners as less supportive as they may require more support from their partners compared to people in better physical health.
depression	Symptoms of depression include feeling like a burden to other people. Therefore, individuals who score higher on depression may experience their partners as less supportive compared to individuals who are lower on depression. Previous research has shown that higher levels of depression predict lower partner support (Canevello and Crocker, 2010).
life satisfaction	Life satisfaction is usually examined as an outcome of partner support (Drigotas, 2002), and shows that higher levels of support in relationships predicts higher life satisfaction. However, it is also likely that individuals who are satisfied with their lives will be more likely to perceive their partners as more supportive.
self-efficacy	Self-efficacy is often examined as an outcome of partner support (e.g., L. Vowels, Carnelley, and Francois-Walcott, 2021). However, we would also expect that individuals higher in self-efficacy perceive their partners as more supportive.
regulatory focus (promotion)	Promotion orientation has been previously positively associated with partner support (Righetti, C. Rusbult, and Finkenauer, 2010).
regulatory focus (prevention)	While we expect promotion orientation to be positively associated with support, prevention orientation was included because it is the other continuum of regulatory focus and has been included in past research (Righetti, C. Rusbult, and Finkenauer, 2010).
social desirability	This was included in the model as a measure of socially desirable responding, we made no specific predictions for the variable.

impression management	This was included in the model as a measure of socially desirable responding, we made no specific predictions for the variable.

3.5.3 Data Analysis

Data Preparation. Any missing variables in each dataset were included in the combined dataset and designated as missing. All continuous variables were scaled to be between 0 and 8 with higher numbers indicating higher levels of the variable (e.g., higher number in self-esteem would indicate higher self-esteem). All categorical variables were dummy coded to be 0 or 1 with each category included in the analyses. A maximum of 0.05% of the data for each variable were missing, and any missing data points were imputed using the scikit-learn package Iterative Imputer (Pedregosa et al., 2011) with a Bayesian ridge estimator. If the variable was missing from an entire subsample, it was not included in an analysis in which the subsample was used.

Analyses. The results were analyzed using Python 3.7 and the code can be found here: `https://github.com/matthewvowels1/Aff_Eff_PPR`. Each dataset was analyzed using a random forest regressor (Breiman, 2001a). A random forest is a type of decision tree that trains on bootstrapped sub-samples of the data to avoid overfitting. The tree can model highly non-linear relationships in the data, and therefore represents a significantly more flexible model than a linear regressor. In general, random forest models are sensitive to hyperparameter settings (such as the number of estimators or the maximum depth of the decision tree). However, tuning hyperparameters requires a separate validation data split which reduces the effective sample size available for training and testing. Therefore, we used the default "scikit learn" random forest regressor with k-fold cross-validation (Pedregosa et al., 2011). The out-of-bag error is a built-in metric frequently used to estimate the performance of random forests (Joel, Eastwick, and E.J. Finkel, 2017; Joel, Eastwick, Allison, et al., 2020), but in some circumstances this metric has been shown to be biased above the true error (Janitza and Hornung, 2018; Mitchell, 2011). By using a k-fold cross-validation approach, instead of the out-of-bag error, we were able to test the model over the entire dataset, and to acquire estimates for the standard error (see below).

A ten-fold cross-validation scheme was used to train and test the model. This meant the total

dataset was randomly split into ten equally sized folds. The model was trained on nine out of ten folds, tested on the tenth, and the test fold performance was recorded. This was repeated until all ten folds had been used as a test set. The average performance, as well as the standard error across the ten folds, provided an estimate of model performance on unseen data. The metrics for test data model performance used were the mean-squared error (which is the averaged squared difference between the prediction and the observed value), the R2, and the variance explained. The last model trained was then saved, and explained using the "SHapley Additive exPlanations" package (SHAP) (Lundberg and S.-I. Lee, 2017; Lundberg, G.G. Erion, and S.-I. Lee, 2017; Lundberg, G. Erion, et al., 2020).

The SHAP is a unified framework for undertaking model explanation, and derives from the seminal game theoretic work of Lloyd Shapley (Shapley, 1953). The framework conceives of predictors as collaborating agents seeking to maximize a common goal (i.e., the regressor performance). The approach involves systematically evaluating changes in model performance in response to including or restricting the influence from different combinations of predictors. For example, the SHAP TreeExplainer function from the SHAP software implementation provides estimations of the per-datapoint, per-predictor impact on model output, as well as the average predictor impacts. These estimations are called 'explanations' because they explain why a particular regressor performs the way it does. The results are provided as feature importances, which describe how important the variable is for the model outcome and how much it changes the outcome. For the analysis, the default settings of the SHAP package TreeExplainer were used, and the entire dataset was fed to the model for explanation. The combination of the powerful function approximation capabilities of random forests with the consistent and meaningful estimations of per-datapoint, per-predictor impact on model output enables a reliable and informative exploration of predictor importance, and the identification of key predictor interactions.

The analyses were conducted separately by first including as many participants as possible in each analysis and then by including as many variables as possible. This resulted in a total of four analyses (three for perceived partner responsiveness, one for affirmation) which were conducted twice: once including only actor effects and once including both actor and partner

effects. The included variables and the results for all analyses can be found on the OSF project page. Random forests in their current form are unable to explicitly model hierarchies in the data and it is possible that hierarchical data can inflate the predictive performance. However, given we were primarily interested in the relative performance of different predictors, which is not affected by hierarchical data, this is less of an issue in the current study.

3.5.4 Self-Efficacy Analysis Results

Total Variance Explained (Research Questions 1-3)

The models with actor effects only were less able to predict self-efficacy with between 23.6% and 28.8% of the variance explained. For self-efficacy, individual variables (72.4% - 73.6%) were more important predictors compared to relational variables (23.0% - 27.6%). Partner effects contributed between 0.6% and 1.9% additional variance in the models for self-efficacy. Partner's individual variables contributed between 16.9% and 19.3% of the variance and partner's relational variables contributed between 4.7% and 8.0% of the variance in the models with both actor and partner effects.

Most Predictive Variables (Research Question 4)

In addition to identifying the most important factors for perceived partner support, we also examined which predictors alongside perceived partner support were the most important predictors of self-efficacy. Affirmation was in the top-10 predictors of self-efficacy with higher perceived partner affirmation predicting greater self-efficacy. Perceived partner responsiveness was also positively associated with self-efficacy; however, it was only in the top-10 most important predictors once out of four models that it was included in. Self-control was the highest predictor of self-efficacy with higher self-control predicting higher self-efficacy.

Other consistently predictive individual variables were self-esteem and life satisfaction with higher scores in both predicting higher self-efficacy. Self-respect was also in the top-10 predictors in six out of eight analyses and higher scores in self-respect predicted higher scores in self-efficacy. There were several individual actor variables that predicted very little variance in self-efficacy including gender, age, education, attachment anxiety or avoidance, prevention

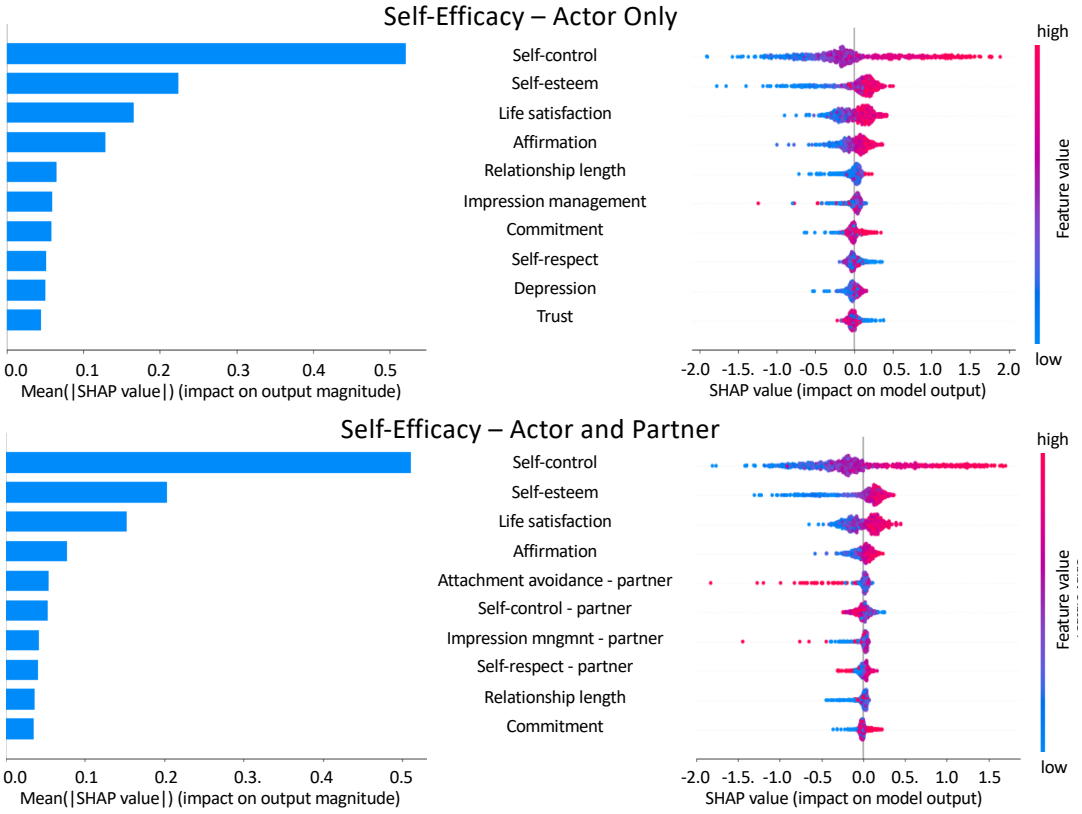
Table 3.5: The Overall Prediction Results for Each Outcome Variable for Individual and Relational Variables and Models with Actor Effects Only and with Actor and Partner Effects

Outcome Self-Efficacy	% Variance M (SE)	MSE M (SE)	R^2 M (SE)	Individual $\%_a/\%_p$	Relational $\%_a/\%_p$
Model 1	23.6 (0.03)	2.14 (0.10)	.22 (0.03)	73.6	26.4
+ Partner	25.5 (0.02)	2.09 (0.10)	.24 (0.02)	58.6 / 18.8	17.9 / 4.7
Model 2	24.4 (0.03)	2.04 (0.10)	.24 (0.03)	73.0	27.0
+ Partner	26.0 (0.04)	1.99 (0.09)	.25 (0.04)	59.9 / 16.9	16.2 / 7.8
Model 3	28.8 (0.03)	1.97 (0.12)	.28 (0.03)	72.4	27.6
+ Partner	30.1 (0.02)	1.94 (0.12)	.29 (0.02)	58.5 / 17.5	16.0 / 8.0
Model 4*	27.2 (0.04)	2.19 (0.13)	.26 (0.03)	77.0	23.0
+ Partner*	27.8 (0.03)	2.18 (0.14)	.26 (0.03)	61.0 / 19.3	12.9 / 6.8

Note. $\%_a$ refers to the percentage of variance explained by actor variables, $\%_p$ refers to the percentage of variance explained by partner variables. The first model for each outcome variable included as many samples as possible and subsequent models included as many variables as possible. The full list of excluded variables and samples can be found on the OSF project page. * Results presented in figures.

orientation, and physical health. There were only two relational variables that consistently predicted self-efficacy: relationship length and commitment. Individuals who had been in their relationship for longer and were more committed in their relationship reported higher levels of self-efficacy. Promotion orientation was among the top-10 predictors in the model it was included in with higher scores predicting higher self-efficacy. Relationship satisfaction was in the top-10 variables in six of the eight models apart from the model with affirmation also included in the model. Higher relationship satisfaction predicted higher self-efficacy. The only consistent partner variable was attachment avoidance: higher partner attachment avoidance predicted lower actor self-efficacy.

Figure 3.3: The Top-10 Most Important Predictors for Self-Efficacy for Models with Actor Effects and Actor and Partner Effects.



Note. The figure presents the results from the model with affirmation as a predictor. Responsiveness was not in the top-10 predictors in the models and therefore was not included in the figures.

CHAPTER 4

Application of Causal Discovery and Causal Inference

In Chapter 2, I recommended that researchers engage with causal methods when undertaking their research. This Chapter contains an example of such an application, specifically with causal discovery and causal inference techniques (with targeted learning for estimation) for investigating the causal links between attachment styles and mental health during the COVID-19 pandemic. The content of this chapter is drawn from the following publication:

Vowels, L.M., **Vowels, M.J.**, Carnelley, K.B., Millings, A. Miller, J.G., Under Review. Toward a Causal Link between Attachment Styles and Mental Health during the COVID-19 Pandemic.

Contribution: All analyses, methodological write-up and presentation of results, manuscript editing.

Abstract: Recent research has shown that insecure attachment, especially attachment anxiety, is associated with poor mental health outcomes, especially during the COVID-19 pandemic. Other research suggests that insecure attachment may be linked to nonadherence to social distancing behaviors during the pandemic. In a nationally representative UK sample (cross-sectional n = 1325; longitudinal n = 950) we examine the causal links between attachment styles (secure, anxious, avoidant), mental health outcomes (depression, anxiety, loneliness), and adherence

to social distancing behaviors during the first several months of the UK lockdown (between April-August 2020). The data were analyzed using state-of-the-art causal discovery and targeted learning algorithms to identify causal processes. The results showed that insecure attachment styles were causally linked to poorer mental health outcomes, mediated by loneliness. Only attachment avoidance was causally linked to nonadherence to social distancing guidelines. Future interventions to improve mental health outcomes should focus on mitigating feelings of loneliness. Limitations include no access to pre-pandemic data and the use of categorical attachment measure.

4.1 Introduction

The COVID-19 pandemic has brought many challenges including navigating how best to protect our health and well-being, while living our lives to the fullest. For some, the circumstances surrounding COVID-19 have been more detrimental to their mental health than for others (Shevlin et al., 2021). In this paper, we test novel hypotheses with important implications for well-being using data from early in the pandemic collected by the COVID-19 Psychological Research Consortium Study (C19PRCS), a longitudinal survey tracking changes in behavior and mental health over the pandemic in a large representative sample of the UK adult population. We aimed to develop a theoretical causal model to better understand how individual differences in attachment styles influence adherence to social distancing behaviors, as well as mental health outcomes (loneliness, depression, anxiety) in the context of the COVID-19 pandemic. Importantly, we used a cutting-edge causal discovery algorithm known as Structural Agnostic Modeling (SAM; Kalainathan et al., 2020) to estimate the causal structure of the model. We then estimated and tested specific causal effects within the model using targeted learning (M. J. van der Laan and S. Rose, 2011). Using these advanced methods allowed us to examine the possible causal relationships between attachment styles, social distancing behaviors and mental health outcomes during the COVID-19 pandemic in a representative sample of UK adults.

4.1.1 Mental Health during the COVID-19 Pandemic

The pandemic has led to an increase in mental health difficulties in many nations (Burkova et al., 2021; Pierce et al., 2020; Randall et al., 2021). For example, a systematic review of 43 cross-sectional studies (Vindegaard and Benros, 2020) showed higher rates of depression, anxiety and post-traumatic stress disorder (PTSD) compared to before the pandemic. Several longitudinal studies support this pattern, for example, (Pierce et al., 2020) compared pre-pandemic levels to one month into lockdown in the UK and showed roughly a 10% increase in depression and anxiety after the pandemic began. Other longitudinal studies that examined outcomes over the course of the pandemic show mixed results. Huckins et al. (2020) found an increase in anxiety and depression over 12 weeks during the pandemic in a student sample, whereas C. Wang et al. (2020) found that the moderate to severe levels of stress, depression, and anxiety assessed at the start of the pandemic and four weeks later in China initially did not change over this time period.

4.1.2 Attachment Styles and Mental Health during the COVID-19 Pandemic

Although many people have found the COVID-19 pandemic stressful, not everyone has experienced worse mental health (e.g., Shevlin et al., 2021). Research has examined several factors that predict who is more likely to experience elevated depression, anxiety, and poor wellbeing due to the pandemic. One such variable is adult attachment style, which describes the human predisposition to form close emotional ties to others, driven by the attachment behavioral system. The primary purpose of forming these ties or 'attachments' is to maintain proximity to our caregivers and thus ensure survival (Bowlby, 1969). The quality of these attachment relationships throughout life becomes internalized over time as 'attachment styles' (Bowlby, 1969; Brennan, C. Clark, and Shaver, 1998). When an individual experiences sensitive, responsive care from their attachment figures (parents, partners, loved ones), they develop attachment security. Attachment security is associated with happiness, life satisfaction, and more positive physical and mental health and well-being outcomes (Mikulincer and Shaver, 2016). Conversely, a history of experiences in which attachment figures are rejecting, or inconsistent, leads to attachment insecurity, conceptualized as avoidance (of intimacy) or anxiety (about abandonment) (Brennan,

C. Clark, and Shaver, 1998). Low scores on both dimensions indicates a prototypically secure attachment style. These individual differences should theoretically link to the ability to adapt behaviorally and emotionally to the demands enforced by the COVID-19 pandemic.

Indeed, research demonstrates that adult attachment styles have influenced well-being during the COVID-19 pandemic. Moccia et al. (2020) found that high levels of need for approval differentiated between those who reported moderate-severe psychological distress versus no distress during the pandemic in Italy, suggesting that attachment anxiety is a risk factor for mental health problems. Similarly, Mazza et al. (2021) found in an Italian sample of healthcare workers that attachment anxiety was positively associated with high stress, depression, and anxiety. Carbajal et al. (2021) also focused on healthcare workers – first-time responders seeking mental health treatment during COVID-19 in the USA– found that both attachment anxiety and avoidance were negatively associated with resilience and positively associated with depression and PTSD, whereas attachment anxiety was positively associated with generalized anxiety and suicidality. Building on this cross-sectional work, L. Vowels, Carnelley, and Stanton (2022) examined the effects of adult attachment on depression and anxiety in two longitudinal studies that assessed depression and anxiety weekly over five weeks near the start of lockdown in the UK. The study found that those high in attachment anxiety experienced higher depression and anxiety than those low in attachment anxiety. Furthermore, those higher in attachment anxiety maintained their elevated levels of depression and anxiety across the five weeks, but those lower in attachment anxiety reported decreasing scores over time. In a study that compared pre-pandemic to during pandemic levels, L. Vowels, Carnelley, and Stanton (2022) found that those higher (versus lower) in attachment anxiety reported increasing depression and anxiety over time. Attachment avoidance did not predict depression or anxiety in either study. The evidence above suggests that insecure attachment, especially attachment anxiety, may be a predictor of poor mental health during the pandemic.

4.1.3 Attachment Styles and Social Distancing Behaviors

The COVID-19 pandemic context presents several threats to the attachment system, most notably separation from loved ones due to enforced national lockdowns and social distancing measures

as well as persistent mortality salience and exposure to illness-related cues. This context also required individuals to enact prescriptive COVID-19-related protective behaviors to prevent infection and/or spreading the disease to others such as hand washing, maintaining a physical distance from others, and wearing face masks.

Attachment style modulates how we respond to stress and threat (Brennan, C. Clark, and Shaver, 1998) including separation from loved ones (Ainsworth et al., 1978; R. Fraley and Shaver, 1998) and cues of danger, such as illness. Consequently, attachment styles are predictive of how people appraise (Meredith, Strong, and J. Feeney, 2005) and cope (Krasuska et al., 2018) with symptoms, manage chronic conditions (Ciechanowski et al., 2004) and take preventive measures, including enacting protective health behaviors (Pietronmonaco, B. Uchino, and Schetter, 2013). Moreover, attachment styles have been found to be predictive of the capacity for prosocial behavior and empathy (Boag and Carnelley, 2016; Mikulincer, Shaver, et al., 2005). Thus, we believe that the requirement to physically and socially distance from others presents the greatest threat to the attachment system that would initiate individual coping responses to regulate this threat, driven by attachment style.

It follows, then, that attachment style is likely to be a key predictor of the enactment of social distancing in the context of COVID-19; this is supported by some initial evidence. In a US context, earlier in the pandemic, Lozano and R. Fraley (2021) examined attachment styles as a predictor of engagement in, and reminding others about, the following protective behaviors: hand washing, social distancing, wearing face masks, refraining from touching face/mouth and disinfection of items. Attachment avoidance was negatively associated with both engagement in, and reminding others about the behaviors, and attachment anxiety was positively associated with reminding others. Brulin et al. (2022) examined the associations between attachment and adherence to the COVID-19 regulations in Sweden. Both attachment anxiety and avoidance were associated with nonadherence to authorities' guidelines, such as social distancing and hand washing. While these findings are consistent with attachment theory and are a first attempt to apply and explore attachment to the COVID-19 context, much work remains to be done to delineate the ways in which individual differences in attachment style affect people's coping and adherence to social distancing behaviors in the pandemic.

4.1.4 Toward Causality in the Present Research

Prior research evidence is derived from non-experimental studies and their authors have understandably refrained from making causal claims about the association between attachment styles, mental health and adherence to health guidance. The well-known phrase "correlation is not causation" cautions researchers in the social and health sciences (M. Hernan, 2018a) to be mindful about the scope and confidence of their conclusions when interpreting results obtained from non-experimental and cross-sectional studies. Well-intentioned caution in this regard often means that cross-sectional data is assumed to tell us nothing about causality. However, recently, several researchers have argued that reluctance to make causal inferences does little to make the interpretations more reliable (Grosz, Rohrer, and Thoemmes, 2020; M. Hernan, 2018a; Rohrer, 2018; M. J. Vowels, 2021). It instead results in a conflation of causal and correlational language, a lack of transparency concerning the (causal) assumptions underlying the research, and a reluctance to adopt robust statistical techniques for improving the validity of our analyses (Grosz, Rohrer, and Thoemmes, 2020; Rohrer, 2018; M. J. Vowels, 2021). Indeed, such techniques do exist, and a vast array of statistical developments can be applied to the estimation of causal quantities from observational data (Imbens and D. Rubin, 2015; Pearl, 2009; Pearl, 2012). Furthermore, there exist techniques for estimation of causal quantities given an assumed structure - a process known as causal inference (Pearl, 2009; Tian and Pearl, 2002; M. J. van der Laan and S. Rose, 2011) - as well as techniques for estimating the structure itself – a process known as causal discovery (C. Glymour, K. Zhang, and Spirtes, 2019; M. Vowels, N. Camgoz, and Bowden, 2022).

Researchers in the field of causal discovery warn against interpreting the output of such algorithms too literally, and thus they should be used to inform theory rather than overrule it (M. J. Vowels, 2021; M. Vowels, N. Camgoz, and Bowden, 2022). These methods nonetheless provide a means to validate certain aspects of theories, to explore data for possible causal structures and thus to help us specify models that reflect the discovered structure, and which can then be used to test hypotheses. In this work, we take advantage of recent progress in the domain of causal discovery, by using a state-of-the-art causal discovery algorithm known as Structural Agnostic Modeling (SAM; Kalainathan et al., 2020). Our aim was to develop

a causal theoretical model between individual differences in attachment styles (i.e., secure, anxious, avoidant), social distancing behaviors (i.e., adherence to government guidelines) and mental health outcomes (i.e., loneliness, depression, anxiety). We then aimed to quantify the causal estimates using a targeted learning approach (M. J. van der Laan and S. Rose, 2011) which sits at the intersection of machine learning and causality. Targeted learning allowed us to estimate the relationships between two target variables (i.e., the causal relationship between attachment anxiety and depression) without making assumptions about the functional form of that relationship (e.g., linear/non-linear).

4.1.5 The Current Research

We hypothesized that those with an insecure attachment style would report greater loneliness, anxiety, and depression compared to those with a secure attachment style; especially so for those with an anxious or fearful attachment style. In addition, we expected those with a secure or avoidant attachment style to better adhere to social isolation/physical distancing than those with an anxious attachment style. Finally, we expected secure individuals to better adhere to social isolation/physical distancing than those with an insecure attachment style. Taking data from two different time points, we also examine the effects of attachment styles on depression and anxiety over time. We examine these hypotheses in a secondary analysis of data from two waves of the COVID-19 Psychological Research Consortium Study (C19PRCS), a longitudinal survey tracking changes in behavior and mental health over the pandemic in a large nationally representative UK sample of adults.

4.2 Method

4.2.1 Participants and Procedure

We conducted a secondary analysis of UK data collected in waves two and three of the longitudinal, internet-based survey COVID-19 Psychological Research Consortium Study (C19PRCS). A detailed methodological account of the C19PRCS is available elsewhere (McBride et al., 2021)

and the data is publicly available on the OSF at this location: `https://osf.io/v2zur/`. Briefly, UK fieldwork of the C19PRC Study was conducted between April/May 2020 for Wave 2 and July/August 2020 for Wave 3. During Wave 2, strict social distancing measures were in place whereas during Wave 3, many of the measures had been lifted for the summer and restaurants and pubs were open and two households were allowed to meet indoors. Quota sampling was used to recruit a panel of adults who were nationally representative of the UK population in terms of age, sex, and household income. Participants were aged 18 years or older at the time of the survey, must have been able to complete the survey in English, and resident in the UK. Consenting adults completed the survey online and were reimbursed by Qualtrics for their time. Ethical approval for this research was provided by a UK University Psychology department (Reference number: 033759). At Wave 2, 1406 participants participated in the survey, but some people did not report on their attachment styles and were thus removed from the analyses. The final sample consisted of 1325 individuals in the cross-sectional analyses and 950 in the longitudinal analyses. The full demographic variables can be found in Table 4.1. This study was not preregistered.

Table 4.1: Demographic Characteristics of Participants in Cross-Sectional and Longitudinal Analyses.

Demographic Variables	**Cross-Sectional (n = 1325)**	**Longitudinal (n = 950)**
Age	M = 49.03 SD = 14.94 Range = 18-88	M = 51.84 SD = 14.45 Range = 18-88
Change in Income	M = -9.5 SD = 26.3 Range = -100 - 100	M = -8.8 SD = 24.2 Range = -100 - 100
	n (%)	**n** (%)
Gender		
Man	683 (51.5%)	521 (54.8%)
Woman	639 (48.2%)	426 (44.8%)
Transgender	1 (0.1%)	1 (0.1%)
Other	2 (0.2%)	2 (0.2%)
Relationship Status		
Married	641 (48.4%)	478 (50.3%)
Single	305 (23.0%)	208 (21.9%)
Cohabiting	159 (12.0%)	101 (10.6%)
Separated	21 (1.6%)	18 (1.9%)
Divorced	108 (8.2%)	83 (8.7%)
Widowed	34 (2.6%)	27 (2.8%)
In a registered same-sex civil partnership	6 (0.5%)	4 (0.4%)
In a relationship but not living together	51 (3.85)	31 (3.3%)
Children		
No	1002 (75.6%)	766 (80.6%)
Yes	323 (24.4%)	184 (19.4%)
Education		
No qualifications	43 (3.2%)	32 (3.4%)
O-Level/GCSE or similar	251 (18.9%)	160 (16.8%)
A-Level or similar	229 (17.3%)	148 (15.6%)
Technical qualification	129 (9.7%)	97 (10.2%)
Undergraduate degree	378 (28.5%)	304 (32.0%)
Diploma	72 (5.4%)	46 (4.8%)
Postgraduate degree	208 15.7%)	150 (15.8%)
Other qualification	15 (1.1%)	13 (1.4%)
Race/Ethnicity		
White British/Irish	1168 (88.2%)	861 (90.6%)
White non-British/Irish	64 (4.8%)	34 (3.6%)
South Asian	43 (3.2%)	24 (2.5%)
Chinese	15 (1.1%)	9 (0.9%)
Caribbean or African	13 (1.0%	9 (0.9%)
Arab	3 (0.2%)	1 (0.1%)
Other	19 (1.4%)	11 (1.2%)
Religion		
Christian	697 (52.6%)	513 (54.0%)
Muslim	34 (2.6%)	16 (1.7%)
Jewish	10 (0.8%)	8 (0.8%)
Hindu	7 (0.5%)	4 (0.4%)
Buddhist	11 (0.8%)	7 (0.7%)
Sikh	7 (0.5%)	5 (0.5%)
Atheist	318 (24.0%)	228 (24.0%)
Agnostic	163 (12.3%)	122 (12.8%)
Other religious conviction.	78 (5.9%)	47 (4.9%)
Employment		
Full time	720 (54.3%)	415 (43.7%)
Part time (regular hours)	152 (11.5%)	107 (11.3%)
Zero hours contract	23 (1.7%)	14 (1.5%)
Other flexible work practice	29 (2.2%)	22 (2.3%)
Unemployed (because of coronavirus)	36 (2.7%)	26 (2.7%)
Unemployed (not because of coronavirus)	204 (15.4%)	128 (13.5%)
Retired	272 (20.5%)	238 (25.1%)
Keyworker		
No	940 (70.9%)	703 (74.0%)
Yes	385 (29.1%)	247 (26.0%)
Chronic illness		
No	1004 (75.8%)	710 (74.7%)
Yes	321 (24.2%)	240 (25.3%)
Pregnant (self or partner)		
No	1301 (98.2%)	940 (98.9%)
Yes	24 (1.8%)	10 (1.1%)

Table 4.2: Means, Standard Deviations, and Correlations with Confidence Intervals for Wave 2.

Variable	M	SD	1	2	3	4	5	6	7
1. Secure (n = 441)	0.33	0.47							
2. Fearful (n = 124)	0.28	0.45	-.44** [-.48, -.39]						
3. Anxious (n = 367)	0.09	0.29	-.23** [-.28, -.18]	-.20** [-.25, -.15]					
4. Avoidant (n = 392)	0.3	0.46	-.46** [-.50, -.41]	-.40** [-.45, -.36]	-.21** [-.26, -.16]				
5. Social distancing	12.84	5.94	-0.03 [-.09, .02]	.13** [.08, .18]	.09** [.04, .15]	-.15** [-.21, -.10]			
6. Depression	5.25	5.94	-.16** [-.21, -.11]	.21** [.16, .27]	.10** [.05, .15]	-.11** [-.16, -.05]	.31** [.26, .36]		
7. GAD	4.41	5.38	-.15** [-.21, -.10]	.21** [.16, .26]	.11** [.05, .16]	-.12** [-.17, -.06]	.26** [.21, .31]	.85** [.84, .87]	
8. Loneliness	4.73	1.81	-.21** [-.26, -.15]	.26** [.21, .31]	.11** [.05, .16]	-.11** [-.16, -.05]	.17** [.11, .22]	.59** [.56, .63]	.53** [.49, .57]

Note. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. GAD = Generalized anxiety disorder. * indicates $p < .05$, ** indicates $p < .01$. Secure: n = 441.

4.2.2 Measures

Attachment style was measured using the Relationships Questionnaire (Bartholomew and Horowitz, 1991) which is a categorical measure of the four attachment styles: secure, anxious, avoidant, and fearful avoidant. Social distancing practices, in accordance with government guidelines during the first UK lockdown, were assessed using a list of 16 statements with respect to the past week. Generalized Anxiety Disorder Scale (GAD-7; Spitzer et al., 2006) was used to measure generalized anxiety and the Patient Health Questionnaire PHQ-9: Kroenke, Spitzer, and J. Williams, 2002) depression. Loneliness was measured using a 3-item Loneliness Scale (Hughes et al., 2004). We also included a set of variables that were theoretically causally related to the central variables in the study that we controlled for in the models. These variables include demographics, COVID-19 related anxiety and perceived one month risk, and hygiene practices. Due to the space limitations, a detailed description of all variables can be found in Supplemental material (accessible here: `https://osf.io/4ypuk/` and provided at the end of this chapter).

4.2.3 Data Analysis

We used a state-of-the-art causal discovery algorithm known as Structural Agnostic Modeling (SAM; Kalainathan et al., 2020) to infer the cross-sectional structure for Wave 2 (17 variables and 1325 participants) and the longitudinal structure across Wave 2 and Wave 3 (19 variables from 895 participants). We included all variables that were expected to be causally linked to the main variables of interest and thus affect the estimation of the causal relationships. We applied a constraint preventing the discovery of causal effects backwards in time, as well as constraints preventing causal links between certain demographics: age and gender cannot be effects; change in income was measured as the change between Waves 1 and 2 and thus was prevented from affecting all demographic variables. We then used a state-of-the-art method at the intersection of machine learning and causality known as targeted learning (M. J. van der Laan and S. Rose, 2011) to estimate the specific effect of attachment styles on social distancing behaviors and mental health outcomes. Details of the data analysis can be found in Supplemental material (accessible here: `https://osf.io/4ypuk/` and provided at the end of this chapter).

4.3 Results

Table 4.2 presents the means and standard deviations as well as the bivariate correlations between the main study variables. The number of people identifying as secure (n = 441), avoidant (n = 392), and anxious (n = 367) were comparable, with fewer people identifying as fearful avoidant (n = 124).

4.3.1 Cross-Sectional Model (Wave 2)

The graphical illustration of the results from the causal discovery algorithm can be found in Figure 4.1 for the cross-sectional data and Figure 4.2 for the longitudinal data. The full adjacency matrices with all causal paths can be found in the Supplemental Material Figures S2 and S3. The directed causal relationships between the cause and effect with a confidence score of at least 0.5 (where this score ranges between 0 and 1) have been included in the graphs. We can see from Figure 4.1 that the only putative cause for attachment styles is participants' gender whereas attachment styles cause relationship status, anxiety, depression, loneliness, and social distancing behaviors. Loneliness was identified as a mediator between attachment styles and anxiety, depression, and social distancing behaviors. We have highlighted the theoretically relevant relationships in bold but have also included the required control variables in Figure 4.1. Precision variables are included in grey as they are not necessary to produce an unbiased estimate but can make the standard errors tighter because they explain variance in the outcome variables.

The algorithm does not provide the direction or the size of the effects and thus we used targeted learning to estimate the causal effects between attachment styles and each of the outcome variables. For the targeted learning, we used only the control variables that were essential in providing accurate causal estimates in line with what are known as the d-separation rules for causal graphs, and precision variables that can help provide tighter estimates of the effect (i.e., smaller standard errors) (Cinelli, Forney, and Pearl, 2022). Thus, our set of control variables consisted of gender as a confounding variable and age, chronic illness, number of children in the household, adults in the household, change in income, keyworker status, one month risk,

COVID-19 anxiety, and pregnancy as precision variables for all outcomes except for loneliness. For loneliness, we used gender as a control variable but only change in income, keyworker status, one month risk, COVID-19 anxiety, and chronic illness as precision variables.

Table 4.3: The Cross-Sectional and Longitudinal Results from the Targeted Learning Analyses.

	Anx.			**Depr.**			**Lonel.**			**Soc. Dist. Beh.**		
Group (0 = Secure)	Effect	SE	p	Effect	SE	p	Effect	SE	p	Effect	SE	p
Cross-sectional												
Naïve estimates												
Fearful	0.14	0.02	<.001	0.13	0.02	<.001	0.21	0.02	<.001	0.04	0.01	0.001
Anxious	0.14	0.02	<.001	0.12	0.02	<.001	0.19	0.02	<.001	0.05	0.01	0.005
Avoidant	0.01	0.02	0.541	0.01	0.01	0.306	0.04	0.02	0.04	-0.03	0.01	<.001
TL estimates												
Fearful	0.06	0.01	<.001	0.05	0.01	<.001	0.18	0.02	<.001	0.01	0.01	0.386
Anxious	0.05	0.03	0.003	0.05	0.02	0.01	0.17	0.03	<.001	0	0.02	0.97
Avoidant	0.01	0.01	0.24	0.01	0.01	0.248	0.05	0.02	0.009	-0.02	0.01	<.001
Longitudinal												
Naïve estimates												
Fearful	0.15	0.02	<.001	0.14	0.02	<.001						
Anxious	0.15	0.04	<.001	0.13	0.03	<.001						
Avoidant	0.02	0.02	0.201	0.02	0.02	0.267						
TL estimates (without W2 control)												
Fearful	0.07	0.02	<.001	0.06	0.01	<.001						
Anxious	0.06	0.02	0.007	0.05	0.02	0.022						
Avoidant	0.03	0.01	0.025	0.02	0.01	0.181						
TL estimates (with W2 control)												
Fearful	0.02	0.01	0.206	0.01	0.01	0.32						
Anxious	0.02	0.02	0.463	0	0.02	0.797						
Avoidant	0.01	0.01	0.375	0	0.01	0.65						

Note. TL = targeted learning.

Table 4.3 shows both the naive estimates as well as the estimates following the targeted learning steps. The naive estimates are essentially the correlations between the two variables without any control variables. We only describe the targeted learning results below. We use Ψ^* to denote the targeted learning estimates within the text. The estimates are scaled to be between -1 and +1 so an estimate can be understood in terms of percentages. For example, a Ψ^* = .10 means a 10% difference in the outcome between two groups. Note that we have opted to maintain a correlational language given that it makes the description of the results easier and readers are more used to interpreting this type of language. Thus, while the language used below is not explicitly causal, we are nonetheless intending for these quantities to be interpreted causally as a change in attachment style causing a corresponding change in the outcome.

We found that compared to secure individuals, fearful-avoidant individuals were higher in anxiety (Ψ^*= .06, $p < .001$), depression ($\Psi^* = .05$, $p < .001$), and loneliness ($\Psi^* = .18$, $p < .001$); but did not report engaging in more social distancing behaviors ($\Psi^* = .01$, $p = .386$). The results were similar for attachment-anxious individuals who were higher in anxiety ($\Psi^* = .05$, $p = .003$), depression ($\Psi^* = .05$, $p = .010$), and loneliness ($\Psi^* = .17$, $p < .001$) than secure individuals; but did not report engaging in more social distancing behaviors ($\Psi^* = .00$, $p = .970$). Avoidant individuals only differed from secure individuals in their social distancing behaviors: Avoidant individuals were significantly less likely to engage in social distancing behaviors compared to secure individuals ($\Psi^* = -.02$, $p < .001$).

4.3.2 Longitudinal Model

The results from the causal discovery algorithm for the longitudinal model can be found in Figure 4.2. The results were largely similar to the results of the cross-sectional model with attachment styles being suggested as putative causes for anxiety, depression, loneliness, and social distancing behaviors. Loneliness was again identified as a mediator between attachment styles and the other outcomes. However, social distancing behaviors were also identified as a mediator between loneliness and anxiety and depression. Attachment styles, anxiety, depression, loneliness, and social distancing behaviors at Wave 2 were also causes of depression and anxiety at Wave 3. Anxiety at Wave 3 was also identified as a cause of depression at Wave 3.

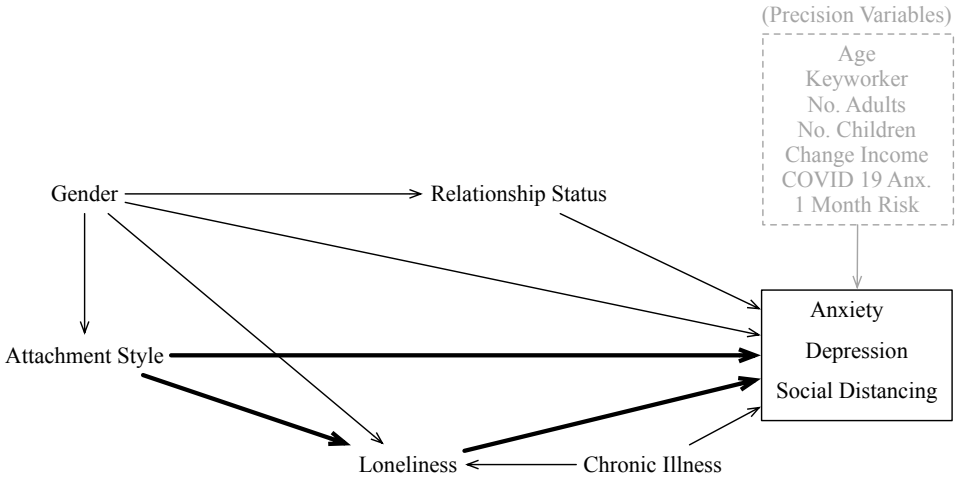
Table 4.3 shows both the naive estimates as well as the estimates following the targeted learning steps for the longitudinal analyses. The naive estimates do not differ between the analyses with and without controlling for time given that naive estimates are estimated without any control variables (including time). For the targeted learning, we used only the control variables that were needed to provide accurate causal estimates in line with the d-separation rules for causal inference. Based on these rules, we needed to control for age, gender, relationship status, keyworker status, number of adults in the household, number of children in the household, change in income, chronic illness, COVID-19 anxiety, and perceived one month risk. We did not include any mediators in the models as we were interested in the total effect of attachment styles on the mental health outcomes. We present the results for the longitudinal estimates with and without controlling for Wave 2 reports of the variables. The estimates without the Wave 2 control refer to how much we can still explain the mental health outcomes at Wave 3 by the participants' self-reported attachment style at Wave 2. The estimates including the Wave 2 control variables refer only to changes in the mental health outcomes from Wave 2 to Wave 3 as a result of attachment styles. However, we would not expect a fixed variable (i.e., a variable which is assumed to not change over time and does not in our models) to cause changes in an outcome over time but have included it in case readers are interested in this outcome.

We found that compared to secure individuals, fearful-avoidant individuals were higher in anxiety ($\Psi^* = .07$, $p < .001$) and depression ($\Psi^* = .06$, $p < .001$) at Wave 3. We also found that compared to secure individuals, anxious individuals were higher in anxiety ($\Psi^* = .06$, $p = .007$) and depression ($\Psi^* = .05$, $p = .022$) at Wave 3. Finally, avoidant individuals differed significantly from secure individuals only in that they reported higher anxiety ($\Psi^* = .03$, $p = .025$) but not higher depression ($\Psi^* = .01$, $p = .181$). Changes in anxiety or depression scores between Waves 2 and 3 were not significantly different between any of the groups.

4.4 Discussion

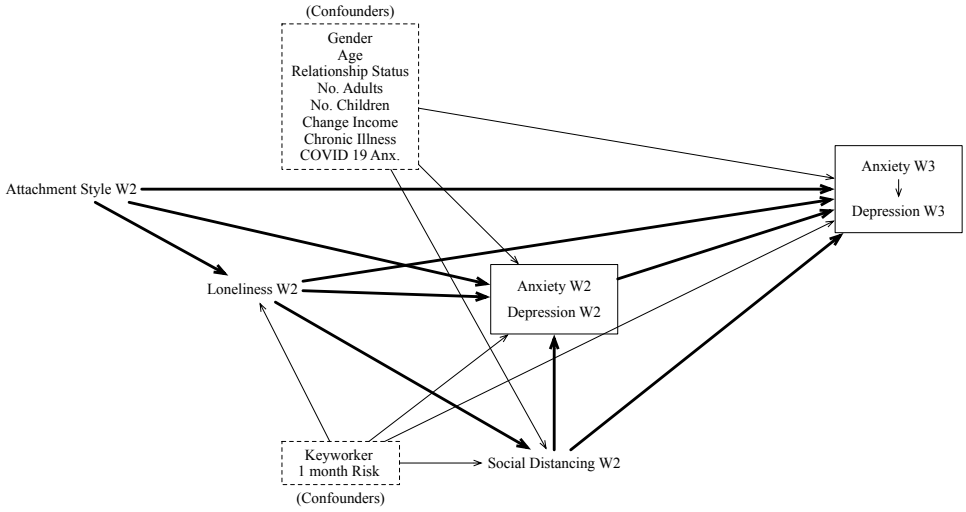
The purpose of the present study was to identify putative causal relationships between attachment styles, social distancing behaviors, and mental health outcomes. As hypothesized, attachment

Figure 4.1: Cross-Sectional Causal Discovery Results.



Note. The directed causal relationships between the cause and effect with a probability of at least 0.5 have been included in the graphs. We have highlighted the theoretically relevant relationships in bold but have included the required control variables in the graph. Precision variables are included in grey as they are not necessary to produce an unbiased estimate but can make the standard errors tighter because they explain variance in the outcome variables.

Figure 4.2: Longitudinal Causal Discovery Results.



Note. The directed causal relationships between the cause and effect with a probability of at least 0.5 have been included in the graphs. We have highlighted the theoretically relevant relationships in bold but have included the required control variables in the graph. There were no precision variables in the longitudinal model given that all the cross-sectional precision variables caused Wave 2 outcomes as well as Wave 3 outcomes meaning they introduced confounding in the data and needed to be included as controls.

insecurity was a risk factor for poor mental health during the COVID-19 pandemic. Specifically, individuals with fearful avoidant or anxious attachment styles were 5-6% higher in depression and generalized anxiety and 17-18% lonelier compared to secure individuals. Avoidant individuals did not differ in their depression or anxiety levels but were also significantly lonelier than secure individuals (albeit by a reduced margin, 5%). The differences in levels of depression and anxiety between attachment anxious and fearful avoidant individuals and secure individuals remained constant over time. This pattern is in line with recent research, which identifies attachment anxiety, rather than avoidance, as being a risk factor for ongoing mental health issues during the pandemic (Mazza et al., 2021; Moccia et al., 2020; L. Vowels, Carnelley, and Stanton, 2022). To this we would add that those individuals with a fearful avoidant attachment are similarly at risk.

The causal discovery algorithm also identified loneliness as a partial mediator of the causal path between attachment styles to mental health outcomes. We found that while depression and anxiety are higher in anxious and fearful avoidant individuals, they are almost four times higher in loneliness than anxiety and depression compared to secure individuals. Again, this pattern has been observed by other researchers; Vismara, Lucarelli, and Sechi (2022) showed that loneliness had a partially mediating role between attachment anxiety and mental health outcomes during the COVID-19 pandemic in a sample of Italian participants. These results suggest that loneliness is particularly prevalent among anxious and fearful avoidant individuals and any interventions that are designed to improve mental health outcomes for individuals with insecure attachment styles should focus on preventing and ameliorating loneliness.

Furthermore, we also examined which attachment styles were causally linked to adherence to social distancing guidelines. In contrast to our hypothesis, we found that avoidant individuals were significantly less likely to follow social distancing guidelines compared to secure. Indeed, recent research has also found the same pattern (Brulin et al., 2022; Lozano and R. Fraley, 2021). In our study however, there was only a small (2%) difference between avoidant and secure individuals, which is not likely to be meaningful behaviorally. Overall, our results suggest that while insecure individuals have worse mental health outcomes and feel lonelier compared to secure individuals, the causal relationship between attachment styles and social distancing

measures, while it exists, is very small.

The present study provided causal evidence of the link between attachment styles and mental health outcomes during the COVID-19 pandemic. The study was the first to our knowledge to use causal methods to examine these relationships. The results from the causal analyses corroborate previous correlational findings but provide a more accurate estimate of the effect size due to the use of targeted learning, which has been shown to produce estimates that are less biased compared to other methods (Luque-Fernandez et al., 2018; M. J. van der Laan and S. Rose, 2011). Furthermore, the data were drawn from a nationally representative UK survey and the results were estimated both cross-sectionally and over time. Thus, we expect that the effect size estimates are a relatively accurate estimate of the real average estimate in the population.

However, there are also several limitations that should be considered. While we were able to establish causal relationships between our variables of interest using state-of-the-art causal discovery and causal inference algorithms, we did not have access to pre-pandemic data. Thus, we were only able to establish the causal relationships between attachment styles, social distancing behaviors, and mental health outcomes during the COVID-19 pandemic but we do not know whether insecure individuals were particularly at risk due to the pandemic or whether they already had higher levels of mental health problems before the pandemic occurred. One study to our knowledge has examined changes in mental health outcomes as a result of the pandemic with pre-pandemic and early pandemic data (L. Vowels, Carnelley, and Stanton, 2022) and showed that individuals higher in attachment anxiety were particularly at risk of worse mental health outcomes over time. However, attachment anxiety has also been linked to worse mental health in general, not just during a pandemic (Mikulincer and Shaver, 2016). What is clear is that attachment insecurity is causally linked to poorer mental health outcomes, especially to loneliness.

Another limitation of the study relates to the measure of attachment. As is the case with most large datasets in nationally representative samples, the choice of variables is limited to what is available in the dataset. In our case, the only measure of attachment styles was categorical and measured attachment on two dimensions: attachment anxiety and attachment avoidance. This meant that participants were forced to place themselves into one of the four categories but there

may be a great deal of heterogeneity within the categories. Arguably a categorical measure of attachment is less likely to suffer from shared method variance with the outcome variable (i.e., be closely correlated because of some third variable such as mood on the day) and thus the actual relationship between the variables is more likely to be indicative of a real causal effect rather than a non-causal correlation due to a third variable. This argument is also supported by the fact that the estimates were the same strength within and between timepoints. However, it is not clear which measures provide a more accurate depiction of attachment styles overall (R. Fraley and Shaver, 1998) and the results may not be directly comparable to other studies which use a continuous measure of attachment.

Finally, the validity of the estimates proposed to correspond with causal quantities relies on four key assumptions generally described in the causal inference literature (Imbens and D. Rubin, 2015; Pearl, 2009; Pearl, 2012). The first assumption is that our theory and graph are correct. Of course, this is a strong assumption, and one which is ideally validated under experimental conditions. The second assumption is known as ignorability (also known as conditional exchangeability), which is closely related to the first, and is the assumption that there exist no unobserved confounders which otherwise bias the effect size estimation. The third assumption is that of positivity, which is the assumption that there exist a sufficient number of people in each attachment style group to adequately estimate the effects (i.e., the probability of being in each comparison group is positive / non-zero). The last is known as the Stable Unit Treatment Value Assumption (SUTVA), which is the assumption that the participants are independent of one another given their control variables (i.e., that the participants do not influence one another). We expect that the latter two assumptions hold in our sample given the relatively large sample size (which helps with positivity) and the participants being independent of one another (which helps the SUTVA). However, it is more difficult to establish whether the first two assumptions hold. We discussed the variables that were included in the study thoroughly among experts in attachment and mental health research and validated the theoretical variables using a causal discovery algorithm. We cannot, however, be certain that there are no unobserved confounders that we should have controlled for. Other researchers may disagree with the variables included and we encourage them to engage in the process of refining our causal theoretical model.

In conclusion, the present study provided causal evidence of the relationship between attachment insecurity and mental health outcomes during the COVID-19 pandemic. Specifically, we showed that anxious and fearful avoidant individuals have higher scores in depression, generalized anxiety, and loneliness whereas avoidant individuals have higher scores in loneliness compared to secure individuals. The results of the study imply that focusing on improving feelings of loneliness and isolation in insecurely attached individuals can help ameliorate mental health symptoms in this population. Many countries introduced lockdown and social distancing measures during the pandemic and these measures are in place periodically in different areas of the world. However, given the burden of social isolation among insecure individuals, it is likely that these measures exacerbate feelings of loneliness. Thus, finding ways to support and maintain social connection is likely to be crucial in ameliorating mental health problems in the population.

4.5 Supplementary Material

4.5.1 A Note on Causality from Cross-Sectional Data

In this work, we used causal discovery and causal inference methods on cross-sectional data (as well as longitudinal data). We understand that this is unusual, particularly in the domain of psychology, and so take the opportunity below to discuss our reasoning.

One of the requirements for causality is that the cause precedes the effect (Pearl, 2009). In other words, in the causal DAG $X \rightarrow Y$, the change in X occurs before the subsequent change in Y. Researchers have, understandably, generalised this rule of causality to data collection, such that if, for example, one wants to estimate the strength of the causal effect between X and Y, one has to measure Y *after* one measures X. Indeed, in causes where X and Y are continually interacting over time - *vis a vis* a medical drug, which *requires* X to occur at some distinct point in time to be able to generate any subsequent effect on Y - it seems especially pertinent to always ensure the measurement of Y occurs after any measurement of X to allow for the cause-effect relationship to occur. Before we get into the details about why this may or may not always be a pertinent consideration when deciding between cross-sectional and longitudinal data, we

should recall that techniques in causal inference and causal discovery disentangle notions of longitudinal/cross-sectional from the features of the data necessary for the associated (causal) task. These features are encoded by conditional independencies, distributional asymmetries, and dynamical causal fingerprints of the joint distribution (M. Vowels, N. Camgoz, and Bowden, 2022; Pearl, 2009; J. Peters, Janzing, and Scholkopf, 2017; Sugihara et al., 2012; M. Vowels, N. Camgoz, and Bowden, 2021). Indeed, there is not always a clear advantage for longitudinal data over cross-sectional data (as commonly assumed in psychology), whereas there are often advantages of statistical power and accuracy of measurement which apply to cross-sectional data.

Based on the temporal requirement separating cause from effect, it would seem that for most, if not all, phenomena in psychology, one should only ever employ longitudinal methodologies for speaking about causal effect estimation. Unfortunately, the time delay between many psychological effects are *completely unknown*. For instance, for any given change in sexual satisfaction in a romantic relationship, how long does it take for the corresponding effect on relationship satisfaction to occur? Furthermore, many of these effects could be argued to be, for all intents and purposes, continually reciprocal. In other words, sexual satisfaction has an almost instantaneous effect on my relationship satisfaction which, in turn, immediately affects the following levels of sexual satisfaction. The summary graph (no longer a DAG) would be $X \leftrightarrow Y$ (Sugihara et al., 2012). Alternatively, the cause-effect time delay may be different for different couples, and/or take an arbitrarily short or long period of time. For instance, it may take a week's worth of low perceived partner support for an effect to manifest in my relationship satisfaction, and two weeks for my partner's. If we had a strong theory for the delay being invariably equal to a week (notwithstanding the potential invalidity of this assumption), we might delay the measurement of Y by a week, following the measurement of X. The two situations - implicating either approximately reciprocal causality, or heterogeneous and arbitrary temporal delay between cause and effect - already make it clear how extremely challenging it is to establish a clear 'best option' for the data collection methodology. Collecting the follow-up at the 'wrong' point in time, may render the estimated effect sizes completely meaningless. This is trivial to demonstrate, and was discussed by M. J. Vowels, L. M. Vowels, and N. Wood (2021).

A possible solution to the difficulty associated with not knowing the time lag of the effect would be to collect data at several time points. Whilst this is indeed possible, it merely shifts the nature of the challenge. When confronted with a series of measurements, one has to either identify whether there exists a single optimal lag (*e.g.*, the effect occurs somewhere between the 4th and 5th timepoint), or to somehow aggregate the causal effect across multiple timepoints, blurring the nature of the cause relative to the auto-correlative effect of the outcome. In addition, this assumes that measurements are taken regularly enough that difficulties with fluctuation and newly introduced confounding do not impact the estimation (see, again M. J. Vowels, 2021; M. J. Vowels, L. M. Vowels, and N. Wood, 2021). These challenges are particularly evident in causal inference applied to dyadic interactions. Even if one can identify (*e.g.*, behavioural) moments from one partner in an interaction which are *predictively* salient of an outcome, it is not clear how these moments cause the outcome over their auto-correlative effects, or over the interaction of these behaviours with those of the partner's. Indeed, one can undertake measurements on a per-microsecond basis, thereby guaranteeing sufficient measurement precision (at least temporal precision), but this does not help us overcome the inherent explosion in complexity. Finally, the researcher would be forced to collapse the dataset down, losing both precision and meaningfulness - its quite possible that there is no *single* meaningful cause-effect relationship.

Going further, let's now make the two additional assumptions discussed above (which may not hold) to make our lives easier, and see how far it takes us: (a) We assume that the cause-effect relationships are not reciprocal, and (b) we assume that there does indeed exist a theoretically justifiable and fixed delay of one week between cause and effect. This would suggest that all one needs to do is to measure Y one week after measuring X. Unfortunately, this creates a myriad of new problems relating to hidden confounding which can easily render the causal effect of interest unidentifiable. This situation is illustrated in Figure 2.5. In the figure, the hidden variable at each timepoint, confounds the estimation of the effect over time of X on Y. Furthermore, longitudinal studies are much more challenging logistically, and often lead to significant participant dropout and censorship.

As such, even under the two substantially simplifying assumptions stated above, the choice for longitudinal data is not clear, and altogether we face the following challenges: (1) longitudinal

data are more expensive to collect than cross-sectional data, leading to problems of dropout and censorship and smaller sample sizes; (2) longitudinal data do not necessarily mitigate problems with hidden confounding, but create their own challenges; (3) the choice of time delay between cause and effect may be arbitrary, heterogeneous, or somewhat reciprocating for many psychological constructs, leading to meaningless effect size estimates regardless of estimand identification. Of course, in an ideal world, the first point about cost would not be relevant - if the only options before us are to either perform a correct but costly experiment, or one which is sub-optimal, we should perhaps cancel the experiment. However, in reality researchers must make a decision, and given that *all* studies will be suboptimal in one sense or another (confronted, as we always are, with numerous simultaneous practical compromises), trading of cost and its associated impact on sample size and statistical power *must* play a role in our decision. Specifically - for any given budget, if we are forced to choose between a longitudinal study with no additional causal benefits, and a cross-sectional study with much higher statistical power but the same causal disadvantages, it is reasonable to choose the latter. Thus, whilst on the one hand, we accept that longitudinal data make sense for causality, it appears that they often do us no real favours in practice.

So what is the case for cross-sectional data? Well, actually it faces many of the same challenges. This is especially the case in psychology where almost everything correlates with everything else (Meehl, 1990; Orben and Lakens, 2020) leading to a near impossibility of ruling out hidden confounding and resulting in meaningless effect size estimates (M. J. Vowels, 2021). As such, it is difficult to argue that there exists a clear benefit for longitudinal data over cross-sectional data when we are practically doomed to non-identification in both cases. Putting longitudinal and cross-sectional data collection on a similar footing with respect to the possibility of identification behooves us to seek other justifications for choosing one over the other. In particular cases, for instance (and these cases overlap strongly with the phenomena under study in this particular chapter) there exist some justifiable cases where cross-sectional data can still serve to yield equivalent levels of (un)identification to longitudinal data. Many variables can be measured without worrying about the temporal aspects of cause and effect. For instance, 'report the age at which you got married' does not (at least not generally) change over a course of days or weeks, and one can measure this almost any time after the person got married. Other variables

which might be assumed to be relatively stable, thus yielding flexibility to the data collection, include things like attachment style, personality traits, or gender. If one is interested in the effect of a hypothetical change in gender on, for instance, current levels of depression and anxiety, one gains nothing from collecting the two variables longitudinally. Of course, if one suspects that there may be some contemporaneous effect of (for example) retrospection on the measurement of an otherwise stable trait, one can argue that longitudinal data may be required, if only to acquire a more robust measure of that trait by averaging the measurements over time. Alternatively, if it is precisely one's suspicion that attachment (or whatever) *is not* a stable trait, then, by definition, it does not fall into the category of traits hereby considered. Otherwise, the justification for cross-sectional data is simple - it is comparatively abundant, logistically simple, and inexpensive to collect. This allows us to divert resources otherwise used for participant follow-ups towards the utilisation of better quality measures and/or larger sample sizes.

Finally, it is worth noting that the empirical samples for the variables and the conditional independencies which exist between them, as reflected in the joint distribution as consequence of the underlying causal (and therefore temporally ordered) structure, do not themselves care about when they were collected. Indeed, it is not necessary to tell a causal discovery algorithm *when* the data were collected, because the causal fingerprints in the data themselves tell this story, and this will be reflected in the putative graph. With our use of SAM, for example, we identified a putative graph *based on the causal fingerprints in the data*, and these fingerprints either tell us whether X cases Y, or whether Y causes X, or whether the direction is unknowable or ambiguous. Of course, and as discussed in the previous paragraph, if the conditional independencies in data which have been recently collected are supposed to reflect a set of conditional independencies from years before (as might be assumed, for example, for structures between trait variables such as attachment), but the person's recent reporting of their attachment is somehow biased/coloured by recent experience, causal discovery may yield biased results in turn. However, if the trait itself is so sensitive to this kind of measurement error, one may have similar concerns for its reliability in a longitudinal context, too. Indeed, herein lies the tradeoff between acquiring larger sample sizes with possibly longer questionnaires in a cross-sectional designs (whilst also inheriting additional error due to, for example, retrospective problems with a participant's memory of past events), and shorter questionnaires, temporal sources of

confounding, smaller sample sizes, and participant censorship in longitudinal designs.

In summary, given (a) that there is nothing inherently about data being cross-sectional that precludes, at least in principle, causal discovery or causal inference (assuming that required consequences of the causal process manifest, as needed, as conditional independencies and/or distributional asymmetries, for example), and (b) that longitudinal methodologies cannot be anymore guaranteed than cross-sectional methodologies to yield identification (in the face of temporal forms of confounding and ambiguities regarding the arbitrary delays between cause and effect), and (c) that cross-sectional data may be higher in quality for the same associated cost as longitudinal data, it seems unreasonable to necessarily prioritise, as a matter of principle, longitudinal data over cross-sectional data. At the very least, we would recommend researchers consider each option based on its own merits, in the context of the specific research questions and phenomena under study.

4.5.2 Details of Measures Included in the Study

Attachment style. Attachment style was measured using the Relationships Questionnaire (Bartholomew and Horowitz, 1991) which included four statements, one for each attachment style. The participants were asked to "Place a checkmark next to the letter corresponding to the style that best describes you or is closest to the way you are". The statements included comfort with emotional closeness and dependence on others. (Secure = It is easy for me to become emotionally close to others. I am comfortable depending on them and having them depend on me. I don't worry about being alone or having others not accept me. Fearful avoidant = I am uncomfortable getting close to others. I want emotionally close relationships, but I find it difficult to trust others completely, or to depend on them. I worry that I will be hurt if I allow myself to become too close to others. Anxious = I want to be completely emotionally intimate with others, but I often find that others are reluctant to get as close as I would like. I am uncomfortable being without close relationships, but I sometimes worry that others don't value me as much as I value them. Avoidant = I am comfortable without close emotional relationships. It is very important to me to feel independent and self-sufficient, and I prefer not to depend on others or have others depend on me).

Social distancing behaviors. Social distancing practices, in accordance with government guidelines during the first UK lockdown, were assessed using a list of 16 statements with respect to the past week, e.g., "Met up with friends or extended family (outside of your home)." Response scales were: not at all; 1-2 days per week; 3-4 days per week; most days; every day. The social distancing items were coded such that higher scoring reflected greater endorsement of social distancing practices, e.g., the item "Engaged in close contact greetings with people outside of your family (e.g., shaking hands, hugging)" was reversed-scored. We performed an exploratory factor analysis to examine the scale items. The results showed that 10/16 variables loaded well on one factor and were thus included as a total score. Variables that did not cluster well with others included variables about keeping a distance, washing hands straight away, and behaviors that were within guidelines. The reliability of the 10-item scale was α = .91.

Generalized Anxiety. Symptoms of GAD were measured using the Generalized Anxiety Disorder 7-item Scale (GAD-7; Spitzer et al., 2006). The GAD-7 has been shown to produce reliable and valid scores in community studies, and the reliability in the current sample was α = .94.

Depression. Depression was measured using the Patient Health Questionnaire PHQ-9: Kroenke, Spitzer, and J. Williams, 2002). The PHQ-9 is a 9-item self-report measure that asks participants the degree to which they have been bothered by depressive symptoms in the last two weeks (items are rated on a 3-point Likert scale ranging from 0 [not bothered at all] to 2 [bothered a lot]). Multiple previous studies attest to the reliability and validity of the PHQ-9 (Hinz et al., 2017). The reliability of the scale in the current sample was α = .93.

Loneliness. Loneliness was measured using a 3-item Loneliness Scale (Hughes et al., 2004). Example items include "How often do you feel that you lack companionship?" The items were measured on scale from 1(Hardly ever), to 2 (Some of the time), to 3 (Often). The reliability of the scale in the current sample was α = .87.

4.5.3 Control variables.

We also included a set of variables that were theoretically causally related to the central variables in the study that we controlled for in the models. These variables include demographics, COVID-

19 related anxiety and perceived one month risk, and hygiene practices and are described below in more detail.

Demographics. The following demographic variables were measured at Wave 2 and included in the analyses: age, gender, relationship status, key/essential worker status, number of adults living in household, number of children living in household, change in monthly household income during pandemic, and currently pregnant – self (partner). Religion, ethnicity, employment status, and education were measured at Wave 1 and only used for descriptive purposes.

COVID-19-related anxiety and perceived one month risk. We also included COVID-19 related anxiety and perceived one month risk in the analyses as control variables. The survey included a question "How anxious are you about the coronavirus COVID-19 pandemic?". The one-month risk included a question "What do you think is your personal percentage risk of being infected with the COVID-19 virus over the following time periods? - In the next month". Both items were rated on a 'slider' (electronic visual analogue scale) to indicate their degree of anxiety/perceived risk with '0' and '100' at the left- and right-hand extremes respectively, and 10-point increments. This produced continuous scores ranging from 0 to 100 with higher scores reflecting higher levels of COVID-19-related anxiety or higher perceived risk.

Hygienic practices. Reasons for maintaining hygiene practices included 18 self-reported statements (e.g., "I knew about why it was important and had a clear idea about how the virus was transmitted" and "I was able to overcome the physical and/or mental barriers that might have stopped me from doing it"). Response scales were 1 (strongly disagree) to 5 (strongly agree). We performed an exploratory factor analysis to examine the scale items. The results showed that 11/18 variables loaded well on one factor and were thus included as a total score. Variables that did not cluster well with others included variables that focused on reminders and support and social pressure to engage in hygiene behaviors. The reliability of the 11-item scale was α = .93.

4.5.4 Full Description of the Data Analysis

Missing Data As mentioned in the main text, we analysed only data for which the key causes and outcomes were available. In the cross-sectional case, we list-wise deleted according to

missing values in the attachment style (wave 2) variable, and in the longitudinal case, we list-wise deleted according to missing values in the attachment style (wave 2), or depression (wave 3), or anxiety (wave 3) variables. The amount of missing data following this for the cross-sectional case was 5.7%, and for the longitudinal case was 32.43% out of 1406 participants total.

Furthermore, we ran two logistic regressions (one for the cross-sectional and one for the longitudinal cases) to see which demographic variables were associated with missingness. For this we constructed two dummy outcome variables. The first was a set of binary labels indicating whether the attachment style variable was missing (=1) or not (=0) for the cross-sectional case, and the second was a set of binary labels indicating whether any of the attachment style (wave 2), or depression (wave 3), or anxiety (wave 3) variables were missing for the longitudinal case. In both logistic regression models, the predictors were as follows: W2_Children_household, W2_Keyworker, W2_Chronic_illness_self, W2_Relationship, W2_COVID19_anxiety, W2_Hygiene_total, W2_Dep_Total, W2_Change_Income, W2_Risk_total, W2_Age_year, W2_Adults_household, W2_Pregnant, W2_GAD_Total, W2_Gender, W2_RISK_1month, W2_Loneliness_Total.

Assuming an alpha of 0.05, in the cross-sectional model, the only predictor which was a significantly predictive of missingness in attachment style were W2_Hygiene_total (B=-0.0472, p=0.000). For the longitudinal model, the predictors which were significantly predictive of missingness in either attachment style (wave 2), or depression (wave 3), or anxiety (wave 3), were W2_Children_household (B=0.2245, p=0.003), W2_Dep_Total (B=0.0468, p=0.020), W2_Age_year (B=-0.0266, p=0.000), and W2_Gender (B=0.2340, p=0.037).

Whilst the cross-sectional missingness is, at least based on these results, relatively minor, the longitudinal results indicate more significant potential issues. It is generally known that missingness has important structural and causal attributes (indeed, causal inference is, itself, a missing data problem) (Shpitser, Mohan, and Pearl, 2015). As such, future work should consider how missingness affects the results of the analyses in the current work.

Data Analysis The code for the associated analysis can be found here: `https://github.com/matthewvowels1/attachment_COVID` We used a state-of-the-art causal discov-

ery algorithm known as Structural Agnostic Modeling (SAM; Kalainathan et al., 2020). The algorithm is based on a Generative Adversarial Network (I. J. Goodfellow et al., 2014) in which one neural network (the adversary) proposes estimations of conditional distributions, and another one (the discriminator) tries to distinguish the estimates from the true data. During optimization, the adversary learns to approximate the true distributions such that the discriminator fails to distinguish the estimates from the originals. The overall model derives a structure that maximizes the fit to the data, whilst enforcing a constraint which encourages acyclicity (no feedback loops in the resulting graph), and a constraint which encourages sparsity.

SAM also takes advantages of a number of structural heuristics which can be used to orient cause-effect directions, thus improving over other contemporary approaches for which model fit statistics are known be insufficient for estimating causal directionality (Pearl, 2009; M. Vowels, N. Camgoz, and Bowden, 2022). By leveraging these heuristics, SAM is thus able to estimate the causal structure of a set of variables under a number of assumptions. The algorithm does not, for example, infer hidden variables for us. In the presence of unobserved confounding, the algorithm may therefore mistake the direction of a causal effect. Nonetheless, it can be used to 'fill in the gaps' of our theories, by proposing structures about which we may have no prior domain expertise. These techniques are not meant to overrule our domain expertise, and so we must evaluate the putative structure for face validity. The validity of the subsequent analysis, which itself is informed by the structure, rests on the assumption that this structure is sufficiently correctly specified. Of course, in reality there may exist some key factors without which our analyses become biased. However, given that we are using causal discovery to specify our structure in addition to domain expertise (whereby the latter is usually applied on its own) we would argue that this approach helps us to robustify the specification of our model and therefore our analysis. In particular, knowledge of the structure helps to guide us when it comes to the selection of good control variables for the estimation of causal effects, and it is well known that the selection of 'bad' control variables can have a dramatic impact on the resulting estimates (Cinelli, Forney, and Pearl, 2022; M. J. Vowels, 2021).

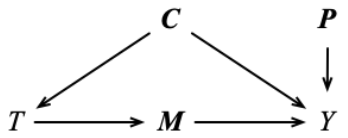
With regards to the specifics of the analysis, we used SAM to infer the cross-sectional structure for Wave 2 (17 variables and 1325 participants), as well as the longitudinal structure across

Wave 2 and Wave 3 (19 variables from 895 participants). We included all variables that were expected to be causally linked to the main variables of interest and thus affect the estimation of the causal relationships. We applied a constraint preventing the discovery of causal effects backwards in time, as well as constraints preventing causal links between certain demographics: age and gender cannot be effects; change in income was measured as the change between Waves 1 and 2 and thus was prevented from affecting all demographic variables.

SAM is a continuously optimized method and can be randomly initialized with a set of starting parameters. This means it does not necessarily converge to the same solution for each initialization. We therefore fit SAM 50 times and took the consensus across the 50 resulting structures. We explored a number of learning rates (0.01 and 0.001), and applied a regularizing penalty of 0.05 which encourages the structure to be acyclic. We found that a learning rate of 0.001 did not converge (the resulting structure was fully saturated), and therefore used the results derived using a learning rate of 0.01. Otherwise, the default hyperparameter settings for SAM were used.

The output of SAM is a Directed Acyclic Graph, which is a structured / graphical model encoding the directions of causal influence without cycles. This structure was used to construct a Structural Equation Model (SEM) using the lavaan package in R. We used only observed variables rather than constructing latent variables of our constructs given the causal discovery algorithm was conducted with observed variables only. The numbers of paths or 'edges' between variables in the cross-sectional and longitudinal graphs were prohibitively high, and the SEM with all variables identified as causally linked to a part of the model would not converge. We thus employed 'd-separation' rules (Koller and Friedman, 2009; Spirtes, C. Glymour, and Scheines, 2000; Pearl, 2009) to reduce the complexity of the graph without impacting the essential structure. An example of the application of the rules can be simply demonstrated by considering the structure $A \rightarrow B \rightarrow C$. Imagine we are concerned with estimating the effect of B on C, then there is no need to estimate the effect of A on B, and the path $A \rightarrow B$ can therefore be removed from the model. This is because of what is known as the Markovicity assumption, which tells us that knowing A tells us nothing about C which is not already contained in B. Formally, the statement is that A is independent of C given B. Similarly, if we are only interested

Figure 4.3: A Directed Acyclic Graph depicting the various components for consideration.



Note. We are interested in the effect of T on Y, where Y is the outcome variable; T is the treatment variable; C is a set of confounders (which must be controlled for in the model); P is a set of precision variables (which do not have to be included but which help explain variance in Y and which can therefore improve estimation precision); M is a set of mediators (which should not be included and which can be ignored unless they are of central importance to the research question).

in the effect of A on C, we can ignore B - a process known as projection (C. Glymour, 2001). These rules can be applied to all the paths in the full graph and used to identify non-causal paths which otherwise affect the estimation of the paths we care about. They can also be used to identify what are known as 'precision variables' which may help in improving the precision of estimation (i.e., to reduce the standard error). Unlike confounders, which are essential to control for, precision variables do not help us debias the estimate of the causal effect. Using these d-separation rules, we can therefore identify variables that are important and variables that can be ignored, and as a result the graph can then be simplified to leave only what is necessary to answer our research questions. Figure 4.3 illustrates an example structure comprising a cause T, an outcome of interest Y, and sets of confounders C, mediators M, and precision variables P, as a Directed Acyclic Graph (DAG). Here, we use bold to indicate sets of multiple variables. We can use the putative structure from the causal discovery stage to identify these variables, and therefore specify our model. SEM is a linear estimator and the causal discovery process used was non-linear and given the large number of variables in the models the fit of the SEM was poor. Thus, the SEM results including all mediations are presented in Tables 4.4 and 4.5.

In addition to SEM, we also used a state-of-the-art method at the intersection of machine learning and causality known as targeted learning (M. J. van der Laan and S. Rose, 2011) which has seen myriad applications and demonstrates across a range of subdomains in epidemiology and biostatistics (H. Li et al., 2022; Luque-Fernandez et al., 2018; Schnitzer et al., 2014). Interested readers are encouraged to consult the accessible introduction by Luque-Fernandez et al. (2018), but essentially the targeted learning frameworks provide us with a means to estimate causal effects of interest without having to make unreasonable assumptions about the

functional or parametric form. The effect itself must be unambiguously specified according to the required confounders and precision variables identified from the graph. Above we provided a list of all included confounders and precision variables used as part of the targeted learning analysis. Targeted learning involves the use of an ensemble of flexible and diverse machine learning algorithms or 'learners' to derive an initial estimate for a target causal effect. The ensemble is known as a SuperLearner (M. van der Laan, Polley, and Hubbard, 2007). It derives estimates from each of the individual learners and, via a process known as k-fold cross-validation, estimates a set of weights across these learners which are used to derive a final linear, weighted combination from each learner. The SuperLearner has been shown to exhibit several desirable properties relating to its optimality and performance. It is more accurate than using any one algorithm alone as it takes a weighted average of many different machine learning algorithms. For modeling the causal effect, we used the following learners for continuous outcome variables: Elastic Net (Zou and Hastie, 2005), Support Vector Regressor (Platt, 1999), linear regressor, linear regressor with quadratic features and moderation effects, Random Forest regressor (Breiman, 2001a), a Multilayer Perceptron regressor (I. Goodfellow, Bengio, and Courville, 2016), and an AdaBoost Regressor (Drucker, 1997).

The process of targeted learning involves an update step that removes a residual bias associated with the causal estimate derived using the Super Learner, and also renders a Gaussian distribution of estimates which is therefore amenable to the derivation of confidence intervals and p-values. The update step requires a second Super Learner model for the cause itself, known as a propensity score model. The propensity score is the likelihood of receiving treatment, and this quantity can be used to help us remove confounding associated with treatment group imbalance. In our case, we were interested in the effect of attachment style (a categorical variable) on a number of continuous outcomes, and thus one can consider attachment style to be equivalent to the treatment in our cause-effect model. The propensity score Super Learner comprises the following algorithms for categorical outcomes: a logistic regressor, a logistic regressor with quadratic and moderation effects, a MultiLayer Perceptron, a Random Forest classifier, a Support Vector classifier, and an AdaBoost classifier. The propensity model is used to generate predictions for the probability of being in a particular attachment category, from a set of predictors. Using the propensity scores, we can derive what is known as a 'clever

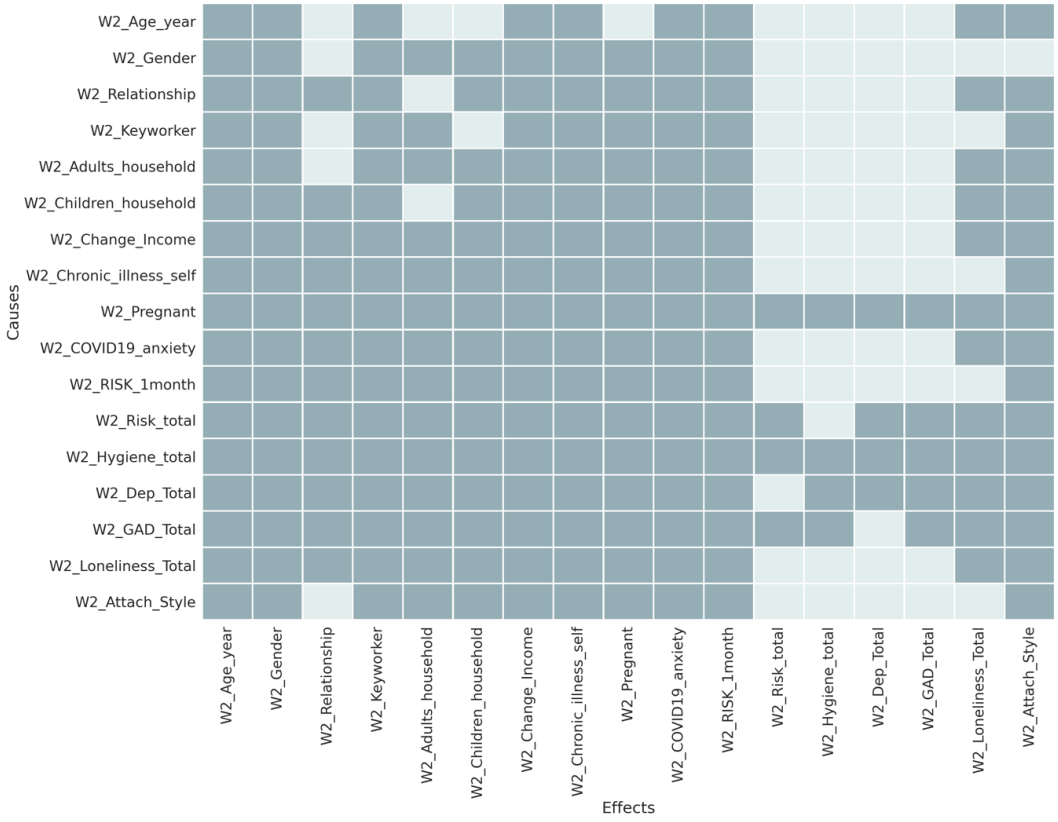
covariate' which quantifies the degree to which the estimate of the causal effect of interest is being biased by the relationship between a set of covariates and the cause. Definitions of clever covariates falls beyond the scope of this paper, but again, interested readers are encouraged to consult Luque-Fernandez et al. (2018). Once we have modeled this bias, we can correct for it by updating the initial estimate. It also provides us with a means to derive the Influence Function (Hampel, 1974; Hines et al., 2021; M. Vowels, Akbari, et al., 2023), which in turn is used to undertake valid statistical inference, despite the fact that our original estimates were derived using non-parametric methods. The power of the targeted learning approach is thus threefold: We can use powerful non-parametric machine learning algorithms to achieve high precision estimates; we can undertake typical statistical inference; and the update step removes residual bias thus improving the estimate. All algorithms in the Super Learner were implemented using the default implementations in the sklearn package (Pedregosa et al., 2011).

We concern ourselves with the estimation of the Average Causal Effect (ACE), which is the average difference in outcomes for participants of different treatment groups. For instance, the ACE for people in group 1 compared with group 0 can be expressed as:

$$\Psi_{1,0} = \mathbb{E}\left[\mathbb{E}[Y|C, P, T = 1] - \mathbb{E}[Y|C, P, T = 0]\right]$$

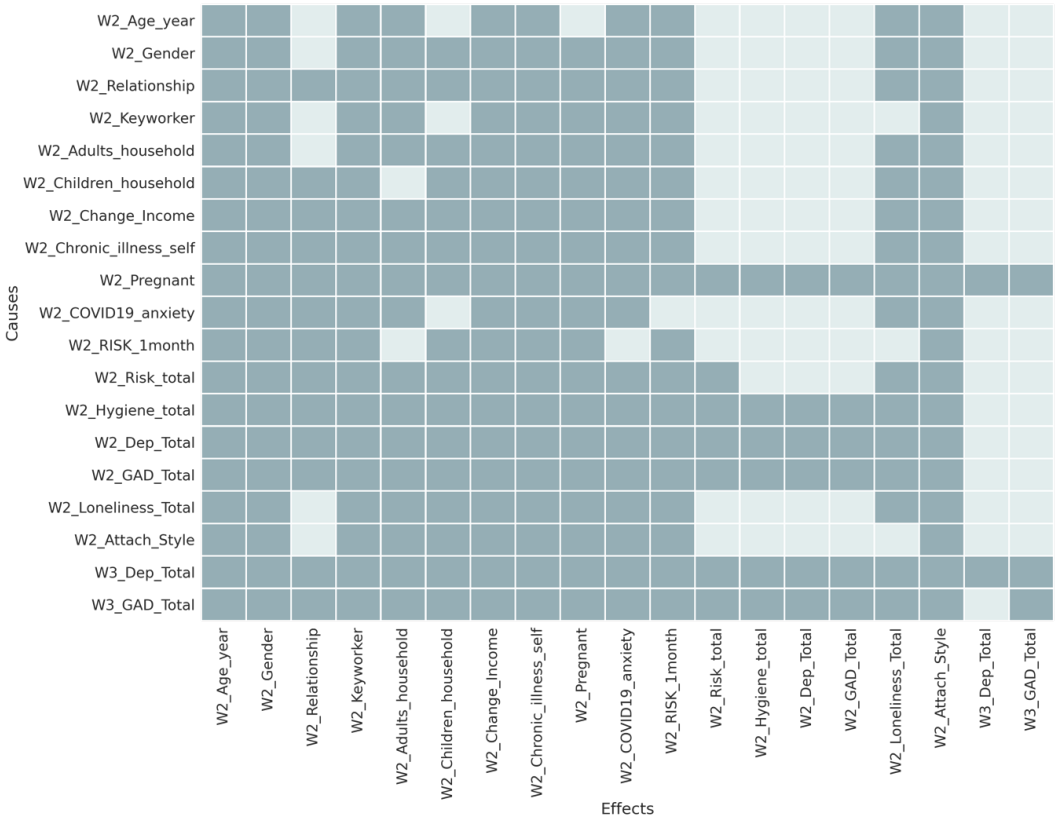
where $\mathbb{E}$ indicates an expectation operator and the bold font denotes that these may be sets of multiple variables. As is usual in causal inference, the validity of our estimates for rests on three key assumptions (Pearl, 2009; Imbens and D. Rubin, 2015): (1) ignorability - we assume that we have sufficiently controlled for confounding such that we can assume that there exist no remaining unobserved confounders, (2) positivity - we assume that the probability of having any attachment style is bounded away from 0 for all participants, and (3) stable unit treatment value assumption - we assume that the outcomes for each participant are independent of the outcomes for any other participant.

Figure 4.4: Cross-Sectional Results for the Causal Discovery Algorithm



Note. The causes can be found on the Y axis and effects on the X axis. The boxes in lighter colors identify a directed causal relationship between the cause and effect with a probability of at least 0.5. For example, we see from the figure that the only cause for attachment styles is participants' gender (there is a lighter colored box with gender on Y axis and attachment style on X axis) whereas attachment styles cause relationship status, anxiety, depression, loneliness, and social distancing behaviors. Attach Style = attachment style; Dep Total = depression, GAD Total = generalized anxiety, Risk total = social distancing behaviors.

Figure 4.5: Longitudinal Results for the Causal Discovery Algorithm



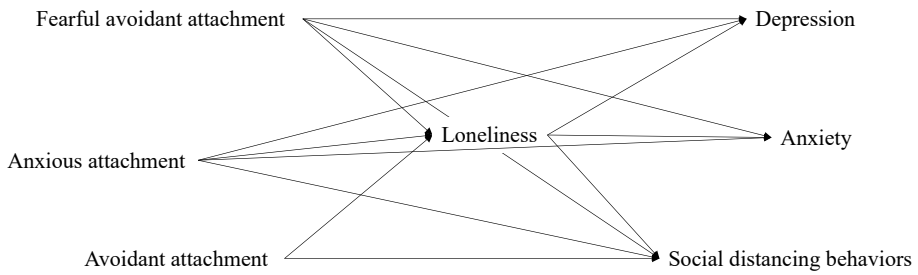
Note. The causes can be found on the Y axis and effects on the X axis. The boxes in lighter colors identify a directed causal relationship between the cause and effect with a probability of at least 0.5. Attach Style = attachment style; Dep Total = depression, GAD Total = generalized anxiety, Risk total = social distancing behaviors.

4.5.5 Structural Equation Modeling (SEM) Results

In addition to the targeted learning approach, we also conducted a more traditional structural equation modeling (SEM) approach which also allowed us to estimate loneliness as a potential mediator. However, SEM is limited in its linear assumptions and given the causal estimates were derived using a non-linear model, the results of the SEM are thus presented in the supplemental file. The full results of the SEM model can be found in Table S1 and the significant results of the relevant variables in Figure 4.6. The model fit the data well: $\chi^2(12) = 35.99$, $p < .000$, CFI = 0.99, TLI = 0.96, RMSEA = 0.04. Based on a sensitivity power analysis, we had a power of .80 to detect a minimum effect size of r = .12 and a power of 1.00 to detect a medium effect of r = .30 with an alpha level of .05. We found that compared to secure individuals, fearful-avoidant individuals were higher in anxiety (B = 1.19, p = .001), depression (B = 1.10, p = .001), and loneliness (B = 1.28, $p < .001$); and reported engaging in more social distancing behaviors (B = 1.28, $p < .001$). The results were similar for attachment-anxious individuals who were also higher in anxiety (B = 1.35, p = .003), depression (B = 1.18, p = .014), and loneliness (B = 1.11, $p < .001$); and reported engaging in more social distancing behaviors (B = 1.42, p = .013) compared to secure individuals. The results for avoidant attachment looked somewhat different with no significant differences between avoidant and secure individuals on anxiety (B = 0.07, p = .825) or depression (B = 0.11, p = .741). Avoidant individuals were higher than secure individuals in loneliness (B = 0.29, p = .014) and reported engaging in less social distancing behaviors (B = -1.01, p = .009).

There was a significant mediation by loneliness on the causal relationship between attachment styles and the other four outcomes (depression, anxiety, risk, and hygiene behaviors). There was a significant indirect effect through loneliness between fearful-avoidant attachment and anxiety (B = 1.74, $p < .001$), depression (B = 2.23, $p < .001$), and social distancing behaviors (B = 0.40, p = .001). There was also a significant indirect effect through loneliness between anxious attachment and anxiety (B = 1.54, $p < .001$) and depression (B = 1.97, $p < .001$), as well as social distancing behaviors (B = 0.36, $p = .002$). There was also a small but significant indirect effect through loneliness between avoidant attachment and anxiety (B = 0.40, p = .015), depression (B = 0.52, $p < .015$), and social distancing behaviors (B = 0.09, p = .043). All

Figure 4.6: The SEM equivalent graph for the Relevant Significant Paths for Cross-Sectional Analyses



Note. Only the main variables of interest were included in the figure. The full results can be found in the tables.

total effects between all causes and effects were significant except the relationship between attachment avoidance and anxiety and depression.

For the longitudinal models, there was a significant effect of fearful-avoidant attachment on anxiety at W3 (B = 0.92, p = .003) but none of the other longitudinal effects of attachment styles on anxiety or depression were significant when accounting for the W2 levels of the outcome variables. Thus, attachment styles did not cause a significant change from W2 to W3 in anxiety or depression except for fearful-avoidant attachment on anxiety.

Table 4.4: The Results of the Structural Equation Modeling for Cross-Sectional Data

	Estimate	Std. Err.	z	p
		Regression Slopes		
W2_anxiety				
Fearful	1.19	0.33	3.66	0
Anxious	1.35	0.46	2.96	0.003
Avoidant	0.07	0.31	0.22	0.825
Loneliness	1.39	0.07	19.48	0
Woman	0.5	0.24	2.03	0.042
Keyworker	0.02	0.27	0.06	0.953
Risk.1month	3.72	0.5	7.51	0
Chronic.illness.self	-0.02	0.29	-0.08	0.932
W2_depression				
Fearful	1.1	0.34	3.21	0.001
Anxious	1.18	0.48	2.45	0.014

Avoidant	0.11	0.33	0.33	0.741
Loneliness	1.78	0.08	23.59	0
Woman	0.24	0.26	0.94	0.345
Keyworker	0.01	0.29	0.02	0.982
Risk.1month	3.39	0.52	6.47	0
Chronic.illness.self	0.03	0.3	0.09	0.932
W2_loneliness				
Fearful	1.25	0.12	10.37	0
Anxious	1.11	0.17	6.42	0
Avoidant	0.29	0.12	2.45	0.014
Woman	0.15	0.09	1.6	0.109
Keyworker	0.15	0.1	1.43	0.152
Risk.1month	0.84	0.19	4.45	0
Chronic.illness.self	0.39	0.11	3.55	0
W2_social distancing				
Fearful	0.92	0.41	2.25	0.024
Anxious	1.42	0.57	2.49	0.013
Avoidant	-1.01	0.39	-2.61	0.009
Loneliness	0.32	0.09	3.6	0
Woman	-0.97	0.31	-3.18	0.001
Keyworker	2.41	0.34	7.04	0
Risk.1month	3.91	0.62	6.29	0
Chronic.illness.self	0.03	0.36	0.08	0.936
W2_hygiene				
Fearful	-2.84	0.52	-5.48	0
Anxious	-4.42	0.73	-6.1	0
Avoidant	-0.97	0.49	-1.98	0.048
Loneliness	-0.12	0.11	-1.05	0.293
Woman	1.74	0.39	4.46	0
Keyworker	-1.37	0.44	-3.15	0.002
Risk.1month	-0.99	0.79	-1.25	0.211
Chronic.illness.self	-0.55	0.46	-1.2	0.229
			Intercepts	
Anxiety	-4.4	0.42	-10.55	0
Depression	-5.13	0.44	-11.65	0
Loneliness	3.64	0.12	29.2	0
Risk	9.4	0.52	18.01	0

Hygiene	48.89	0.66	73.65	0
Fearful	0.28	0.01	22.53	0
Anxious	0.09	0.01	11.75	0
Avoidant	0.3	0.01	23.59	0
Woman	0.48†			
Keyworker	0.29†			
Risk.1month	0.40†			
Chronic.illness.self	0.24†			
		Indirect Effects through Loneliness		
Fearful to anxiety	1.74	0.19	9.15	0
Anxious to anxiety	1.54	0.25	6.1	0
Avoidance to anxiety	0.4	0.17	2.43	0.015
Fearful to depression	2.23	0.23	9.49	0
Anxious to depression	1.97	0.32	6.2	0
Avoidance to depression	0.52	0.21	2.43	0.015
Fearful to social distancing	0.4	0.12	3.4	0.001
Anxious to social distancing	0.36	0.11	3.14	0.002
Avoidance to social distancing	0.09	0.05	2.02	0.043
		Total Effects		
Total.fear.anxiety	2.93	0.36	8.25	0
Total.anx.anxiety	2.89	0.51	5.68	0
Total.avo.anxiety	0.47	0.35	1.35	0.177
Total.fear.depression	3.33	0.39	8.44	0
Total.anx.depression	3.15	0.56	5.58	0
Total.avo.depression	0.63	0.39	1.61	0.107
Total.fear.social distancing	1.32	0.39	3.35	0.001
Total.anx.social distancing	1.78	0.56	3.15	0.002
Total.avo.social distancing	-0.92	0.39	-2.36	0.018
		Fit Indices		
χ^2	35.99(11)			0
CFI	0.99			
TLI	0.96			
RMSEA	0.04			
†Fixed parameter				

Note. Gender, keyworker status, 1 month risk, and chronic illness were used as control variables as identified by the causal discovery algorithm and minSEM.

Table 4.5: The Results of the Structural Equation Modeling for Longitudinal Data

	Estimate	Std. Err.	z	p
			Regression Slopes	
W3_GAD_Total				
W2.fearful	0.92	0.31	3.01	.003
W2.anxious	0.76	0.46	1.64	.100
W2.avoidant	0.30	0.28	1.09	.276
W2.Loneliness.Total	0.39	0.08	4.80	.000
W2.Social distancing.total	0.00	0.02	0.18	.856
W2.Hygiene.total	-0.01	0.02	-0.38	.706
W2.GAD.Total	0.62	0.03	21.74	.000
W2.Age.year	0.00	0.01	0.34	.735
W2.woman	0.31	0.23	1.35	.177
W2.married	0.24	0.35	0.69	.490
W2.single	0.61	0.39	1.57	.116
W2.cohabiting	0.61	0.46	1.32	.188
W2.Keyworker	-0.10	0.27	-0.37	.714
W2.Adults.household	0.09	0.14	0.66	.507
W2.Children.household	-0.01	0.18	-0.08	.936
W2.Change.Income	0.00	0.00	0.60	.552
W2.Chronic.illness.self	0.46	0.26	1.80	.072
W2.COVID19.anxiety	0.01	0.00	2.52	.012
W2.RISK.1month	0.00	0.00	0.57	.566
W3_Dep_Total				
W2.fearful	0.04	0.29	0.15	.880
W2.anxious	-0.24	0.43	-0.57	.571
W2.avoidant	-0.16	0.26	-0.61	.544
W2.Loneliness.Total	0.33	0.08	4.16	.000
W2.Social_distancing.total	0.02	0.02	1.10	.271
W2.Hygiene.total	-0.03	0.02	-2.08	.038
W2.Dep.Total	0.26	0.03	9.20	.000
W3.GAD.Total	0.73	0.03	25.80	.000
W2.Age.year	0.00	0.01	0.19	.849
W2.woman	-0.36	0.22	-1.66	.098
W2.married	0.48	0.33	1.46	.144
W2.single	0.04	0.36	0.11	.913

W2.cohabiting	-0.21	0.43	-0.49	.621
W2.Keyworker	-0.43	0.25	-1.71	.088
W2.Adults.household	-0.24	0.13	-1.76	.078
W2.Children.household	-0.31	0.17	-1.86	.063
W2.Change.Income	-0.00	0.00	-0.53	.595
W2.Chronic.illness.self	-0.02	0.24	-0.10	.921
W2.COVID19.anxiety	-0.01	0.00	-1.86	.063
W2.RISK.1month	0.01	0.00	1.12	.263
W2_GAD_Total				
W2.fearful	0.86	0.35	2.47	.013
W2.anxious	1.15	0.52	2.20	.028
W2.avoidant	0.35	0.31	1.12	.264
W2.Loneliness.Total	1.19	0.09	13.88	.000
W2.Social_distancing.total	0.10	0.02	4.12	.000
W2.Hygiene.total	-0.02	0.02	-1.25	.210
W2.Age.year	-0.05	0.01	-4.49	.000
W2.woman	0.55	0.26	2.08	.037
W2.married	0.80	0.40	2.01	.044
W2.single	0.07	0.44	0.15	.880
W2.cohabiting	0.33	0.52	0.64	.525
W2.Keyworker	-0.38	0.31	-1.23	.218
W2.Adults.household	0.23	0.16	1.41	.157
W2.Children.household	-0.54	0.20	-2.63	.008
W2.Change.Income	-0.01	0.01	-0.95	.341
W2.Chronic.illness.self	0.03	0.29	0.11	.912
W2.COVID19.anxiety	0.05	0.01	9.42	.000
W2.RISK.1month	0.01	0.01	2.49	.013
W2_Dep_Total				
W2.fearful	0.88	0.37	2.36	.018
W2.anxious	1.25	0.57	2.21	.027
W2.avoidant	0.45	0.34	1.33	.184
W2.Loneliness.Total	1.49	0.09	16.10	.000
W2.Social_distancing.total	0.15	0.03	5.68	.000
W2.Hygiene.total	-0.04	0.02	-1.80	.071
W2.Age.year	-0.04	0.01	-2.94	.003
W2.woman	0.12	0.28	0.41	.680
W2.married	0.49	0.43	1.13	.257

W2.single	0.66	0.48	1.38	.169
W2.cohabiting	0.42	0.57	0.75	.455
W2.Keyworker	-0.50	0.33	-1.48	.139
W2.Adults.household	0.12	0.18	0.67	.503
W2.Children.household	-0.24	0.22	-1.08	.281
W2.Change.Income	-0.01	0.01	-1.27	.204
W2.Chronic.illness.self	-0.05	0.32	-0.17	.864
W2.COVID19.anxiety	0.03	0.01	5.55	.000
W2.RISK.1month	0.01	0.01	1.68	.093
W2_Loneliness_Total				
W2.fearful	1.31	0.14	9.37	.000
W2.anxious	1.23	0.22	5.60	.000
W2.avoidant	0.30	0.13	2.28	.022
W2.woman	0.16	0.11	1.44	.150
W2_Risk_total				
W2.fearful	0.66	0.46	1.43	.152
W2.anxious	1.77	0.69	2.55	.011
W2.avoidant	-0.76	0.42	-1.82	.069
W2.Loneliness.Total	0.29	0.11	2.54	.011
W2.Age.year	-0.04	0.02	-2.92	.003
W2.woman	-0.97	0.35	-2.80	.005
W2.married	0.48	0.53	0.91	.364
W2.single	-0.29	0.59	-0.49	.622
W2.cohabiting	-0.20	0.70	-0.28	.779
W2.Keyworker	1.59	0.41	3.91	.000
W2.Adults.household	-0.46	0.21	-2.17	.030
W2.Children.household	0.78	0.27	2.89	.004
W2.COVID19.anxiety	-0.00	0.01	-0.36	.716
W2.RISK.1month	0.03	0.01	4.57	.000
W2_Hygiene_total				
W2.fearful	-2.15	0.59	-3.62	.000
W2.anxious	-2.33	0.90	-2.59	.010
W2.avoidant	-0.68	0.54	-1.26	.209
W2.Loneliness.Total	-0.04	0.15	-0.28	.779
W2.Social_distancing.total	-0.28	0.04	-6.66	.000
W2.Age.year	0.06	0.02	3.20	.001
W2.woman	1.99	0.45	4.44	.000

W2.married	0.78	0.69	1.13	.257
W2.single	-0.06	0.77	-0.08	.934
W2.cohabiting	0.58	0.90	0.65	.518
W2.Keyworker	0.27	0.53	0.50	.619
W2.Adults.household	0.10	0.28	0.35	.726
W2.Children.household	-0.37	0.35	-1.05	.295
W2.Change.Income	-0.01	0.01	-1.22	.224
W2.Chronic.illness.self	-0.88	0.51	-1.73	.083
W2.COVID19.anxiety	0.04	0.01	4.30	.000
W2.RISK.1month	-0.01	0.01	-0.74	.461
W2_Children_household				
W2.Age.year	-0.01	0.00	-8.13	.000
W2.Keyworker	0.07	0.05	1.41	.158
W2.RISK.1month	0.00	0.00	1.58	.114
W2_Adults_household				
W2.Children.household	0.03	0.04	0.85	.393
W2.RISK.1month	-0.00	0.00	-1.49	.137
W2_married				
W2.Keyworker	0.03	0.03	0.97	.332
W2.Adults.household	0.18	0.02	10.00	.000
W2.fearful	-0.18	0.04	-4.55	.000
W2.anxious	-0.10	0.06	-1.64	.102
W2.avoidant	-0.08	0.04	-2.26	.024
W2_single				
W2.Keyworker	0.01	0.03	0.45	.652
W2.Adults.household	-0.06	0.02	-4.19	.000
W2.fearful	0.16	0.03	4.73	.000
W2.anxious	0.16	0.05	2.94	.003
W2.avoidant	0.11	0.03	3.44	.001
W2_cohabiting				
W2.Keyworker	0.00	0.02	0.04	.969
W2.Adults.household	0.01	0.01	0.94	.347
W2.fearful	0.01	0.03	0.29	.768
W2.anxious	-0.01	0.04	-0.25	.801
W2.avoidant	-0.02	0.02	-0.88	.380
			Intercepts	
W3.GAD.Total	-2.19	1.20	-1.82	.069

W3.Dep.Total	1.42	1.13	1.26	.206
W2.GAD.Total	-3.95	1.36	-2.89	.004
W2.Dep.Total	-3.85	1.47	-2.61	.009
W2.Loneliness.Total	4.04	0.10	39.50	.000
W2.Social_distancing.total	12.80	1.37	9.37	.000
W2.Hygiene.total	44.89	1.85	24.26	.000
W2.Children.household	0.90	0.10	9.33	.000
W2.Adults.household	2.03	0.05	38.03	.000
W2.married	0.22	0.04	4.95	.000
W2.single	0.26	0.04	6.67	.000
W2.cohabiting	0.09	0.03	3.06	.002
W2.fearful	0.25	0.01	17.57	.000
W2.anxious	0.07	0.01	8.63	.000
W2.avoidant	0.30	0.01	20.13	.000
W2.woman	0.45+			
W2.Age.year	51.84†			
W2.Keyworker	0.26†			
W2.Change.Income	-8.80†			
W2.Chronic.illness.self	0.25†			
W2.COVID19.anxiety	60.41†			
W2.RISK.1month	38.50†			
			Indirect Effects*	
FearfultoLtoGAD	0.52	0.12	4.27	.000
AnxioustoLtoGAD	0.49	0.13	3.65	.000
AvoidancetoLtoGAD	0.12	0.06	2.06	.039
FearfultoLtoDep	0.44	0.11	3.81	.000
AnxioustoLtoDep	0.41	0.12	3.34	.001
AvoidancetoLtoDep	0.10	0.05	2.00	.045
FearfultoLtoHtoGAD	0.00	0.00	0.23	.822
AnxioustoLtoHtoGAD	0.00	0.00	0.22	.823
FearfultoLtoHtoDep	0.00	0.01	0.28	.781
AnxioustoLtoHtoDep	0.00	0.01	0.28	.781
AvoidancetoLtoHtoDep	0.00	0.00	0.28	.783
FearfultoLtoRtoHtoGAD	0.00	0.00	0.37	.710
AnxioustoLtoRtoHtoGAD	0.00	0.00	0.37	.710
AvoidancetoLtoRtoHtoGAD	0.00	0.00	0.37	.713
earfultoLtoRtoHtoDep	0.00	0.00	1.54	.123

AnxioustoLtoRtoHtoDep	0.00	0.00	1.51	.132
AvoidancetoLtoRtoHtoDep	0.00	0.00	1.29	.197
FearfultoRtoGAD	0.00	0.01	0.18	.857
AnxioustoRtoGAD	0.01	0.04	0.18	.857
AvoidancetoRtoGAD	-0.00	0.02	-0.18	.857
FearfultoRtoDep	0.02	0.02	0.87	.382
AnxioustoRtoDep	0.04	0.04	1.01	.312
AvoidancetoRtoDep	-0.02	0.02	-0.94	.346
FearfultoRtoHtoGAD	0.00	0.00	0.36	.716
AnxioustoRtoHtoGAD	0.00	0.01	0.37	.710
AvoidancetoRtoHtoGAD	-0.00	0.00	-0.37	.712
FearfultoRtoHtoDep	0.01	0.01	1.16	.245
AnxioustoRtoHtoDep	0.02	0.01	1.57	.117
AvoidancetoRtoHtoDep	-0.01	0.01	-1.34	.180
			Total Effects	
total.fear.GAD	1.44	0.31	4.64	.000
total.anx.GAD	1.25	0.47	2.68	.007
total.avo.GAD	0.42	0.28	1.48	.139
total.fear.Dep	0.51	0.29	1.72	.086
total.anx.Dep	0.20	0.44	0.46	.649
total.avo.Dep	-0.05	0.26	-0.17	.861
			Fit Indices	
χ^2	2245.54(77)			.000
CFI	0.63			
TLI	0.00			
RMSEA	0.17			
†Fixed parameter				

Note. The full model results are presented in the table for transparency but only the bolded paths are of interest. All indirect effects are italicized. The fit effects of this model are poor but given fit statistics are inherently predictive metrics rather than causal, we have not modified the model. *Indirect effects: L = loneliness, H = hygiene, R = risk behaviors. The indirect paths can be interpreted as: AvoidancetoLtoRtoHtoGAD = indirect affect of attachment avoidance to anxiety through loneliness, risk behavior, and hygiene (serial mediation).

4.5.6 Model Specification for Targeted Learning

Outcome: W3_Dep_Total

Confounders: Empty Set

Precision Variables: W2_Chronic_illness_self, W2_Adults_household, W2_Keyworker, W2_Change_Income, W2_Age_year, W2_Children_household, W2_RISK_1month, W2_COVID19_anxiety, W2_Pregnant, W2_Gender

Outcome: W3_GAD_Total

Confounders: Empty Set

Precision Variables: W2_Chronic_illness_self, W2_Adults_household, W2_Keyworker, W2_Change_Income, W2_Age_year, W2_Children_household, W2_RISK_1month, W2_COVID19_anxiety, W2_Pregnant, W2_Gender

Outcome: W2_Risk_Total

Confounders: W2_Gender

Precision Variables: W2_Chronic_illness_self, W2_Keyworker, W2_Change_Income, W2_Age_year, W2_Children_household, W2_RISK_1month, W2_COVID19_anxiety, W2_Pregnant

Outcome: W2_Dep_Total

Confounders: W2_Gender

Precision Variables: W2_Chronic_illness_self, W2_Keyworker, W2_Change_Income, W2_Age_year, W2_Children_household, W2_RISK_1month, W2_COVID19_anxiety, W2_Pregnant

Outcome: W2_GAD_Total

Confounders: W2_Gender

Precision Variables: W2_Chronic_illness_self, W2_Keyworker, W2_Change_Income, W2_Age_year, W2_Children_household, W2_RISK_1month, W2_COVID19_anxiety, W2_Pregnant

Outcome: W2_Loneliness_Total

Confounders: W2_Gender

Precision Variables: W2_Chronic_illness_self, W2_Keyworker, W2_Change_Income, W2_RISK_1month, W2_COVID19_anxiety

Intentional Mediation Included:

Mediator: W2_Dep_Total

Outcome: W3_Dep_Total

Confounders: Empty Set Precision Variables: W2_Chronic_illness_self, W2_Adults_household, W2_Keyworker, W2_Change_Income, W2_Age_year, W2_Children_household, W2_RISK_1month, W2_COVID19_anxiety, W2_Pregnant,W2_Gender

Mediator: W2_GAD_Total

Outcome: W3_GAD_Total

Confounders: Empty Set Precision Variables: W2_Chronic_illness_self, W2_Adults_household, W2_Keyworker, W2_Change_Income, W2_Age_year, W2_Children_household, W2_RISK_1month, W2_COVID19_anxiety, W2_Pregnant,W2_Gender

CHAPTER 5

Prespecification of Structure for the Optimization of Data Collection and Analysis

> *"Causal effects are not binary signals that are either detected or undetected; causal effects of numerical quantities that need to be estimated. Because the goal is to quantify the effect as unbiasedly and precisely as possible, the solution to observational analyses with imprecise effect estimates is not avoiding observational analyses with imprecise estimates, but rather encouraging the conduct of many observational analyses... meta-analyze them and provide a more precise pooled estimate."*
>
> M.A. Hernan (2022)

The content of this chapter is drawn from the following publication:

Vowels, M.J., 2023, Prespecification of Structure for the Optimization of Data Collection and Analysis. *Collabra: Psychology*.

Abstract: Data collection and research methodology represents a critical part of the research pipeline. On the one hand, it is important that we collect data in a way that maximises the validity of what we are measuring, which may involve the use of long scales with many items.

On the other hand, collecting a large number of items across multiple scales results in participant fatigue, and expensive and time consuming data collection. It is therefore important that we use the available resources optimally. In this work, we consider how the representation of a theory as a causal/structural model can help us to streamline data collection and analysis procedures by not wasting time collecting data for variables which are not causally critical for answering the research question. This not only saves time and enables us to redirect resources to attend to other variables which are more important, but also increases research transparency and the reliability of theory testing. To achieve this, we leverage structural models and the Markov conditional independency structures implicit in these models, to identify the substructures which are critical for a particular research question. To demonstrate the benefits of this streamlining we review the relevant concepts and present a number of didactic examples, including a real-world example.

5.1 Introduction

Imagine you want to estimate the effect of a therapeutic treatment on depressive symptoms, and how this effect may be mediated via another variable, say, therapeutic alliance. One might suspect that these variables are linked through a complex causal web involving multiple other factors - but which of these other factors are necessary, in terms of data collection, for estimating the main effect of interest? Collecting too many variables increases the cost and time required to complete data collection, having an impact on participant fatigue (Lavrakas, 2008) as well as draining valuable project resources. Conversely, collecting too few may make render the results of the statistical tests invalid. In this manuscript, we describe how to identify those variables which are strictly necessary to arrive at unbiased answers to pre-specified questions. Of course, other interests may influence data collection (such as subsequent applications and usage), but knowing what is strictly necessary allows one to make more informed decisions about what to include.

In this paper, we argue that the data collection and research project methodology can be optimized by specifying the causal structure underlying a theory in graphical form. Using rules from the structural modeling framework, one can then use the graph to identify variables or scales which are either causally necessary or which can be omitted from the data collection

process. This liberates resources to either improve the quality of the remaining scales (*e.g.*, by using scales with a more comprehensive set of items), and/or to reduce participant fatigue by shortening the duration of a questionnaire and using these resources to increase the overall sample size. Indeed, concerns about inadequate statistical power are growing in response to the replication crisis (Sassenberg and Ditrich, 2019; Baker et al., 2020; Correll et al., 2020; Aarts et al., 2015), and researchers are thus encouraged to make sure they have sufficient data to estimate the effects of interest.

Furthermore, even if a researcher decides not to undertake any analyses (perhaps they are not able to collect data, for whatever reason) the process of reflecting a theory graphically nonetheless helps with transparency, reproducibility, and the meaningfulness of subsequent interpretation. Psychology has been accused of being 'not even wrong' (Scheel, 2022) on the basis that the theories are too vague to be adequately tested. By reflecting our theories in a graphical form, we thus improve the clarity and reduce the one-to-many relationship between our theories and our statistical models. Translating our theories to graphs also forces researchers to think carefully about the underlying process, and the concomitant implications for data collection. The specification can then be made explicit, preregistered (Nozek et al., 2018), and compared unambiguously against other work. This, in turn, facilitates more precise replication by subsequent researchers, as well as a clearer understanding of the relationships between the hypotheses being tested and the assumptions and theory which underpin the model specification and results (Navarro, 2021; Haslbeck et al., 2021; Grosz, Rohrer, and Thoemmes, 2020).

In this work we show how four related concepts - conditional independencies, Markov Blankets, projection, and causal identification - can be used to judiciously shrink the number of variables required to answer a research question, without impacting downstream analyses and without impacting the congruity of the model with the underlying theory. The process is not data-driven and is not the same as seeking model 'parsimony' - our approach does not fundamentally change the complexity of the underlying processes reflected by the 'full' model. Instead, using a set of rules which are consistent with the assumptions of the original graph being specified, our initial graphical representation can be reduced to focus in on the effects we really care about. Thus whilst the complexity of the statistical model reduces, it does so without introducing any

additional simplifying assumptions beyond those which already existed in the original theory.

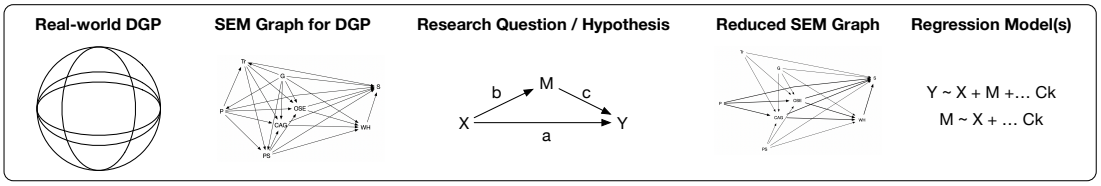
The techniques are relevant to a broad range of problems amenable to specification in graphical form. For example, the didactic examples given by Rohrer (2018) involve health problems and work satisfaction, genetics and child's depressiveness, or educational attainment and income. Additionally, social psychologists interested in complex, mediated processes and multiple baseline control variables could also benefit from the proposal presented here. To this end, as well as providing a set of experimental results to demonstrate the performance characteristics in a general and non-domain-specific way, we also provide an example application to a graph used in organizational behavior (Spurk and Abele, 2011). Our hope is that researchers can use the techniques presented in this work so that they can optimize their data collection and analysis in a more transparent way which is tailored specifically to the particular relationships of interest.

We begin by motivating the specification of our theories in graphical form. Then, we introduce the relevant statistical/structural concepts needed to understand the process for reducing this model. We then walk through a number of didactic examples, comparing an assumed 'real-world' or Data Generating Process (DGP) against the minimal required model for estimating a set of causal effects of interest. We also provide the associated multiple linear regression models where a single regression model can be used to provide the same information, and present a real-world example. In supplementary, we also provide simulation results to demonstrate that the approach does not introduce bias, and in some cases can improve model fit and reduce standard error. Finally, in the supplementary we also provide the code for an automatic tool for reducing the graph (along with a description of the associated algorithm). The code for reproducing the simulations as well as the automatic tool are also provided here: `https://github.com/matthewvowels1/minSEM`.

5.1.1 Terminology and Conceptual Overview

In this work, we assume that psychologists/researchers are principally concerned with estimating a particular causal effect (*e.g.*, the effect of treatment on an outcome). Indeed, this goal aligns with the causal nature of psychological theories (which, in general, describe causal processes), as well as the goal to design and implement effective interventions which improve peoples' lives.

Figure 5.1: Top level terminology.



Note. We assume (left) there exists a real-world causal Data Generating Process (DGP), which we wish to model using a structural model. This structural model can be represented graphically (see SEM graph for the DGP in the figure). Using our proposed approach, this SEM can be simplified in such a way that does not jeopardise the estimation of a particular (causal) effect size which is of interest to our research. For example, we may be interested in estimating path coefficients/effects a, b, and c in a mediation model. Finally, the effect sizes may be estimated using straightforward regression models.

As such, we assume that a researcher wishes to test a particular hypothesis which concerns a (causal) effect size of interest.

We will refer to a number of objects which deserve to be defined up-front. In Figure 5.1 we present examples of these objects for reference. Firstly, we assume that there exists some (potentially highly complex) real-world *Data Generating Process* (DGP). According to our existing theories, we wish to model this DGP in such a way that we are able to meaningfully represent it. One option for doing so involves the use of *Structural Equation Models* (SEMs). SEM provides us with a powerful and popular (Blanca, Alarcon, and Bono, 2018) statistical framework to unambiguously reflect and test causal theories and relationships (M. J. Vowels, 2021; Rohrer, 2018; Grosz, Rohrer, and Thoemmes, 2020; S. Wright, 1921; S. Wright, 1923; Pearl, 2009). In particular, the SEM can be represented in an intuitive *graphical* (and therefore visual) way, thereby specifying our domain knowledge about the DGP.

The graphical representation of the theory, which we will refer to as the graphical or structural model, can be used early on in the research pipeline to inform the data collection methodology, by helping us specify which constructs we need to measure. Furthermore, early specification of a statistical models helps us with preregistration and research transparency (Wagenmakers et al., 2012). Such transparency is increasingly important in the fields of psychology and social science, where attention has been drawn to numerous problems with theory testing, research methodology, and analytical practice (M. J. Vowels, 2021; Flake and Fried, 2020; Scheel et al., in press; Gigerenzer, 2018; McShane et al., 2019; Aarts et al., 2015; Marsman et al., 2017).

As we will discuss, we will apply the rules of a type of graphical model known as a *Directed Acyclic Graph* (DAG) to the graphical representations of our SEM. These rules are actually more general than those specific to SEM, because whilst SEM assumes linear relationships between variables, the rules we use are applicable to problems with almost arbitrarily non-linear relationships. Using these rules, and in combination with a *Research Question* expressed as a set of target causal effects of interest, we can reduce its complexity (which we refer to as the *Reduced* SEM) without sacrificing our ability to estimate what we care about for a particular research question or hypothesis. This reduced model then determines which variables we are required to collect data for. In some cases, we may not need to use the typical SEM estimation techniques to answer our research questions, and a simple multiple regression model may suffice. However, it is worth emphasising that this work is not concerned with the estimation of the coefficients themselves, but rather how we can use the graphical modeling rules to simplify the representation of a theory, and in turn streamline our data collection and study design.

5.2 Motivation

In this section we provide two principal motivations for our proposed approach: Statistical power, and model under- or mis-specification. In light of these motivations, we then provide a top-level overview of our proposal.

5.2.1 Statistical Power and Model Specification

Psychological research is frequently *underpowered* (Vankov, Bowers, and Munafo, 2014; Maxwell, 2004; Crutzen and G. Peters, 2017), and the theory and analysis are often poorly specified (Scheel et al., in press; Scheel, 2022; Grosz, Rohrer, and Thoemmes, 2020; Rohrer, 2018; M. J. Vowels, 2021). The studies are underpowered to the extent that the sample sizes are insufficient to test a target hypothesis. For example, for a minimum assumed *true* effect size of interest, it is generally recommended that enough data are collected to yield a power of 80%, meaning that there is an 80% probability that we will find a statistically significant result (at a given threshold such as 0.05) (Gelman, Hill, and Vehtari, 2021). Researchers are

thus encouraged to ensure that their studies are adequately powered, and have been encouraged to do so for some time (Vankov, Bowers, and Munafo, 2014; Sedlmeier and Gigerenzer, 1989). However, depending on the complexity of the theory under test, researchers may need to measure a large number of constructs, each with a large number of items. For example, depending on the format, the IPIP-NEO Big 5 inventory contains between 120-300 items (Goldberg, 1999; Goldberg et al., 2006) and therefore takes considerable time to complete. Besides the associated cost and time required to measure constructs using such comprehensive scales, the participants may also experience fatigue, lowering the quality of the responses (Lavrakas, 2008).

The second problem of under-specification has prompted meta-researchers to describe research in psychology as 'not even wrong' (Scheel, 2022). That is to say, if the theories are too vague to be specified unambiguously, then it is not clear what it is that any particular statistical test is actually testing. If we are considered with understanding the real-time process of dyadic support, for instance, we might need to develop a statistical model which can capture the intricacies of back-and-forth, multi-modal (verbal, para-verbal, non-verbal) interactions between partners. Without unambiguously reflecting the complexity of the process in our statistical model, it is not clear what a typical model in psychology (*e.g.*, a multiple linear regression model) is really doing for us. The structural representation of this process can be a helpful aid to understand (a) what data we need to collect, and (b) whether the data can even be collected in principle (the acquisition of real-time, multi-modal data may in some cases be infeasible).

Furthermore, a single theory may admit multiple statistical models, each of which tests something slightly different but all of which are valid given the malleability of the underlying theory. Few psychological theories make it clear which variables are necessary to include as control variables, for instance. And yet, the inclusion of different control variables can have a large impact on the resulting parameter estimates, and it is not usually clear how these control variables are chosen or how they relate to the tested theory (M. J. Vowels, 2021; Hullman et al., 2022; Cinelli, Forney, and Pearl, 2022). As an example, in medical studies older patients may be more likely to choose medication over surgery, but also be less likely to recover. This makes age a key confounder that must be controlled/adjusted for to evaluate the treatment effects. However, perhaps there exist other, less obvious confounders which we have not collected and

which we can therefore not adjust for. Some variables may need to be controlled for but be unattainable, some may be inconsequential (and can be omitted without consequence), and still others may actually be detrimentally biasing the model. In order to determine which control variables should or should not be included, and to therefore avoid what is known as structural misspecification (M. J. Vowels, 2021), researchers need to somehow formalise their theories.

5.2.2 The Proposed Solution

With respect to statistical power, there exists a need for compromise - maximising the quality of a survey such that it measures all that we need, at a sufficient level of quality, for a sufficient number of participants. Of course, we acknowledge that there often exist multiple goals for studies in which new data will be collected - they may have either confirmatory or exploratory research questions, or both; they may wish to compare and contrast multiple competing hypothesized structures; they may want to 'future-proof' the study, such that additional variables are collected with a view that they may be necessary for answering research questions which are not yet specified.

At the same time, and in order to correctly specify a model with respect to a psychological theory, it is important that psychologists consider not only the structure between the primary constructs central to their theory, but also the full data-generating process (DGP) which leads to a set of observations. The theory can then be translated into a graphical/structural model which reflects this DGP, which we can use to make sure we are not missing variables which are key to answering a particular research question. The process of deriving a structural model from our theory has been previously discussed by Rohrer (2018) and others (Kline, 2005; Loehlin and Beaujean, 2017), and we do not describe the procedure in this work, but note that the graphical framework (more about this in later sections) makes the process quite intuitive.

The advantages of reflecting the theory unambiguously in a structural model include reproducibility (it is clear what exactly is being tested) and an increase in the interpretability and validity of the resulting effect sizes. Rather than the effect sizes being arbitrary consequences of *ad hoc* models loosely connected to theory, they reflect specific causal effects within a fully specified structural/causal process. Whilst the causal validity of effect sizes estimated using

these models still depends on whether a number of strong assumptions hold (*e.g.*, whether the hypothesized structure is correctly specified with respect to the actual, real-world structure), the transparent specification of the model makes subsequent criticisms and revisions more precise. The task of translating our theories may also highlight possible weaknesses in the theory, or call attention to possibly insurmountable difficulties for data collection. For instance, theories which involve dynamic processes that unfold at irregular intervals over time may require very specific, expensive, and challenging data collection procedures (Hilpert et al., 2019). Identifying the specifics of such challenges in advance could save a lot of wasted time and effort.

Unfortunately, the task of identifying all relevant variables will likely implicate a large number of secondary variables (such as demographics and other theoretically related constructs), and thus require longer questionnaires. The problems of statistical power, comprehensive scale inventories, and the need to collect a broad range of variables and constructs relevant to our theory puts a lot of pressure on researchers to find a suitable 'Goldilocks' design, and one or multiple methodological facets are likely to be compromised as a consequence. As such, after the specification of the full DGP, we should examine the resulting model to identify possible shortcuts in the data collection process. Indeed, and as we will show, even if a variable or construct is relevant to a particular causal process, it may not be required for the actual analysis. To know this, however, the variable needs to be transparently situated in a causal model for us to understand whether it is essential for answering a target research question, or not.

Once the structure of the DGP is fully specified, and as we will describe in detail below, we are able to identify essential substructures which are sufficient for testing our intended hypotheses. The substructures, by definition, exclude certain variables. Thus, if we can identify these substructures in advance of data collection, we may be able to significantly reduce the number of constructs we need to measure. Indeed, in example 2i in Figure 5.4) below, we show that it is possible to reduce the number of variables/constructs by two thirds, although this depends on how much of the causal process we are interested in testing. It goes without saying that any simplification must be done carefully. Indeed, the potential consequences of any resultant model misspecification can be severe, and includes heavily biased parameter estimates which are almost impossible to meaningfully interpret (M. J. Vowels, 2021; Hullman et al., 2022).

However, there are no requirements for researchers to 'go all the way' with the simplification, and the proposal is flexible insofar as the degree of desired reduction can be determined by the researcher and their specific requirements.

We thus advocate that researchers consider the DGP upfront, before the data collection stage. Such prespecification in the form of a structural (or, as we will present, graphical) model represents a beneficial step in terms of preregistration and transparency, helps researchers distill their theories into testable models, thereby increasing the validity and meaningfulness of downstream statistical inference and results interpretations, and provides us with an opportunity to 'prune' the structure to optimize for statistical power during data collection.

5.3 Background

In this section, we introduce a number of relevant technical concepts for reducing our structural models. In general, we assume that the model is being specified in graphical form as a path model, or a Structural Equation Model (SEM), where directed paths/arrows correspond with causal links. As we mention above, the techniques we use are more general than the SEM framework, and come from the graphical models literature.[1] A number of existing resources discuss the implications of changes in causal structures on statistical estimation. For example, M. J. Vowels (2021) discusses the problems that arise due to misspecification of causal models, and notes the potential to focus on specific effects within a causal process; and Cinelli, Forney, and Pearl (2022) provides a laconic summary of how to choose control variables such that the choice does not induce bias in our parameter estimation. Unfortunately, these resources do not discuss the possibility of reducing our SEMs to the most simple model which can still yield unbiased estimates of (possibly multiple) causal effects.

To best communicate our approach, we begin with a brief review of the relevant background. We aim to review four related concepts in particular: causal identification, conditional independence,

[1] Both path and SEMs represent subtypes of the graphical modeling framework known as Probabilistic Graphical Models (Koller and Friedman, 2009; Pearl, 2009), and the relevant concepts are adequately reflected in SEMs which are already popular in psychology (Blanca, Alarcon, and Bono, 2018). In order to avoid terminological pedantry we thus assume researchers are using SEMs, but note that the ideas here generalise to other structural frameworks as well (such as Directed Acyclic Graphs and Structural Causal Models).

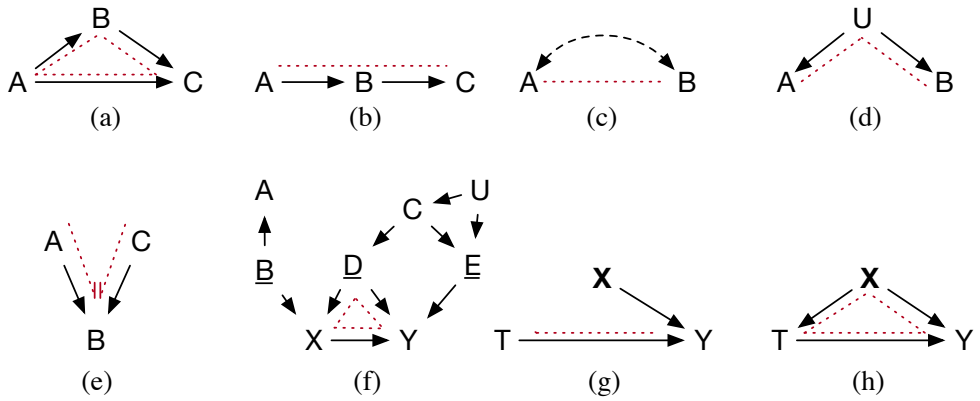
Markov Blankets, and projection. Briefly, identification is the goal of isolating causal from non-causal statistical dependencies, and, when possible, facilitates the estimation of causal effects. It relies on conditional independencies, which describe how statistical dependencies arise due to the underlying causal process, and how conditioning on these variables enables us to isolate or disentangle different sources of dependence. Markov blankets show that, through the use of conditional independencies, we can completely isolate an entire substructure in a graph, thereby making it clear that not all variables are necessarily required for a particular research question. Finally, projection enables us to combine/reduce the number of paths. This is particularly true in the case of mediation, where a mediator can be excluded entirely if the researcher is not interested in estimating the mediation *per se*.

Interested readers are encouraged to consult useful resources by Hünermund and Bareinboim (2021), M. Vowels, N. Camgoz, and Bowden (2022), Cinelli, Forney, and Pearl (2022), J. Peters, Janzing, and Scholkopf (2017), Koller and Friedman (2009), Kline (2005), Pearl (2009), Pearl, M. Glymour, and Jewell (2016), and Loehlin and Beaujean (2017). In terms of notation, we use X (or, *e.g.* A, B, C etc.) to denote a random variable, and bold font $\mathbf{X}$ (or, *e.g.* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ etc.) to denote a set of random variables. We use the symbols $\perp\!\!\!\perp$ and $\not\perp\!\!\!\perp$ to denote statistical independence and statistical dependence, respectively. For linear systems, such statistical dependence may be identified using correlation, but the majority of our discussions are general and non-parametric. We use directed arrows to denote a directional structural/causal dependence, and U (or $\mathbf{U}$) for a single (or set of) unobserved variable(s). [2]

For example, in SCM terminology $A := f(B, C, U_A)$ indicates that A is some general function f of B and C. Here, U_A tells us that A is also a function of exogenous random process U_A. Indeed, it is this U_A which prevents the relationship between A and B and C from being deterministic. Structural Equation Models (SEMs), on the other hand, assume that all endogenous variables are the result of a linear weighted sum of others, such that $A := \beta_{BA}B + \beta_{CA}C + U_A$. Here,

[2]Note that the theory we discuss is applicable to models with latent constructs (such as factor or measurement models), as well as those without (such as path and structural models), and generalises beyond linear models. The theory we discuss is part of the general Structural Causal Modeling (SCM) and Directed Acyclic Graph (DAG) frameworks (Pearl, 2009). Path models and SEMs both represent a subset of the family of SCM and DAG models, where the *functional* relationships between variables are assumed to be linear. In other words SCMs and DAGs make no assumptions about whether one variable is an arbitrarily complex function of another (strictly, there are exceptions to this, as discussed by Maclaren and Nicholson, 2020).

Figure 5.2: A set of demonstrative graphs.



Note. This figure provide a number of example graphical models. Solid black lines indicate causal dependencies, dashed red lines indicate statistical dependence, parallel red bars indicate a 'break' in statistical dependence (example (e)), **boldfont** indicates a set of variables, and the letter U is reserved to denote unobserved variables.

the βs are structural parameters (also called path coefficients or effect sizes) which we wish to estimate. The walrus-shaped assignment operator $:=$ tells us that the left hand side is a structural outcome of the right hand side; the equations are not intended to be rearranged and there is very much a directional relationship involved.

As we construct system of equations representing our SEM (or, indeed, our SCM) it is often convenient to represent these relationships graphically/visually. For example, consider the following set of (linear) structural equations:

$$\begin{aligned} A &:= U_A, \\ B &:= \beta_{AB}A + U_B, \\ C &:= \beta_{AC}A + \beta_{BC}B + U_C. \end{aligned} \tag{5.1}$$

These can be represented simply as the mediation model depicted in black, solid arrows in Figure 5.2(a). The variables U are generally not included unless they are statistically dependent. Of course, they frequently *are* dependent in psychology, and this may be denoted using a curved, bidirected edge, as between variables A and B in Figure 5.2(c), or by explicitly including the relationship as in Figure 5.2(d). Such relationships can, of course, also be included in the system

of equations comprising the SEM. Note that, as a result of the causal structures present in the DGP, there are induced a number of statistical dependencies indicated in Figure 5.2 by the red dashed lines. By induced statistical dependency, we mean that the variables are correlated, or, more generally, statistically dependent, by consequence of the causal relationships between the variables in the underlying causal process.

5.3.1 The Data Generating Process

It is worth maintaining conceptual separation between: (1) the process occurring in the real world, which we consider to be the true Data Generating Process (DGP), (2) Our SEM, which we generally want to sufficiently capture the process in the real world, and (3) the specification of a multiple linear regression. Note that (1) and (2) do not have to match precisely. Indeed, when we create our SEM we expect it to be a significant simplification of the real-world process, but it needs to be somewhat *consistent* with the true process (and the degree to which this is achieved is one of the primary aims of our research). If it is not sufficiently consistent, we might deem it to be *misspecified*, and it will not yield meaningful statistical estimates.

For example, if we have a strong theory that the true DGP can be adequately represented by a fully mediated process $A \rightarrow B \rightarrow C$, then we would be advised to employ an SEM which is consistent with this structure. By consistent we mean that the model we use facilitates the unbiased estimation of the parameters of interest, and that these estimated parameters correspond with something meaningful in the real-world (*e.g.*, causal effects sizes).[3] One option we have is to specify everything about our theory explicitly using an SEM, and this can be done in graphical form to aid formalisation. However, what we aim to show is that if we are primarily concerned with a subset of parameters (*vis a vis* all path coefficients in the model), then in some cases we can significantly reduce the complexity of our model without affecting the consistency of our resulting model. In the case of the *full* mediation, it is interesting to note, for example, that including a direct path in the SEM (in addition to the indirect effect) does not bias our estimates of the indirect path parameters. This is because the direct path will have an estimated

[3] For the estimation task itself, we can either use the SEM estimation framework (and estimate all the included paths), or alternatively, we can derive a set of equivalent regression equations.

effect of zero if it does not exist in the real-world, and its inclusion does not influence the value of the coefficient estimated for the indirect path. This is an example of how *increasing* the complexity of the SEM does not necessarily result in 'disagreement' or misspecification with respect to the SEM and the real-world DGP. In contrast, failing to include a direct path which *does* exist in the real-world DGP, can affect the resulting path estimates. As such, in some cases assumptions which simplify the graph can be more 'dangerous' than those which increase the complexity of the graph, and it is especially important any simplification be done with care to avoid biasing the estimates of the remaining path coefficients.

Finally, note that the effect sizes of interest in the final SEM can be estimated using multiple regression. Indeed, the specification of an SEM using the popular *lavaan* R library (Rosseel, 2012) follows a very similar syntax to that used to estimate each path using the *lm* regression library. Note that this may not always be possible, particularly if one needs to estimate latent factors. However, we provide the equivalent regression syntax to highlight the equivalence between the techniques, and to show that even if a structural model is used to specify the DGP, it may be possible to use a straightforward linear regression model for the actual estimation.

5.3.2 Identification and Disentangling Statistical Influence

Identification is the goal of isolating causal from non-causal statistical dependencies, and, when possible, facilitates the estimation of causal effects. It concerns whether or not, for a given graph, the causal effect we are interested in is actually estimable from the observed data, even in the absence of an experiment (Huang and Valtorta, 2006; Shpitser and Pearl, 2008). In the case where the full graph is given and there are no unobserved confounders, all causal effects are technically identifiable from the data. This means that there exists a mathematical expression which expresses the causal effect(s) of interest as a function of the observed statistical associations. If a causal effect is identifiable, it may be possible to estimate it with only a fraction of all the observed variables. Furthermore, if researchers are only interested in estimating a single path coefficient in a structural model, it may not be necessary to run the full SEM estimation process, and instead researchers can run a multiple regression (possibly employing machine learning techniques) to directly estimate the effect of interest (M. J. Vowels,

N. Camgoz, and Bowden, 2021; M. J. van der Laan and S. Rose, 2011).

In the case where researchers *are* interested in the estimation of multiple paths (for example, in a mediation model), one can choose either to undertake a series of multiple regression analyses (and we provide examples of this below), or to estimate them simultaneously using the SEM estimation framework. In both cases, however, all effects of interests must fulfil the requirements for identification. In other words, the estimation multiple causal effects (*e.g.*, from treatment to mediator and from mediator to outcome) requires that all effects can be identified from the data, which is obviously entails more stringent requirements than does the estimation of only one of these paths.

A detailed description of how to use identification is beyond the scope of this paper, but we describe below how to isolate/disentangle statistical influence using the conditional independency properties below. For now, let us consider the case where we are interested in estimating only one path coefficient / causal effect - the rules generalize to multiple coefficients. Consider the graphs in Figure 5.2(g) and (h). Graph (g) represents the canonical Randomized Control Trial setup, where T represents some treatment, Y some outcome, and $\mathbf{X}$ some set of covariates which help to explain the outcome Y. In this graph, the covariates $\mathbf{X}$ are independent of treatment T because of the random assignment of treatment. Such a structure means the only statistical dependence that exists between the treatment and the outcome is a result of the treatment itself. This statistical dependence is thus equivalent to the causal dependence we are interested in. As such, the effect can be directly estimated by comparing the outcome under different treatments. Note that one may still wish to consider $\mathbf{X}$ too - it can be used to explain additional variance in Y in order to tighten the estimate of the treatment effect. In other words, the inclusion of these variables may reduce the standard errors associated with a particular causal effect size estimate.

In contrast, in observational studies patients may select their own treatment, and graph Figure 5.2(h) is more appropriate. For instance, if age is one of the covariates, older patients may prefer medication and have a lower chance of recovery, whilst younger patients may prefer surgery and have a higher chance of recovery. Thus, if we wish to estimate the *causal* influence of treatment T on the outcome Y, we cannot simply compare the outcomes of the two treatment groups, but now also need to somehow adjust or 'control' for the additional statistical

dependence that exists between Y and T which results from the 'backdoor' non-causal path $T \leftarrow \mathbf{X} \rightarrow Y$. This is non-causal because there is no directed path between T and Y via X (the arrow points from X to T, not the other way around). Knowing the rules of conditional independencies described below, we will be able to isolate the causal effect of interest such that the remaining statistical dependence between T and Y corresponds with the causal dependence we actually wish to estimate.

Note that we will use the term control variables to mean variables which we wish to adjust for to identify causal effects of interest, and which would otherwise leave an opening for non-causal, statistical association. For example, the set of variables $\mathbf{X}$ in Figure 5.2(h) could be considered to be a set of relevant control variables which enables us to get unbiased estimation of the effect of treatment T on the outcome Y. However, it is worth considering that a set of control variables itself may comprise a complicated structure in its own right, and we consider two cases in the examples section below.

5.3.3 Conditional Independencies

The visual graphs provide us with a way to directly read off the conditional independency structure of the model. Conditional independencies tell us whether the inclusion of additional information changes anything about our knowledge. For instance, consider the (illustrative) fully mediated model Testosterone $\rightarrow$ Bone Length $\rightarrow$ Height. This model tells us that, in the absence of a direct path from Testosterone to Height, if we already know someone's Bone Length, knowing their Testosterone in addition changes nothing about their likely height. In other words, no more of the statistical dependency between Testosterone and Height is left to explain once Bone Length is known. Equivalently, if we condition our knowledge on Bone Length, Testosterone is rendered *conditionally independent* of Height. Indeed, if a linear regression is used to estimate the effect of Testosterone on Height, but we include Bone Length as a control variable, the coefficient on Testosterone will tend towards zero. This is a useful example which highlights the importance of a consideration for structure and the associated conditional independencies - if we do not already know that the process is fully mediated, we might incorrectly arrive at the conclusion that Testosterone is unrelated to Height.

If our graph Testosterone $\rightarrow$ Bone Length $\rightarrow$ Height is a sufficient representation of the process in reality, and if the statistical relationships hold in the data we observe, then the graph is also said to be *Markovian* (*i.e.*, the 'Markov condition' holds). In fact a Markovian graph is simply a graph for which its implied conditional independencies hold in the data it is being used to model. Conversely, if their exists one or more unobserved variables which we have failed to include in our model, and which influence the statistical dependencies in our data such that the Markov condition no longer holds, the graph is said to be *semi-Markovian*. If we suspect a graph is semi-Markovian because of the presence of some unobserved confounder(s), we should do our best to update our graph and include this unobserved factor, so that the rules apply to our (now Markovian) model. If we find this unobserved variable is necessary for identification, but we simply cannot collect data for it (it might not be an easily measurably factor), then it may not be possible to estimate the causal effects of interest.[4] Whether or not a causal effect of interest is identifiable is important to understand early on, because it may determine the feasibility of the study. This is another reason why a graphical specification of a theory can be useful.

We can use conditional independencies to isolate causal from non-causal statistical dependence (the task of identification described above), as well as to identify which variables we need to include or exclude in our SEM. Starting with the example in the full mediation model of Figure 5.2(b), we see that variable C cannot contain information about A which does not already 'pass' through B. Therefore, if we already know B, knowing A tells us nothing more about C than we already knew. This renders A statistically independent of C given B, which can be expressed as: $A \perp\!\!\!\perp C|B$. This is known as a conditional independence statement, because it tells us which sets of variables are independent of each other given a set of conditioning variables. It is worth noting that when we run a regression (logistic or otherwise) we are estimating some expected outcome *conditioned on* some set of predictors. Running the regression to estimate $\mathbb{E}[C|B, A]$ (*i.e.*, the expected value of C, controlling for B and A) from data generated according to a fully mediated DGP will result in the same consequences as above: the fact we have included B means that the importance given to A will be zero (notwithstanding finite sample deviations). Clearly, therefore, an understanding of the structure is therefore absolutely

[4] One might consider sensitivity analysis as a means to quantify the extent to which a causal effect can be explained by unobserved third variables (Diaz and M. van der Laan, 2013).

crucial for constructing the regression models (M. Vowels, 2022). For instance, if A is a treatment variable and we do not recognise B as a mediator, the inclusion of B in the model will result in a negligible coefficient estimate for A which may well mislead us to think the treatment is ineffective.

To generalise this result to other graph structures, it is worth committing some rules to memory. If a graph contains these two substructures:

$$\begin{aligned} A \rightarrow B \rightarrow C, \\ A \leftarrow B \rightarrow C, \end{aligned} \tag{5.2}$$

then knowing/conditioning on B renders A and C statistically independent. Of course, without this conditioning, A, B, and C are all statistically dependent. These two graphs are known, respectively, as a chain and a fork. One can start to write the complete list of conditional independencies which are implied by *both* of these two graphs is:

$$A \not\perp\!\!\!\perp B, \;\; A \not\perp\!\!\!\perp C, \;\; B \not\perp\!\!\!\perp C, \;\; C \perp\!\!\!\perp A|B, \;\; C \not\perp\!\!\!\perp B|A, \;\; B \not\perp\!\!\!\perp C|A. \tag{5.3}$$

The first, $A \not\perp\!\!\!\perp B$, means that A is not statistically independent of B (because A causes B), the second means that A is not statistically independent of C (because A causes C through B), and so on. Importantly, both of the graphs in Eq. 5.2 imply the same set of conditional independencies, and therefore there is no way to tell them apart using statistical dependencies alone.[5] Alternatively, if a graph is structured as follows:

$$A \rightarrow B \leftarrow C, \tag{5.4}$$

we have what is known as a *collider*. Unlike the examples in Eq. 5.2, variables A and C are

[5] Given that the chain and the fork are yield statistically equivalent data, it is worth considering the implications for testing for mediation structures.

actually already independent such that $A \perp\!\!\!\perp C$. A collider is also depicted in Figure 5.2(e), and the parallel vertical red lines depict the 'break' in statistical dependence between A and C. Furthermore, conditioning on B in this structure actually *induces* statistical dependence between A and C - a phenomenon known as explaining away (Pearl, 2009; Pearl, M. Glymour, and Jewell, 2016). A corresponding list of conditional independency statements for this collider is therefore:

$$A \not\perp\!\!\!\perp B, \quad B \not\perp\!\!\!\perp C, A \perp\!\!\!\perp C, \quad A \not\perp\!\!\!\perp C|B, \tag{5.5}$$

Variables are known as *ancestors* of downstream *descendants* if there exists a directed path between the variables. A direct descendent is also called a child, and the direct ascendant is called a parent. Note that conditioning on *descendants* of the variable B in the two graphs depicted in Eq. 5.2 can *partially* render A and C independent (because it essentially contains critical information from A via B). Similarly, conditioning on a descendent of the collider variable B in Eq. 5.4 can also render variables A and C *partially* dependent. Of course, two variables are either dependent or not, and the partial terminology is used here to communicate that the effect of conditioning is not as strong as would be the case using B itself, as opposed to one of its descendants. We can actually test for these conditional independencies using conditional independence tests (which, in the linear Gaussian setting are essentially partial correlations). These tests can then be used to *discover* the underlying structure in the data - a task known as causal discovery, for which many methods exist (M. Vowels, N. Camgoz, and Bowden, 2022).

Finally, returning to Figure 5.2(h), which was discussed above in relation to estimating the effect of treatment T on outcome Y given some confounders $\mathbf{X}$, we know that for the substructure $T \leftarrow \mathbf{X} \rightarrow Y$, we can achieve $Y \perp\!\!\!\perp T|\mathbf{X}$ in order to essentially simulate the structure of the graph for the RCT in Figure 5.2(g). In other words, by conditioning on $\mathbf{X}$ we 'block the backdoor' path of *confounding* statistical dependence which 'flows' from treatment to outcome by conditioning on $\mathbf{X}$. This leaves only the one statistical path, which is also the causal path

we care about. In this case, the statistical dependence is equivalent to the causal dependence we wish to estimate. Thus, we have used conditional independency rules to isolate the causal statistical dependencies, and disentangle them from the non-causal statistical dependencies.

5.3.4 Markov Blanket

The conditional independency rules introduced above can be used to define a Markov Blanket. Essentially, the blanket constitutes a set of variables which yield conditional independence between variables 'within' the blanket, and those outside it. The notion of a Markov Blanket confirms the idea that not all variables are necessarily needed to estimate or identify a particular causal effect. The implication of this is that if we have knowledge of a set of conditioning variables, other variables which are causally 'downstream' of these conditioning variables become effectively 'disconnected' from those which are upstream.[6]

Consider Figure 5.2(f) which depicts a Markov blanket around variables X and Y. The underlined variables $\underline{B}$, $\underline{D}$, and $\underline{E}$ constitute the Markov blanket - knowing or conditioning on these variables renders X and Y independent of variables A and C, which are outside of the blanket.

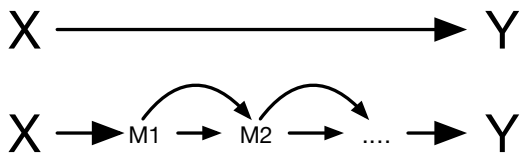
An SEM model can be reduced in size to comprise only the variables and paths necessary to estimate set of paths of interest. Considering, again, Figure 5.2(f), if we are only interested in the path coefficients proximal to the variables X and Y, we do not need variables A or C, thus reducing the number of estimated paths from ten (if we include the paths from unobserved U) to five. We discuss more opportunities below.

5.3.5 Projection

A cause-effect relationship can often be broken down into smaller and smaller subdivisions, until one starts talking about the effect of one molecule on the next to explain a simple game

[6] It is possible to have variables which fall into the set of defining Markov blanket variables but which do not need to be explicitly conditioned on. This can occur, for example, in the presence of a collider structure which may already render upstream variables (which are outside of the blanket) as statistically independent of those within the blanket, without conditioning (recall that conditioning on a collider can open up an otherwise 'closed' path.

Figure 5.3: An illustration of 'infinite mediation'.



Note. This figure illustrates that between any two cause-effect pairs, there exists an almost infinitely decomposable chain of intermediate mediators.

of billiards. As per Figure 5.3, each subdivision of the cause-effect relationships between X and Y could be represented as a mediating path with an infinite number of intermediate mediating paths. By consequence of the Markov assumption (described above) it is thankfully not necessary to model all these intermediate mediators, and it suffices to abstract to the key 'beginning and end points'. For instance, it is not necessary to know the intermediate position and velocity of a billiard ball (assuming these are well known), but it may be important to know when/if it changes course following a collision. One can, for example, reduce $X \rightarrow M \rightarrow Y$ simply to $X \rightarrow Y$ (C. Glymour, 2001, pp.40). Of course, if one is specifically interested in a mediating variable then one can collect the relevant data and explore the process (such examples are provided below). Of course, some reductions may yield an intractably blunt abstraction, or, in the extreme, a form of infinite causal regress (*e.g.* regressing all first causes to our birth or the beginning of time), and one might instead consider more modest examples, such as whether a treatment is mediated by some psychological mechanism(s). In this case, one can nonetheless reduce the problem (via projection) to an estimation of the total effect of treatment on the outcome, thus aggregating the intermediate direct and indirect effects and thereby reducing the complexity of the graphical representation.

5.4 Reducing SEMs - Worked Examples

In the previous section we reviewed four concepts which we will use for simplifying our SEMs without introducing bias into our effect estimates: (1) causal identification, (2) conditional independencies, (3) Markov Blankets, and (4) projection. In order to demonstrate these various techniques, we will walk through a number of examples which are presented in Figure 5.4.

For each example, we specify (a) a full DGP as our starting point which we assume to be true and complete ('Full DGP' in Figure 5.4), (b) a set of causal effects of interest, that must be identifiable for subsequent estimation ('Research Question' in Figure 5.4), (c) a minimal SEM (denoted Reduced in Figure 5.4), and (d) syntax for the R *lm()* function for a multiple regression. Five example DGPs are shown in Figure 5.4. Again, whilst we are not concerned with the estimation itself, note that one can choose to either use the SEM framework to estimate all the path coefficients in the resulting model, or one can undertake (possibly multiple) regressions to arrive at the same goal. In both cases, the graphical representation of the theory is what enables us to reduce the model in a way which does not invalidate the subsequent analysis (as well as increasing transparency, helping us to think more deeply and concretely about the causal process, etc.).

In practice, the graphical representation of our DGP will be developed using domain knowledge and/or causal discovery techniques (M. Vowels, N. Camgoz, and Bowden, 2022; M. J. Vowels, 2021; C. Glymour, K. Zhang, and Spirtes, 2019). For now, we provide general examples with a view to demonstrating the ways in which the concepts reviewed above can be used to reduce our SEM. Similarly, in practice the set of paths of interest will be determined by our research questions and our hypotheses. Note that it may be possible to simplify SEMs bearing in mind other techniques which are applicable to linear models (such as instrumental variables) (Bollen, 2019), but we focus on those techniques reviewed above because they are generally applicable to a much broader family of problems. Finally, it is worth remembering that if a set of variables and paths are not needed for the SEM, then we also do not need to collect these variables to begin with, thus saving additional time and expense which could be used to, for example, collect more samples of the variables that really matter. Note that some variables may not strictly be necessary for the estimation of the effect but may nonetheless be worthwhile including. For example, proximal causes of an outcome which do not interfere with our estimation of other desired causes can increase the precision/tightness of our estimates, and may therefore still be worth including (Cinelli, Forney, and Pearl, 2022).

Unobserved variables and/or latent constructs may also be integrated into the specification of the graph. In terms of the planning, these objects can be considered in the same way as other

observed variables, at least insofar as they relate to the estimation of the causal dependence we are interested in. One may find, for example, that the existence of certain unobserved variables fundamentally preclude identification (*i.e.*, the estimation of the target effect), perhaps because they induce a backdoor/confounding path between the 'treatment' and the outcome. Conversely, one may find that either certain unobserved variables, or particular latent constructs are not necessary for the identification of the target effect. We later consider a number of worked examples involving unobserved variables (Examples 3 and 4).

To motivate the examples, we will attempt to describe semi-plausible DGPs for psychological processes, but note that these examples are likely to be overly simplistic, and are only intended to illustrate the process. We will discuss each of the examples in Figure 5.4 in turn. Finally, in the supplementary material we also provide simulation results for DGPs 2-5 in Figures 5.7-5.9.[7]

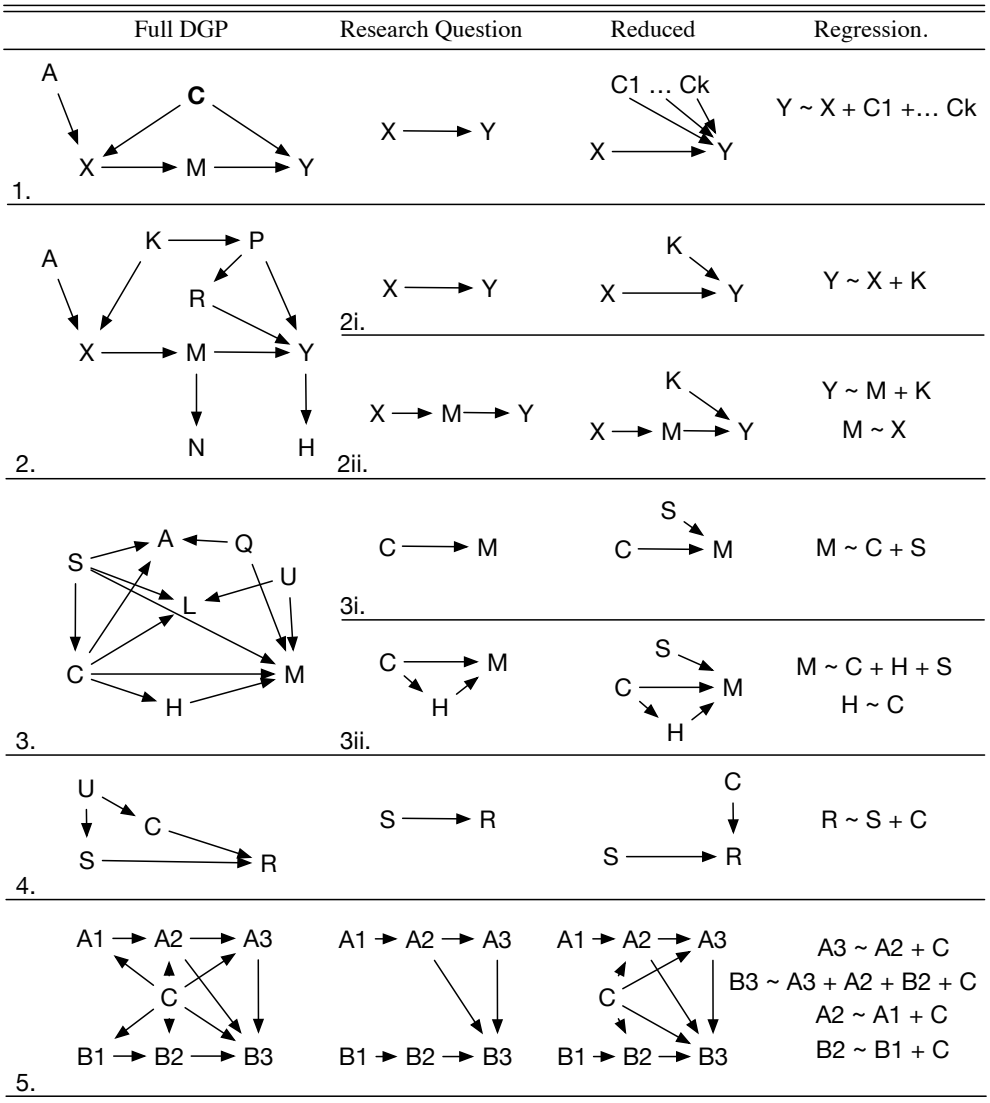
5.4.1 Example 1: Mediated Treatment

Starting with the first example depicted in Figure 5.4, let us begin by considering what this graph could possibly represent. Variable Y could be an outcome (*e.g.* depressive symptoms) for a therapy X, the effect of which is mediated by therapeutic alliance M. The set $\mathbf{C}$ represents covariates that influence the choice of therapy modality as well as the likelihood of recovery, and includes factors such as age, gender, history of mental health problems, and so on. Finally, variable A could represent a personal attitude which influences the choice of treatment but which does not influence whether the person recovers.

For this example, let us assume that our research question concerns estimation of the efficacy of treatment on the outcome, *i.e.*, $X \rightarrow Y$. The reduced model (denoted in Figure 5.4 as Reduced) requires three fewer paths to estimate this effect. Firstly, if we are not interested in the particulars of the mediated path $X \rightarrow M \rightarrow Y$ then we do not need to include $X \rightarrow M \rightarrow Y$, or to therefore collected data for M (afforded by the projection concept reviewed above). Secondly, even though there exists a spurious/confounding/backdoor path $X \leftarrow \mathbf{C} \rightarrow Y$, we do not *need* to estimate the actual path $X \leftarrow \mathbf{C}$ so long as we include the path $\mathbf{C} \rightarrow Y$. The inclusion of $\mathbf{C}$

[7] We omit simulations for DGP 1 because it represents a reduction of the other examples, and so including it is somewhat redundant.

Figure 5.4: Finding the reduced model.



Note. This figure presents a number of examples for taking the full 'true' Data Generating Process (DGP) and finding the reduced graph and minimal linear/logistic regression required to answer a given research question.

facilitates identification of the principal effect of interest $X \rightarrow Y$. Note that in this case we do not have to use SEM for the estimation procedure. Indeed, in this example we are not interested in the path coefficients linking $\mathbf{C}$ to Y either, even though these paths must be included to acknowledge the dependence that Y has on $\mathbf{C}$ and to block the backdoor path. Given we are only interested in the path from X to Y, we can simply run a multiple regression, using $\mathbf{C}$ as control variables and restricting interpretation to the coefficient on X. Note that the resulting *lm()* syntax contains only the two necessary components as predictors - X and the set of control variables $\mathbf{C}$.

Finally, we do not need to include A in the model (neither do we need to collect data for A) because it is not necessary for the causal identification of the target causal effect of interest. Adding the path $A \rightarrow X$ into the model is superfluous to the effect we are interested in.[8]

5.4.2 Example 2: Structured Controls

The first graph with structured controls is given as example 2 in Figure 5.4. We can consider the meaning of variables A, X, M, and Y to be the same as in Example 1, that is attitude, treatment, treatment-outcome mediator, and outcome, respectively. The difference now is that we also have a mediation child N, an outcome child H, and a structured set of control variables K, P, and R. If, as indicated in example 2i, we are only interested in estimating the effect of X on Y then, as in the first example, we can ignore A and M. Similarly, we can also exclude N and H for our reduced model, as their existence in the DGP does not change the principal relationship we are interested in.

There still exists a backdoor path through the control variables K, P, R, and Y, and so we need to understand which of the associated variables and paths to include in our reduced model to adjust for this spurious path. There exist the following options which block this path: $K \rightarrow Y$, $K \rightarrow P \rightarrow Y$, and $R \rightarrow Y \leftarrow P$. Note that $R \rightarrow Y$ is not an option by itself because this would leave the path through $P \rightarrow Y$ open. Note also that we do not need to estimate the path $K \rightarrow X$ because we are not interested in this effect. Thus, overall, our initial/complete model

[8] Indeed, its inclusion can even increase the standard errors on the effect of $X \rightarrow Y$ because it makes it 'harder' to disentangle the variance in Y that stems from X and the residual variation of A which is also contained in Y.

reduces to the estimation of only two paths (reduced from ten), as in the previous example. The linear regression also remains equivalent.

If our research question involved the estimation of the mediation, as in example 2ii in Figure 5.4, then the only change to the model needs to be the inclusion of the mediation $X \rightarrow M \rightarrow Y$. The linear regression now involves two stages to decompose the problem into two sets of paths (one from $X \rightarrow M$, and the other comprising the paths $M \rightarrow Y$ and $K \rightarrow Y$).

5.4.3 Example 3: Colliding Controls

One might be forgiven for thinking that the safest thing to do with a set of control variables is to always include them in the model to make sure we are blocking the backdoor paths. In the previous example, for instance, we could just play it safe by including $\{K, P, R\}$. However, example 3 in Figure 5.4 shows that some putative control variables may include collider structures. Let us consider that variables C, M, and L are class-size, math exam score, and language exam score, respectively. H represents a mediator such as whether a student does their homework, S represents Social Economic Status (SES) - perhaps children with higher SES attend schools with smaller class sizes and have better grades overall - U represents an unobserved attribute of intelligence Q a measured attribute of intelligence, and A musical ability.

Based on example 3i we are interested in the effect of class size on math exam score. It might be tempting to include the paths concerning the other related scores (such as language score, or musical ability). In the case of musical ability, we *could* include the paths $C \rightarrow A \leftarrow Q \rightarrow M$ without causing any problems, but it doesn't actually help us estimate the effect we are interested in. Indeed, the collider structure $C \rightarrow A \leftarrow Q$ prevents any backdoor information affecting our estimation of $C \rightarrow M$, so we do not need these paths for causal identification. Another collider exists between $C \rightarrow L \leftarrow U \rightarrow M$, and even though the structure is the same, the fact that U is unobserved means we cannot and should not include L in the model. Indeed, if L were to be included (without U as U is unobserved) we would induce a spurious path linking C to M through L and U. Thus, whilst these might appear to be tempting control variables which we

might think would, at best increase precision and at worst do nothing, in fact they should not be included owing to the collider structure with an unobserved variable.

We have no need to include paths relevant to A or L in our model. Including the path $Q \rightarrow M$ may improve the precision of our desired estimate, but it is not necessary. The partial mediation through H, if not part of our research question, does also not need to be included. The only path we have to be concerned about is $C \leftarrow S \rightarrow M$, and we can deal with the induced statistical path by simply including the path $S \rightarrow M$. In this case, the the reduced model contains two paths, whilst the full model (including the unobserved paths) involves thirteen. The corresponding linear regression is equally simple, and only includes C and S as predictors.

If we are interested in the partial mediation of class size, homework, and math exam score, then we can simply augment the reduced model from example 3i to include this additional structure. The linear regression also changes to accommodate the estimation of the additional paths, as with example 2ii.

5.4.4 Example 4: Simple Unobserved Confounding

The fourth example is relatively straightforward. Here, R, S, and C could represent relationship satisfaction, partner support, and communication style, respectively, where the unobserved confounder U between support and communication. The unobserved confounder induces a non-causal statistical dependence between S and R through C, and the reduced model therefore needs to include the path $C \rightarrow R$. The linear regression, similarly, needs only S and C as predictors.

5.4.5 Example 5: Longitudinal Dyadic Effects

The final example concerns a longitudinal dyadic process, whereby variables for *e.g.* relationship satisfaction for two individuals A and B are collected at three timepoints, but there exist intermediate opportunities where confounding could occur. This confounding could represent, for example, shared stressful events. The target causal effects involve all of the 'actor effects' (that is, autocorrelation in each individual's variables which results in similar values across

consecutive timepoints), as well as two partner effects from $A_2 \rightarrow B_3$ and a 'concurrent' effect $A_3 \rightarrow B_3$.[9] This example demonstrates when the use of SEM may be less complicated than undertaking a series of multiple regression tasks; our research question concerns the estimation of six separate causal effects, all of which have to be identified.

We do not need to estimate the paths $C \rightarrow A_1$, so long as we include the path $C \rightarrow A_2$, which enables us to block the backdoor path from A_1 to A_2 via C and thereby identify the effect $A_1 \rightarrow A_2$. For the same reasons, we do not need to estimate the path $C \rightarrow B_1$. In this example, we are not able to make any data collection savings (*i.e.*, we need to collect all variables), even though some of the path coefficients are not needed for estimation of the principal causal effects of interest.

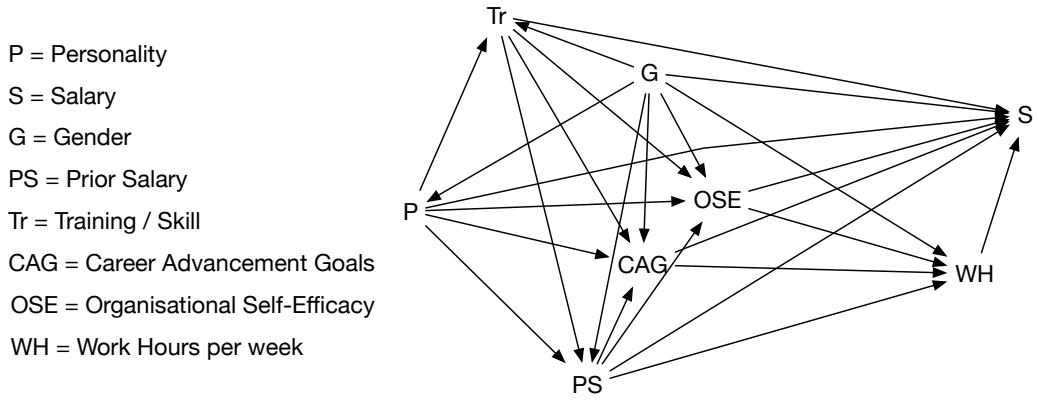
5.4.6 Real World Example

To motivate the application of the techniques to non-synthetic examples, we have chosen a graph adapted from a paper published in the domain of business psychology and organizational behaviour. The graph is shown in Figure 5.5, and was presented to test the relationship between personality ('P' in the graph), and salary ('S'). First, let us consider the model required in the case where our research question solely concerns $P \rightarrow S$. The only non-causal path from personality to salary, assuming the graph shown in Figure 5.5, is via gender: $P \leftarrow G \rightarrow S$. The reduced graph is shown in Figure 5.6i. In this case, the simple regression $S \sim P + G$ would suffice, and the graphical representation of the SEM would be $P \rightarrow S \leftarrow G$. Once again, it is only possible to confirm this if we already have a representation of our model which enables us to identify the required control variables.

In the original work (Spurk and Abele, 2011), the researchers were specifically interested in a double-mediation by occupational self-efficacy ('OSE') and career advancement goals ('CAG'), which represent the first set of mediating variables, and working hours ('WH') which represents a second mediation of the effect of personality on salary. In this case, all variables

[9] Even though the causal framework does not strictly admit simultaneity (there must be some time delay between the case and the effect), we assume that this concurrence is permitted according to the data collection procedure (*i.e.*, within wave three, partner A can influence partner B with some arbitrary time delay which is not distorted by the otherwise cross-sectional nature of the data collection methodology).

Figure 5.5: Real-world example graph.



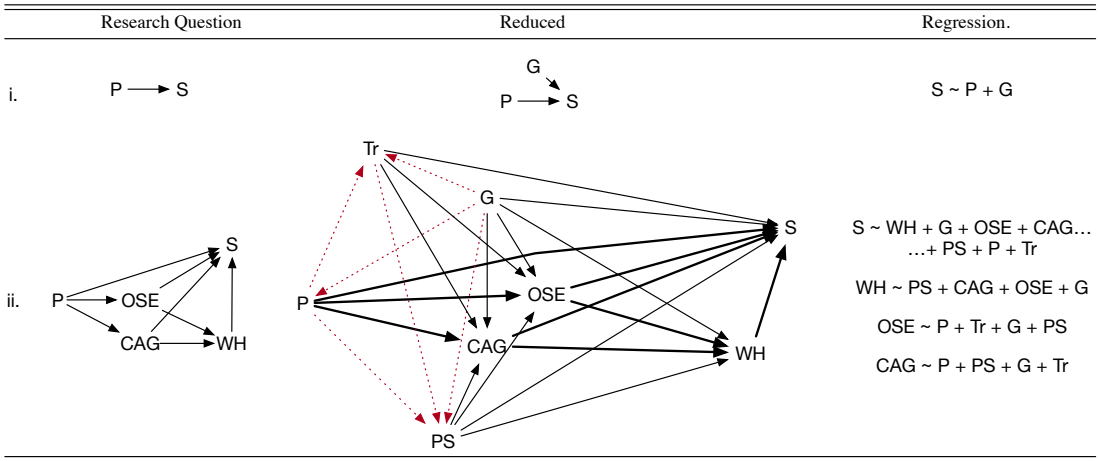
Note. Real-world example graph adapted from Spurk and Abele (2011).

are required for the analysis, and no savings can be made at the data collection stage, but we can nonetheless reduce the number of paths to be estimated. The reduced graph is shown in Figure 5.6ii. Identifying this reduced solution by eye is already becoming challenging, and automated tools (such as the one provided in supplementary material) are helpful in ensuring the reduction is correct. In addition, identifying the set of multiple regressions which can yield unbiased estimates of each of the target paths is also quite involved, and this example demonstrates how the SEM estimation framework might provide a more convenient alternative. In any case, it can be seen that six out of a total of 24 paths were not required.

5.5 Discussion

We have provided a number of didactic examples showing that if we are presented with a specific question regarding a relatively complex process, we can simplify our SEMs considerably. The simplification process takes advantage of a number of graphical rules, and does not introduce any additional assumptions to those which already apply to the full model. Furthermore, researchers are also free to choose whether they actually wish to estimate all the path coefficients using SEM framework itself, or whether a multiple regression would be more straightforward. Indeed, in cases where only a single causal effect needs to be estimated, one might consider using the graphical representation first, and then estimating it using a multiple regression instead. In

Figure 5.6: Reduced real-world example graph.



Note. Reduced real-world example graphs for the real-world DGP assumed in Figure 5.5. Bold black lines are those key to a multiple-mediation research question, whereas red dashed lines are those that may be excluded from a graphically specified SEM without affecting the estimation of the target paths.

this work we provided both the graphical representation of the SEM that one needs to estimate in order to answer a research question relating to one or more causal effects, as well as the equivalent multiple regression equation(s).

In one of the demonstrative examples, an SEM with upwards of thirteen paths was reduced to only two. The simulation results provided and discussed in the supplementary material highlight unsurprising improvements in adjusted model fit metrics (unsurprising because simpler models are penalised less than complex models according to such metrics). Importantly, note that the simplification process does not bias the effect size estimates.

Even without the simplification process, translating a psychological theory into a graph is a worthy exercise, particularly when undertaken *before* the data collection stage. It helps us be transparent and unambiguous about our model and assumptions, increases specificity for preregistration, and can highlight potential methodological challenges and difficulties before any resources have been expended. It may even highlight cases where estimation is not possible, and this relates to the problem of causal identification. For example, if there exists an unobserved confounder between X and Y in the graph $X \rightarrow Y$, *i.e.* $X \leftarrow U \rightarrow Y$, the causal effect cannot be estimated because the non-causal statistical association induced by the confounder cannot be adjusted for without access to U. These problems can, again, be seen by an inspection of

the graph, and it is worthwhile identifying these problems sooner rather than later. In practice, such problems may be common, and either a researcher must do all they can to account for the possible unobserved confounders, or they must assume that a sufficient number have been already collected to assume that the problem is 'ignorable' (Pearl, 2009). In general, it is important to remember that the goal of estimating causal effects rests on a number of strong (and often untestable) assumptions. However, it is only by taking causality seriously that we can understand what these assumptions are and whether they are reasonable.

5.5.1 Limitations

We have used SEM throughout the text because researchers in psychology may be familiar with this framework (Blanca, Alarcon, and Bono, 2018). Furthermore, if they wish/need to estimate latent variables, the SEM framework readily facilitates this. Note, however, that SEM is generally considered to be an estimation framework, rather than a means to graphically represent one's causal theory. Furthermore, SEM usually assumes linear (or at least pre-specified) functional relationships between variables. Fortunately, and as we briefly discussed earlier, all the rules and techniques discussed in this work belong to a broader class of graphical model known as Directed Acyclic Graphs (DAGs). DAGs do not make assumptions about the parametric (*e.g.*, normally distributed vs. non-parametric) form of the variables, nor about the functional (linear vs. non-linear) form relating variables. This means that when one uses our proposed method to construct and subsequently simplify a graphical structure, they can also consider themselves to be working directly with a DAG. If the researcher then wishes to avoid making assumptions about the functions and distributions, they do not have to use the SEM framework to do the estimation, but can instead use non-parametric regression or machine learning techniques (a discussion about which is beyond the scope of this paper). Indeed, another reason that we provide the multiple regression syntax is because its specification can be generalized relatively straightforwardly to non-parametric settings. For example, the specification of the regression $Y \sim X + C$ relates to the estimation of $\mathbb{E}[Y|X, C]$, which is the conditional expectation of Y given X and C. The conditioning set given on the right hand side of the tilde in the regression syntax, or the right hand side of the conditional expectation, are the variables/predictors in the

regression which are being used to identify the causal effect(s) of interest, and this can be done in both linear parametric as well as non-linear, non-parametric settings.

The reduction which is achievable depends on the research questions being asked, as well as the requirements of the researcher. We foresee that some researchers may wish to collect more variables than are strictly required for identification to future-proof their datasets, thereby facilitating the testing of currently unspecified hypotheses. The collection of extra variables can not only provide the opportunity for researchers to answer potentially unforeseen research questions, but it also enables researchers to include 'hedge' variables, in cases where the theory specification is uncertain and researchers do not want to risk variable omission. Indeed, if the researcher is contending with multiple hypothesized graph structures, they may wish to avoid putting all their eggs in one basket by collecting only the smallest set of variables relevant for one particular graph and one particular research question. Furthermore, researchers may also be able to promote the project, and thereby indirectly achieve higher statistical power and better measurement precision, if they agree to collect variables beyond those which are strictly necessary for the specific research question but which may be relevant to collaborators. Finally, by 'over-collecting' variables, they may also open up opportunities to undertake causal discovery - a data driven approach to the validation of putative causal structures. Without the extra variables, researchers would be somewhat stuck with what they have. In any eventuality, being able to determine which variables are strictly necessary for a particular research question is not only helpful in *optionally* streamlining data collection, but also in ensuring that none of the essential variables are otherwise excluded (even in the case where many non-essential variables are collected).

Finally, researchers should be mindful that the success of the approach rests on the degree of correct specification achieved when the DGP model is constructed.[10] However, this limitation applies to *all* statistical approaches which concern the estimation of interpretable / causal effects, and this approach does not alleviate the consequences of model misspecification. Furthermore,

[10]Note that the use of a reduced graph reduces the chances that the associated structural components are misspecified (there are fewer opportunities for misspecification in a smaller graph). Put differently, if I have the choice between estimating two effects or four effects, the estimation of four effects puts stronger requirements on identification than the estimation of two does (requiring, as it does, all four effects to be identified, rather than only two).

reducing model complexity may reduce the precision of the estimation because less explanatory power may be available to estimate an effect. For instance, if a set of risk/precision variables are included (risk variables being parents of the outcome), the ability to estimate the target effect does not change, but the efficiency may improve. This is because risk/precision variables can explain variation in the outcome which derives from other causes than the one of interest, and the model is thereby able to obtain more reliable estimates for the target effect. This is evidenced by a review of the simulation results for the p-values, as well as the error bars on the coefficient estimation plots in the supplementary material. This downside is somewhat offset by the possibility that, with a simpler model, a larger sample size may be acquired for equivalent cost. For example, if the simplification process indicates that a number of constructs with large inventories are no longer required, we may gain back significant data collection time which can be put towards the recruitment of more participants. Such possibilities therefore enable us to increase statistical power for estimating the effects we really care about. Furthermore, the specification of larger models increases the chances of misspecification (simply put, in the specification of larger graphical models, there is more opportunity for error). Reducing the model and being specific and less ambitious about the number of primary effect sizes of interest (as opposed to wishing to estimate as many effects as possible) increases the likelihood that, at the end of the project, we have estimated something meaningful.

5.5.2 Related Options

It is worth noting that other approaches for streamlining data collection and reducing study cost, such as the tools for the development of short-form scale design (Greer and J. Liu, 2016; G. Smith, Combs, and Pearson, 2012) and planned missingness design (J. Wood et al., 2019). In the case of the former, researchers can use statistical techniques to identify reduced scale designs which provide similar performance in terms of certain scale quality measures, such as validity. In the case of the latter, there are a number of planned missingness techniques which enable researchers to amortize data collection cost over the course of a longitudinal design, or to leverage statistical associations to compensate for foreseen missing data. These methods differ significantly from our proposal, and can even be used in combination with ours. Specifically,

the short-form scale design approaches are motivated by the fact that there may exist redundant information in a scale which is already represented by other items (or combinations, thereof). In contrast, our approach is concerned with the assumptions about and formal specification of the causal structure of data generating process itself, and does not concern redundancies in the scales used to measure the constructs/variables within this structure. The data generating process can therefore be considered independently of scale-item redundancy. Similarly, planned missingness techniques include split form designs (Raghunathan and Grizzle, 1995) which split large questionnaires into multiple smaller blocks, each of which is completed by participants at different stages of a longitudinal design. Alternatively, multiple imputation provides researchers with a way to leverage statistical associations to compensate for instances of missing data. Again, in contrast with our proposal, this approach does not consider the opportunities already implicit in the specification of our theory.

5.5.3 Conclusion

In summary, graphical representations of our theories provide us with an opportunity to encode our domain knowledge about a particular phenomenon of interest. In this paper we showed that, by using graphical modeling rules (in particular, the concept of conditional independencies), we can significantly shrink the required causal structural model without affecting the validity of the associated estimates, thereby reducing the required sample size and enabling us to redirect resources and funds towards the collection of variables which are critical to answering the questions we care about.

5.6 Supplementary: Simulation Results

The purpose of the simulation is to illustrate the differences in χ^2, Root Mean Squared Error of Approximation (RMSEA), Comparative Fit Index (CFI), Mean Absolute Error (MAE) and p-values, between two models which differ in complexity but which are otherwise correctly specified (with respect to the true, underlying DGP). It is worth noting that χ^2 is known as an 'absolute' fit index, and is not adjusted for model complexity. A lower χ^2 value indicates

better fit and provides a measure of how much our sample covariance matrix differs from our fitted covariance matrix. In contrast, RMSEA adjusts for the model complexity (favouring model parsimony), and here a lower value is preferred. Finally, CFI is not adjusted for model complexity, and higher values are preferred. For more information on these metrics, readers are pointed towards works by Maruyama (1998) and Hoyle and Panter (1995).

It is important to note that under these conditions (and when researchers use the process/tools presented in this work), the causal effect size estimates are unbiased regardless of whether the full model or the reduced models are used. As such, even though the use of these tools can have an effect on the standard errors (and therefore also the p-values and null-hypothesis significance testing), it does not affect the large-sample performance of the model. Indeed, this is evidence in the lower four plots of Figure 5.7, which confirm that the choice of model does not affect the effect size estimates (all are unbiased). Nonetheless, it is important to understand the possible impact on the various model metrics to understand that two different correctly specified models can yield different finite-sample behaviours. These differences are discussed in more detail in this section.

Simulation results for DGP examples 2-5 in Figure 5.4 are shown in Figures 5.7-5.9. We use the *sem* function in the *lavaan* library (Rosseel, 2012) to estimate a single target effect for each variant. For the MAE and the p-values, we provide results for a single effect of interest. For example, for the DGP research question 2ii in Figure 5.4, we specify the SEM models given in the 'Full DGP' and 'Reduced' columns and generate MAEs and p-values for the total effect of X on Y. Similarly, for DGP research question 3ii, we specify the SEM models given in the 'Full DGP' and 'Reduced' columns, and generate MAEs and p-values for the total effect of C on M. Finally, for example 5, we specify the SEM models given in the 'Full DGP' and 'Reduced' columns, and generate MAEs and p-values for the total effect of $A1$ on $B3$.

For each of the example DGPs, we generate data across a range of sample sizes (10-200), and for each sample size we undertake 100 simulations. The results of these 100 simulations are used to derive means and standard deviations for each of the metrics, thus allow us to compare the results when specifying the full DGP model compared with the reduced models.

Starting with the results for the model fit metrics χ^2 in Figure 5.7, we see that for DGPs 2-4 the

reduced models have better fit (lower χ^2 indicates better fit). This comes as no surprise because here the complexity of the model impacts our ability to reduce error for the path coefficients we are estimating (reducing the degrees of freedom). For similar reasons, it is also not surprising that the differences for the full and reduced models for DGP 5 were not different - the reduced model did not differ greatly in its reduction of complexity. In this sense, reducing the complexity of the model can have an effect on the resulting χ^2, in such a way that yields a value which is considered desirable (of course, in practice we should specify theories based on more than just the resulting fit-statistics).

In Figure 5.7 we provide estimates for the target effect size 'Coefs', on top of the true effect size 'True Coef'. Importantly, the results confirm that the simplification process does not bias the estimates - all model variants correct estimate the effect size.

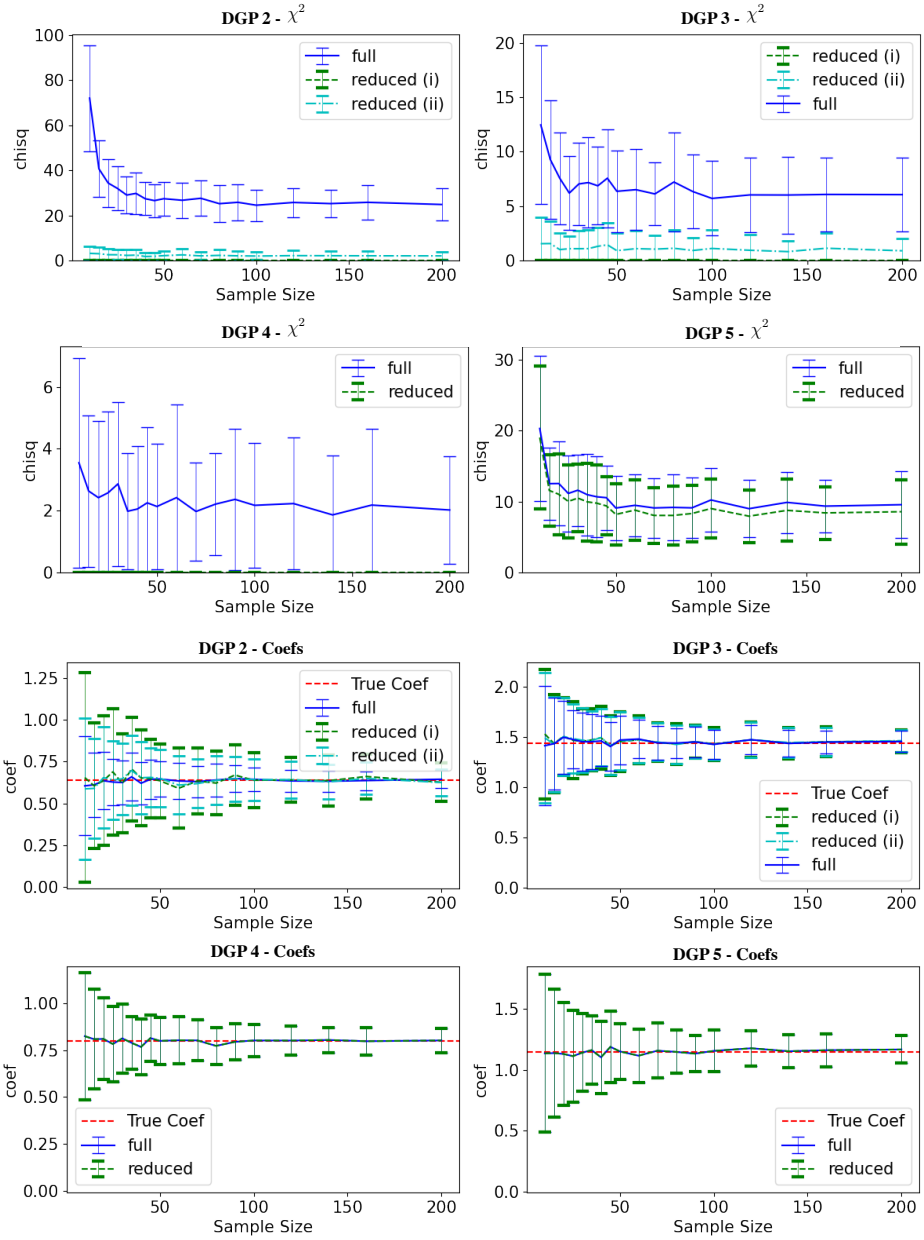
Results for CFI (higher is better) and RMSEA (lower is better) are shown in Figure 5.8. Once again, the smaller models are preferred and yield higher CFI values. This again comes as a consequence of the complexity of the larger models and the concomitant impact on estimation. This notwithstanding, as the sample size increases, the results converge fairly quickly. The RMSEA results indicate a great improvement with the use of the reduced models, particularly for smaller sample sizes. This is not surprising beacuse RMSEA is an adjusted metric, and so the results are consistent with the expectation that lower RMSEA values are associated with smaller models.

Finally, the p-values and MAEs for the target effect size estimates are shown in Figure 5.9. For DGP 2 (top left plot), the p-values are higher for the reduced model than the complete model. This is consistent with the expectation that the inclusion of more variables can help increase the precision of our estimates. Indeed, in general we expect that the inclusion of variables into a structural equation model will reduce the standard error and, by the mathematical expressions relating these quantities, also reduce the p-values. However, this is only reliably the case if the model is correctly specified, and the reason it happens is because we are able to partial out the variance more completely. For example, consider the graph $X_1 \to Y \leftarrow X_2$. Here, Y has two causes, but let's say that we actually only care about the link $X_1 \to Y$. In this case we have two options: create an SEM which includes $X_2 \to Y$ (in addition to the $X_1 \to Y$ link), or create

an SEM which does not. Note, however, that the inclusion of $X_2 \rightarrow Y$ can help us estimate $X_1 \rightarrow Y$ because it partials out variance in Y which, in a finite sample, might otherwise be attributable to X_1. Unfortunately, in practice it may not be as simple as this, because every time we include a new variable and a new path, we also increase the chances that we incorrectly specify the graph. Thus, whilst the option to reduce standard error by the inclusion of more paths is perhaps still a good thing to consider/understand in general, doing so requires us to be more and more confident that our specification is correct as we include more and more paths in our model.

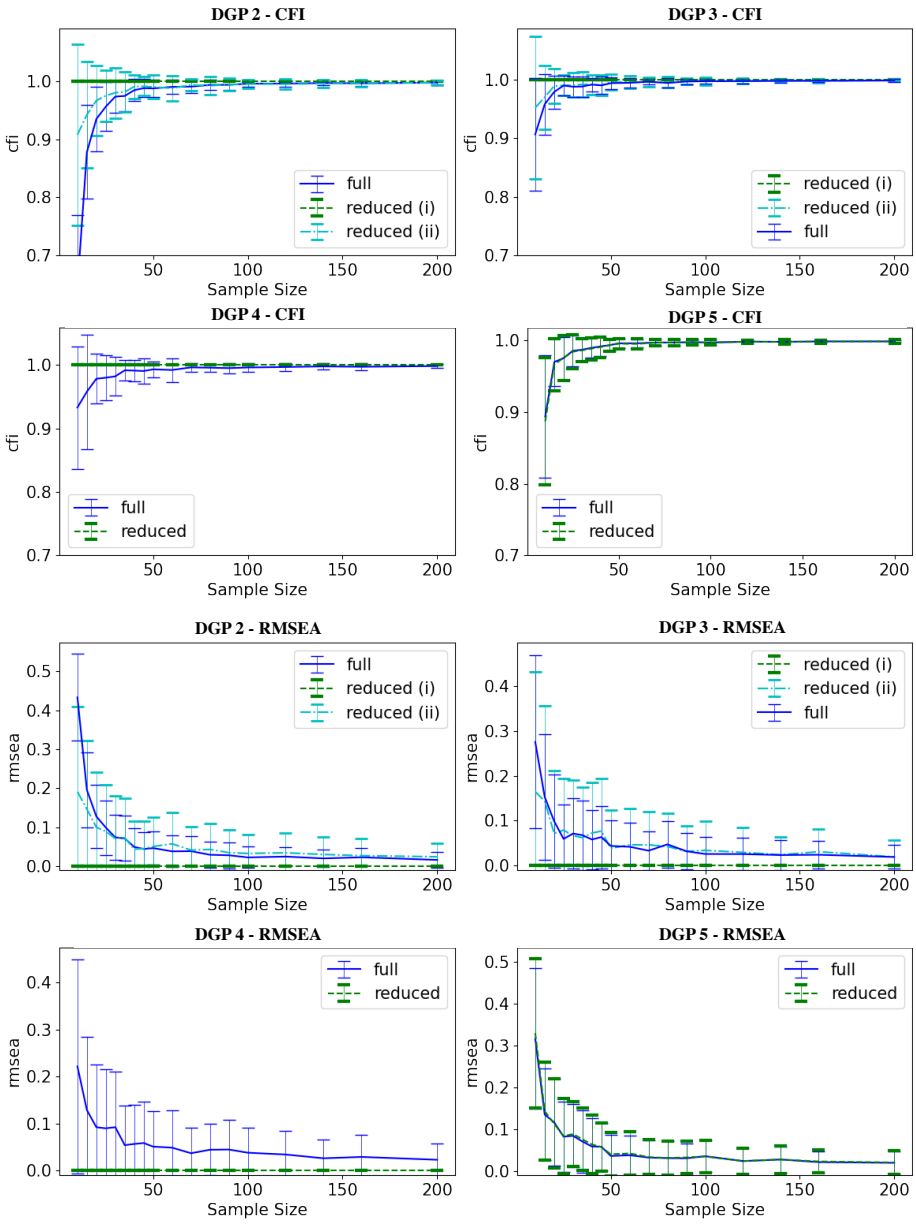
Returning to the examples in the figure, the reduced model in DGP 2i only includes two effects of the outcome Y, which is X and K. However, other more proximal variables P and R exist, and their inclusion would improve the quality of the estimate. In this case, R and P would be doubling as both control variables (adjusting for the backdoor path from X to Y, as well as variables which aid in precision (Cinelli, Forney, and Pearl, 2022). Note also that the standard deviation of these *p*-values is higher, indicating greater variation across simulations. This increased variance also results in a higher MAE, which is also evidence in the DGP2 - MAE plot in Figure 5.9 (third row, first column). Thus, even though the effect size estimates will be unbiased (owing to correct specification of the reduced model with respect to the full DGP), the removal of explanatory variables can impact the precision of the estimates. In order to compensate for this, one can choose to retain variables which have explanatory power so long as their inclusion does not contradict the full, underlying model. DGP 2 represents a useful example insofar as variables R and P can be included (optionally in addition to K), to help explain the effect of X on Y.

Figure 5.7: Simulation χ^2 and Coefficient Estimation Results.



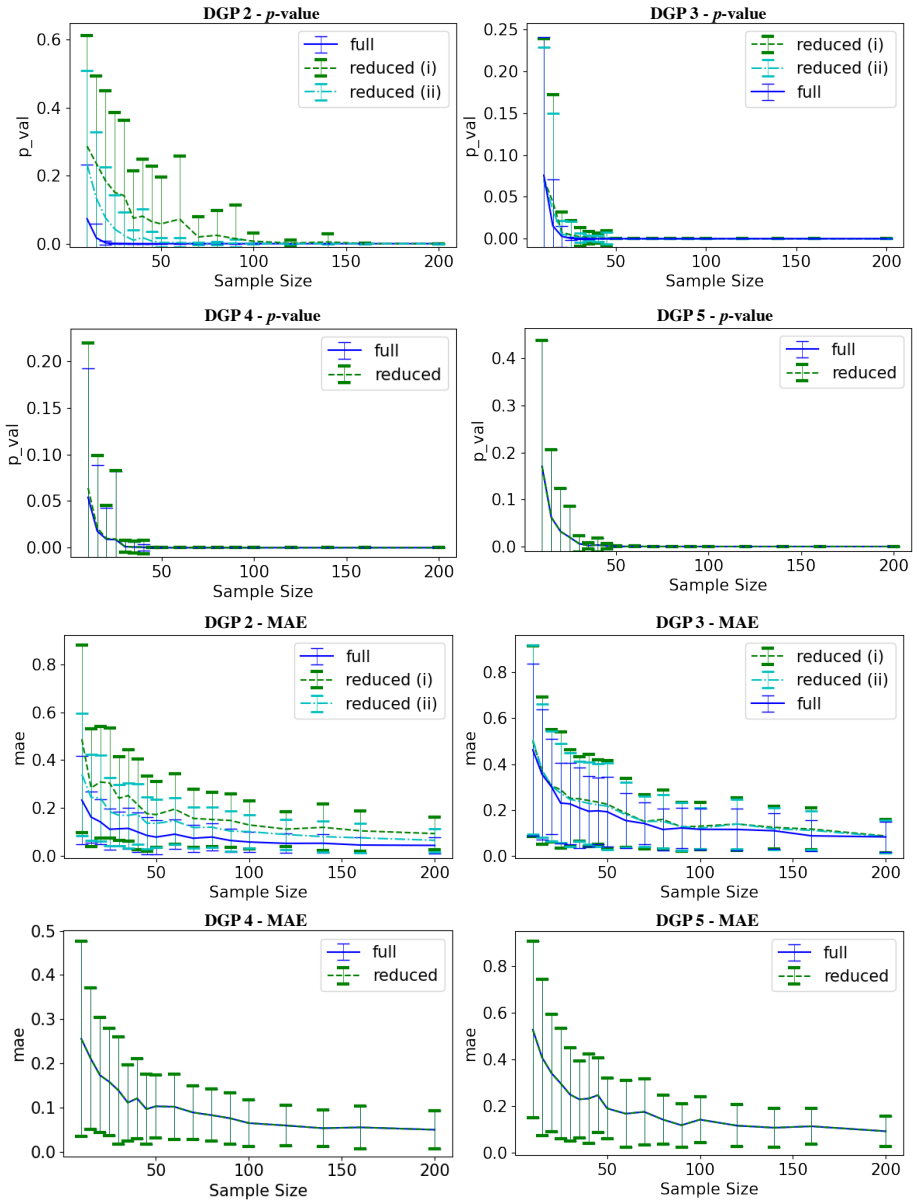
Note. Averages and standard errors over 100 simulations with varying sample sizes for χ^2 and estimated coefficient values for data generated from Data Generating Processes (DGPs) 2-5 in Figure 5.4.

Figure 5.8: Simulation CFI and RMSEA Results.



Note. Averages and standard errors over 100 simulations with varying sample sizes for Comparative Fit Index (CFI) and Root Mean Squared Error of Approximation (RMSEA) for data generated from Data Generating Processes (DGPs) 2-5 in Figure 5.4.

Figure 5.9: Simulation p-value and MAE Results.



Note. Averages and standard errors over 100 simulations with varying sample sizes for p-values and Mean Absolute Error (MAE) for data generated from Data Generating Processes (DGPs) 2-5 in Figure 5.4.

CHAPTER 6

Trying to Outrun Causality with Machine Learning: Limitations of Model Explainability Techniques for Exploratory Research

> *"This [causal] ingredient should allow computer systems to choreograph a parsimonious and modular representation of their environment, interrogate that representation, distort it by acts of imagination and finally answer "What if?" kind of questions. Examples are interventional questions: "What if I make it happen?" and retrospective or explanatory questions: "What if I had acted differently?" or "what if my flight had not been late?" Such questions cannot be articulated, let alone answered by systems that operate in purely statistical mode, as do most learning machines today."*
>
> Pearl (2018)

The content of this chapter is drawn from the following publication:

Vowels, M.J., Under Review. Trying to Outrun Causality with Machine Learning: Limitations of Model Explainability Techniques for Exploratory Research.

Abstract: Machine Learning explainability techniques have been proposed as a means for psychologists to 'explain' or interrogate a model in order to gain an understanding about a phenomenon of interest. Researchers concerned with imposing overly restrictive functional

form (*e.g.*, as would be the case in a linear regression) may be motivated to use machine learning algorithms in conjunction with explainability techniques, as part of exploratory research, with the goal of identifying important variables which are associated with / predictive of an outcome of interest. However, and as we demonstrate, machine learning algorithms are highly sensitive to the underlying causal structure in the data. The consequences of this are that predictors which are deemed by the explainability technique to be unrelated/unimportant/unpredictive, may actually be highly associated with the outcome. Rather than this being a limitation of explainability techniques *per se*, we show that it is rather a consequence of the mathematical implications of regression, and the interaction of these implications with the associated conditional independencies of the underlying causal structure. We provide some alternative recommendations for psychologists wanting to explore the data for important variables.

6.1 Introduction

Researchers in psychological and social science seem to be aware that machine learning cannot be used 'naively' to yield causal quantities, and that causal conclusions cannot be drawn from the results of predictive algorithms. This is evidenced by the adage 'correlation is not causation', which is well-baked (almost to a fault - see M. Hernan, 2018b) into the research zeitgeist, and has been widely discussed in commentaries by Yarkoni and Westfall (2017), Pearl (2009), M. J. Vowels (2021), Grosz, Rohrer, and Thoemmes (2020), and Shmueli (2010). Seemingly without contradicting this mantra, it is becoming increasingly popular to use machine learning techniques to infer 'important' or 'predictive' variables. Indeed, random forests with explainability methods such as random forest importances, or Shapley values (Lundberg, G. Erion, et al., 2020) have seen application in the domains of psychology (L. Vowels, M. Vowels, and K.P. Mark, 2021; L. M. Vowels, M. J. Vowels, and K.P. Mark, 2020; L. M. Vowels, M. J. Vowels, and K.P. Mark, 2021a; Joel, Eastwick, Allison, et al., 2020), genetics (Goldstein, Polley, and Briggs, 2011), epidemiology (Orlenko and Moore, 2021; Khalilia, Chakraborty, and Popescu, 2011), drug-discovery (Jiménez-Luna, Grisoni, and Schneider, 2020), and many others (C. Strobl et al., 2007). Other commentaries have encouraged similar practice. For instance, Yarkoni and Westfall (2017) discuss ways to interpret predictive machine learning models for psychology,

and argue that 'psychologists stand to gain a lot by relaxing their emphasis on identifying the causal mechanisms governing behavior, and focusing to a greater extent on predictive accuracy'. Similarly, M. J. Vowels (2021) recommends machine learning in order to avoid imposing restrictive, unrealistic limitations on the relationships between variables.

Given a general trend towards the incorporation of machine learning and explainability techniques into empirical research in psychology, it is important to understand the associated behavior of such techniques and whether they can be used productively to guide research. Whilst researchers may be interested in identifying predictive variables to inform the design of intervention or to guide theory development and future research, explainability techniques may actually yield conflicting or unhelpful evidence. In fact, due to the causal structure underlying the data generating process, the identification of predictive variables depends heavily on which other variables are included in the model. Machine learning algorithms are just as sensitive to the 'partialling' out of variance deriving from other variables as linear regression is, and yet this sensitivity is not well acknowledged by psychologists wishing to use these techniques to guide their research. Without strong prior knowledge of the underlying structure, this sensitivity makes it difficult to use explainability and variable importance techniques to draw conclusions about anything other than the behavior of the algorithm itself, and we would thus question the meaningfulness of any associated conclusions.

In this paper we adopt a causal perspective to help us understand and explain the limitations of machine learning and model explainability techniques. In particular, we examine the sensitivity of these techniques to the underlying structure of the Data Generating Process (DGP). We empirically demonstrate the almost inescapable dependence that linear models (and their coefficients), random forests (both variable importance measures and Shapley values), and MultiLayer Perceptrons (and the associated Shapley values) have on the underlying structure, highlighting how highly correlated/predictive variables may nonetheless be deemed unimportant, even when using powerful ML algorithms and state-of-the-art explainability techniques. Whilst the sensitivity of the coefficients of linear models can be well understood given knowledge of the underlying structure (M. J. Vowels, 2021), the fact that this sensitivity translates to machine learning explainability techniques has, to the best of our knowledge, not been described

before. Our conclusion is that it is not possible to 'outrun causality with machine learning', and that it always important to understand the possible interactions between the algorithm and the underlying causal structure in the data, whenever our goal is to use the predictor importances to guide our research and theory development.

The paper is structured as follows: In the Motivation section, we review some of the discussion surrounding the use of machine learning approaches to research in psychology and social science. In the Background, we review some relevant background theory relating to causality, Directed Acyclic Graphs, *d*-separation, regression, random forests, multilayer perceptrons, and explainability. Readers already familiar with this background can skip this and proceed directly to the Methods section, where we provide details on the experiments, including the datasets, algorithms, and explainability techniques. Then, in the Results section[1], using three different datasets of increasing structural complexity, we demonstrate how both random forest importances and Shapley value techniques for machine learning algorithms cannot be used to reliably infer anything about the presence of correlations/associations in the data. Following a discussion of the results, we discuss a number of methods which could be used to develop an understanding of a phenomenon at the exploratory stages of a research project. Finally, we summarise the work in the Conclusion.

6.2 Motivation

Psychological phenomena are well known to be complex, making causal inference and causal discovery exceptionally challenging (Meehl, 1990; Eronen and Bringmann, 2021) even in experimental scenarios, where causal interventions often suffer from 'fat-handedness' (Eronen, 2020). Accordingly, it is often difficult to properly develop and test theories, resulting in theories which are assumed to be true *a priori* and which are underspecified and often vague (Scheel et al., in press). Simultaneously, it is well accepted that theories are essential to our science (Oberauer and Lewandowsky, 2019; VanderWeele, 2020; Muthukrishna and Henrich, 2019b; Fiedler, 2017). Recently, and perhaps with some acknowledgement of the difficulties associated with

[1]Complete code can be found here: `https://github.com/matthewvowels1/ML_structural_interactions`

taking causal approaches, researchers have proposed taking a predictive approach to help validate and test theories utilising machine learning techniques. This approach was recommended by Yarkoni and Westfall (2017), who argued that it can be used to 'help gain a deeper understanding of the general structure of one's data.' Similarly, M. J. Vowels (2021) argued that machine learning can be used to explore the data to identify strong associations and predictive variables. Researchers have already started implementing this advice. For example, in large scale analysis, Joel, Eastwick, and E.J. Finkel (2017) established sets of variables which are robustly predictive of relationship quality, with the aim of using these results to guide future modeling; L. M. Vowels, M. J. Vowels, and K.P. Mark (2021b) and L. M. Vowels, M. J. Vowels, and K.P. Mark (2021a) identified predictive variables and the associated per-variable importances for sexual desire and infidelity, and suggested that these variables could be focus of future research and interventions; Mun and Geng (2019) identified variables predictive of post-experiment fatigue and suggested that the results highlighted a network of connections associated with health behaviors; and Plonsky et al. (2016) proposed a way to predictive human behaviours, and demonstrated that their method could also be used to identify important contributing factors. In all cases, the researchers used some combination of machine learning algorithms and model explainability techniques to infer something about the phenomena under study. Specifically, rather than simply establishing predictive validity, they used explainability techniques to read further into the underlying structure of the data. Furthermore, all of these studies used a combination of random forests and random forest importances, feature selection techniques, or Shapley value techniques to identify predictive or important variables.

Unfortunately, methods for explainability are only intended to be used to understand the algorithmic decisions themselves, and care must be taken when using model explanations to draw inference about external reality (Kumar et al., 2020; Lundberg, G. Erion, et al., 2020). Indeed, a variable is 'important' or 'predictive' insofar as it is useful to an algorithm for reducing prediction error. As we will demonstrate, whilst it is possible to infer that predictive/important variables are only predictive because the associated construct is linked in some way with the outcome (and therefore arguably theoretically relevant), the converse is not true. Variables which are deemed to *not* be predictive, may nonetheless contain just as much relevant information as the variables that are. Thus, it is at least not possible and even ill-advised to draw conclusions

about which variables are *not* important, because one may be inadvertently discounting the relevance of highly predictive variables. This comes as a consequence of the underling structure in the Data Generating Process (DGP), which is something that is unlikely to be known *a priori*. For instance, if a treatment T is fully mediated by variable M which have an effect on outcome Y, the combined inclusion of T and M as predictors will result in zero predictive importance being placed on T. These kinds of interactions are not limited to mediation structures, and depend broadly on the underlying structure governing the DGP.

Even though it is well known that similar consequences occur with partial correlation coefficients, whereby the shared variance between the outcome and the mediator renders the treatment to be ostensibly unimportant, the reasons behind this phenomenon, and the associated implications, are not generally well understood by practitioners, particularly from a causal perspective (Pearl, 2009; Rohrer, 2018; Cinelli, Forney, and Pearl, 2022; M. J. Vowels, 2021; VanderWeele, 2019). The fact that this phenomenon extends beyond linear regression to machine learning explainability techniques seems to be even less well appreciated, at least judging from the research in which they are used. Indeed, if explainability techniques cannot be reliably used to explore the data to uncover theoretically relevant predictive variables, we would encourage researchers to ask themselves whether algorithmic explainability measures of variable importance can reliably be used to inform and guide them in the ways that they expect. Consider the case where an explainability technique indicates that one particular set of variables $\mathbf{A}$ is important whilst another $\mathbf{B}$ is not. Would researchers be willing to make recommendations about the scope of further research if they also understand that set $\mathbf{B}$ may, in fact, be just as important as set $\mathbf{A}$? Even if the importances accurately reflect the use of information by the predictive algorithm, we should temper any conclusions relating to the nature of the psychological phenomenon itself. In particular, we should be reluctant to use the explainability results to influence the development of our theory. As such, it is important that researchers understand the behaviour of explainability techniques. To the best of our knowledge, the implications have yet to be discussed.

6.3 Background

As described above, the interaction between the underlying (usually unknown) causal structure and the output of machine learning explainability techniques is not generally well understood by practitioners, and has important implications for any subsequent interpretations. The Directed Acyclic Graph (DAG) framework and the associated d-separation rules make the consequences of structure visually intuitive, and recent commentaries have encouraged psychologists and social scientists to engage with these concepts (Hünermund and Bareinboim, 2021; Grosz, Rohrer, and Thoemmes, 2020; Rohrer, 2018; M. Hernan, 2018b). We believe that understanding these concepts is key to understanding the behaviour of machine learning explainability techniques, and we introduce them in this section. We also introduce some background concepts relating to regression and machine learning explainability techniques, in particular, the techniques we subsequently evaluate empirically.

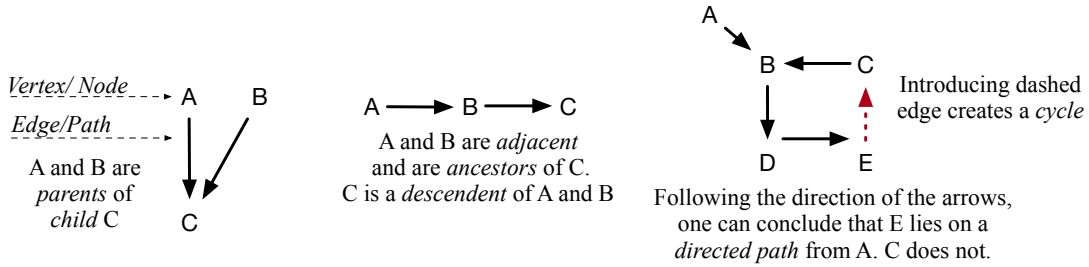
6.3.1 Structure

In order to explore the inability of machine learning models and explainability techniques to reliably inform us of correlations in the data, we take a causal perspective, which means we adopt the machinery developed for Structural Equation Modeling, probabilistic graphical models, and graphs. In particular, we use Directed Acyclic Graphs (DAGs) to operationalize the dependence of machine learning models on the underlying structure in a general, non-parametric way. In this section we introduce the relevant concepts for the subsequent exploration. Interested readers are encouraged to consult other resources on graphical models and causal inference such as Hünermund and Bareinboim (2021), M. Vowels, N. Camgoz, and Bowden (2022), J. Peters, Janzing, and Scholkopf (2017), and Koller and Friedman (2009).

The Link with Structural Equation Modeling

We follow a similar formalism to J. Peters, Janzing, and Scholkopf (2017) and E. Strobl (2018). These definitions are quite dense, and we therefore provide a figure illustrating them in Fig. 6.1.

Figure 6.1: Notational conventions.



Note. This figure aims to visually communicate some of the notational conventions defined in the text.

Many psychologists may already be familiar with Structural Equation Models (SEMs), which represent a popular subclass of Structural Causal Models (SCMs), whereby the SEMs are usually constrained to encode linear dependencies. SCMs enable an expression-based (rather than graphical) representation of structure. For example, the chain structure represented as the graph $X_i \rightarrow X_j \rightarrow X_k$ tells us that X_i causes X_k via a mediator X_j. The variables $X_{i,j,k}$ are graphically represented as *vertices* or *nodes* which are corrected via directed edges or paths which represent the flow of cause-effect. The chain structure can be specified as a system of equations in an SEM or SCM:

$$\begin{aligned} X_i &:= f_i(U_i), \\ X_j &:= f_j(X_i, U_j), \\ X_k &:= f_k(X_j, U_K). \end{aligned} \tag{6.1}$$

In the case of SEM, the functions f are linear. Furthermore, the use of the assignment operator ':=' makes explicit the asymmetric nature of these equations. In other words, they are not to be rearranged to solve for their inputs. Also of note are the $U_{\{i,j,k\}}$ terms, which represent unobserved exogenous noise variables which are usually omitted from the graphical representation. Including them would involve additional causes for each of the X nodes *e.g.*, $U_i \rightarrow X_i$.

Graphs

Working with the graphical portrayals of structural relationships provides an intuitive and immediately visually comprehensible representation. As such, we devote some time to defining the relevant notation and terminology. A directed graph $\mathcal{G}$ represents a joint distribution $\mathcal{P}$ as a factorization of d variables $\mathbf{X} = \{X_1, ..., X_d\}$ using d corresponding *nodes/vertices* and connecting, directed edges. If two variables or nodes X_i and X_j are directly connected by an edge we call them *adjacent* (think of them as being 'next to each other' in the graph), and, can also denote this in terms of the corresponding graph $\mathbf{X}$ as $X_i \rightarrow X_j$ or $X_i \leftarrow X_j$. If all edges are directed, and there are no cycles (see the right hand graph in Fig. 6.1 for an example of how a cycle might be induced), we have the class of *Directed Acyclic Graphs* (DAGs).

We can define a *parent* variable pa_j as one which has a *child* which is connected by a directed edge *e.g.* $X_i \rightarrow X_j$. Further upstream parents are *ancestors* of downstream *descendants* if there exists a directed path constituting $i_k \rightarrow j_{k+1}$ for all k in a sequence of vertices. An *immorality* or *v-structure* describes when two non-adjacent vertices are parents of a common child. A *collider* is a vertex where incoming directed arrows converge.

DAGs are assumed to fulfil the Markov property, such that the implied joint distribution factorizes according to the following recursive decomposition, characteristic of Bayesian networks (Pearl, 2009):

$$P(\mathbf{X}) = \prod_{i}^{d} P(X_i|pa_i). \tag{6.2}$$

Eq. 6.2 tells us that the joint probability of the system can be calculated as the product of the probabilities of each of the d variables, conditional on its parents. Taking the $\log$ of this expression makes the quantity computatable as a sum, instead of a product. By way of example, the likelihood of the graph on the left hand side of Fig. 6.1 can be computed as: $P(A, B, C) = P(A)P(B)P(C|A, B)$, and the log-likelihood can be computed as: $\log P(A, B, C) = \log P(A) + \log P(B) + \log P(C|A, B)$.

d-Separation Rules

The decomposition relates to the rules of d-separation and the implied conditional independencies implied by the graph. Two vertices X_i and X_k are d-separated, by the set of vertices $\mathbf{S}$ if $X_j \in \mathbf{S}$ in any of the following structural scenarios (J. Peters, Janzing, and Scholkopf, 2017):

$$\begin{aligned} X_i \to X_j \to X_k \quad &\text{(chain)} \\ X_i \leftarrow X_j \leftarrow X_k \quad &\text{(chain)} \\ X_i \leftarrow X_j \to X_k \quad &\text{(fork)} \end{aligned} \tag{6.3}$$

That is to say that there exists a 'flow' of statistical dependence between X_i and X_k *i.e.*, $X_i \not\perp\!\!\!\perp X_k$ for all three graphs above, but this flow is 'blocked' if we condition on the middle node X_j, in which case the two outer variables become statistically independent *i.e.*, $X_i \perp\!\!\!\perp X_k | X_j$. They are also d-separated if neither X_j *nor any of the descendants of* X_j are in set $\mathbf{S}$ in the following structural scenario:

$$X_i \to X_j \leftarrow X_k \quad \text{(collider)} \tag{6.4}$$

This means that if we condition on X_j, there is an induced statistical dependence between X_i and X_k which otherwise would not exist. Essentially, the 'flow' of statistical dependence is already 'blocked' by the collider structure, and conditioning on the collider itself unblocks it. As we will see, these rules have important implications for undertaking regression, which includes conditioning on a set of variables.

To transform these relationships from graphical/mathematical relationships to *causal* relations, the *Causal* Markov Condition is imposed, which simply assumes that the arrows represent causal dependencies and that there are no unobserved or unmodelled confounders (J. Peters, Janzing, and Scholkopf, 2017, p.105-6). It is then common to use the DAG framework as a means to represent domain knowledge relating to the underlying Data Generating Process

(DGP). The ultimate benefit of the graphical and structural model frameworks is that they, at least in principle and under some strong assumptions, enable us to use observational data to answer scientific questions such as 'how?', 'why?', and 'what if?' (Pearl and Mackenzie, 2018). If a domain expert has a theory about the structure underlying a given phenomenon, they may represent this theory graphically using a DAG.

In the presence of statistically dependent unobserved variables, the graph is said to be *semi*-Markovian, because some of the implied graphical conditional independencies may not hold in practice as a result of the additional dependencies induced by the unobserved variables. These unobserved variables are usually denoted with U, as in Eq. 6.1, but if the Markov condition holds they are usually omitted from the graph for convenience (their presence does not affect the entailing machinery of the graph). However, when the graph is semi-Markovian, it is actually more convenient to indicate these relationships graphically than in a system of equations, by including the dependence between U vertices graphically. For example, a curved, dashed, bidirectional link between the observed variables can be used to indicate the presence of an unobserved confounder U.

6.3.2 Regression

As psychologists wishing to undertake an analysis, we may be confronted with a dataset of N samples from random variables Y, an outcome of interest, and $\mathbf{X}$, a set of predictors. In terms of notation, we use **bold** to denote that we have multiple predictors, lower-case to denote a realisation y^i of upper-case random-variable Y. This notation is compatible with the notation used above for graphs, such that x_j^i is a specific datapoint i of variable X_j, which in turn may constitute a variable in a graph.

In the regression setting, we might be concerned with estimating the conditional expectation of Y using the set of variables X with regression model having parameters θ. Specifically, we wish to estimate $\mathbb{E}[Y|\mathbf{X}]$, *i.e.*, the expected value of Y given a particular set of values for the set of predictor variables. It is important to highlight the conditioning statement in this expectation $Y|\mathbf{X}$ (*i.e.*, Y conditioned on X). If Y were binary, the expectation would be equivalent to $P(Y|\mathbf{X} = \mathbf{x})$. This has important implications for *d*-separation as we described in the previous

section because conditioning on variables can induce independence or dependence (depending on the structure).

We can use a regression model (*e.g.*, a random forest) m_θ to approximate $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \approx m_\theta(\mathbf{x})$ with an empirical sample or subset of our dataset $\mathcal{D}$. Of course, with parametric assumptions we can fit a linear function m_θ via Ordinary Least Squares. However, we assume readers are already familiar with the construction of such as model (following the usual $Y = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K + \epsilon$ form where K is the number of variables in the set $\mathbf{X}$ and ϵ is exogenous noise, or model error). It is, however, useful to review some theory behind much more flexible models, because it is in light of their flexibility that the results in the experiment section may seem unintuitive. We therefore provide a brief review of both the random forest, and the MultiLayer Perceptron (MLP), below.

Random Forests

Random forests (Breiman, 2001a) are a type of adaptive decision tree. Decision trees partition the input space recursively according a set of thresholds. They determine a set of variables and thresholds with which to 'split' up the input and to logically derive a prediction of the outcome. The model can be trained to identify the best variable and thresholds to split on according to a variation of the squared error (K. P. Murphy, 2012, pp.548) $\sum_i^N (y_i - \bar{y})^2$ where here, $\bar{y}$ represents the average of the outcome variable for the associated set of data.

The popular extension of the decision tree, namely the random forest, reduces the variance of estimates from a decision tree by averaging together the predictions from many similar trees. By training J different trees using different subsamples of our dataset (subsamples of both datapoints and variables), we can derive a powerful algorithm for prediction. As such, the random forest 'carves' up the input space according to these splits, thereby deriving a highly non-linear mapping from the predictors to the outcome.

MultiLayer Perceptrons (MLPs)

The MLP (Haykin, 1999) is a type of neural network, comprising multiple linear layers and non-linear activations. Each layer serves not only as a non-linear transformation of the input, but also as a means to change (expand or shrink) the dimensionality of the set of predictors. In other words, if we start with *e.g.* five predictors, we can take 20 different functions (where these functions are learned by the algorithm) of each of these predictors to achieve a 100-dimensional expansion of these predictors. We take a set of predictors and recursively process them according to a set of these non-linear layers, until they can be mapped to the desired outcome. By stacking any number of these layers (we use two layers in the experiments), one after the other, and by having the inner layers operating at a much higher dimensionality than the dimensionality of the set of predictors we started with (we use a dimensionality of 100 in the experiments). MLPs thus facilitate the learning of highly complex, non-linear functions, and have been shown to be 'universal function approximators' (Hornik, Stinchcombe, and White, 1989), which, loosely, means that they can approximate any function.

The MLP is generally trained according to gradient descent, where the parameters are updated according to their impact on a specified loss function (*e.g.* mean squared error). Using relatively elementary calculus (chain rule) one can compute the extent to which each weight or bias parameter and update them according to a learning rate. The principal consideration that arises is that these algorithms may not converge to the global optimum, and may even 'get stuck' at arbitrary local optima. Interested readers are directed to the accessible overview of deep learning by I. Goodfellow, Bengio, and Courville (2016) and additional information in the Supplementary material for this Chapter.

6.3.3 Model Explainability

Random Forest Importances for Model Explainability

One of the most common ways to derive feature importances for random forests is via the use of impurity measures. In this work, we consider continuous outcome variables (*i.e.*, we use

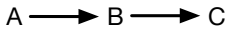
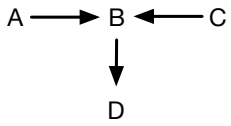
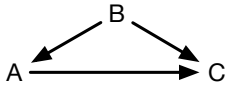
regressors, rather than classifiers), and in this case the measure of impurity is usually the mean squared error (others are possible, *e.g.*, the mean absolute error). For each of the decision trees in the random forest, the importances can be calculated based on the decrease in impurity (*i.e.*, the improvement in performance) for each node, weighted by the probability of using that node in a particular tree, and these improvements can be averaged across data samples (Breiman, 2001a; C. Strobl et al., 2007).

In the empirical literature, it is common to use these importances as proxies for variable importance *outside* the model. For example Joel, Eastwick, Allison, et al. (2020) used machine learning importance measures to infer the presence of associations between variables pertinent to relationship quality in the real-world. The logic seems to be that if a random forest finds a variable useful in making a prediction, then there exists some (potentially non-linear) association in the real world. Indeed, given the bootstrapped nature of random forests, and the way that they can carve up the input space to define highly non-linear mappings between predictors and outcome, we can understand this thought process. Unfortunately, in spite of their flexibility, they are nonetheless constrained according to the structure in the data, and variables which appear unimportant, may actually be important and highly statistically associated.

Shapley Values for Model Explainability

The Shapley value explainability methods derive from the seminal game theoretic work of Lloyd Shapley (Shapley, 1953). The methods conceive of a regression task as a collaborative game, where each of the predictor variables represents a player. The goal of the game is to maximise the regression performance (or, equivalently, to minimize the regression error), and the explanation quantifies the degree to which each player (*i.e.*, predictor) contributes to this goal. Of course, the role that each predictor plays is difficult to directly ascertain, because it is 'collaborating' with other predictors at the same time (in the form of multiplicative interactions $X_1 \cdot X_2$, for example) in specific and complex ways which are largely determined by the ML algorithm itself. The Shapley value methods therefore disentangle these complex contributions by evaluating the impact that each possible combination of predictors has on the model output. The result is a per-predictor, per-datapoint estimation of the impact on model performance, thus

Figure 6.2: Example d-separation implications.

Graph:	Regression:	Implications:
	$\mathbb{E}[C\|A,B]$	$A \perp\!\!\!\perp C\|B$
	$\mathbb{E}[C\|A]$	$A \perp\!\!\!\perp C$
	$\mathbb{E}[C\|A,D]$	$A \not\perp\!\!\!\perp C\|D$
	$\mathbb{E}[C\|A,B]$	$A \perp\!\!\!\perp B$

Note. In the first graph (a chain structure), all the variables are statistically dependent. However, if we are interested in predicting C from A and B, we 'block' the path from A to C when we condition on B, inducing independence. This would result in A being unimportant for the regression $\mathbb{E}[C|A,B]$. In the middle graph, A is already independent of C because of the collider structure at B. The regression $\mathbb{E}[C|A]$ does nothing to affect this independence structure. However, by conditioning on D, we are conditioning on a descendent of a collider, which renders A important for the regression $\mathbb{E}[C|A,D]$. Finally, whilst the regression $\mathbb{E}[C|A,B]$ renders A independent of C via B (which is useful for 'blocking the non-causal backdoor path'), the conditions do not prevent B or A from being important, as predictors, to the regressor.

providing a fine-grained summary of model behaviour. Interested readers are directed to the recent papers by Lundberg and colleagues (Lundberg and S.-I. Lee, 2017; Lundberg, G.G. Erion, and S.-I. Lee, 2017; Lundberg, G. Erion, et al., 2020). It is this level of fine-grained information which gives Shapley values a distinct advantage over random forest importances. Furthermore, they have also been shown to be more reliable, in general, than random forest importances (Lundberg, G. Erion, et al., 2020), and can be derived for almost arbitrary model classes. Indeed, in the experiment section we will use the Shapley value techniques to derive estimates of variable importances from neural networks. Further details on explainability methods and SHAP in particular can be found in the supplementary material for this chapter.

6.3.4 Regression and Structure - Possible Explanations

In order to glean an understanding for why the results reported in later sections are seemingly so structurally dependent requires an understanding of the conditional independencies implied by the underlying graph. Before we present the results, we therefore devote some time to discuss a couple of simple examples and, in addition, Fig. 6.2 provides some other instances for

consideration. First, let us look more closely at the important implications of the *d*-separation rules.

Consider the graphs given by:

$$\begin{aligned} X_1 &\xrightarrow{\beta_1} Y \xleftarrow{\beta_2} X_2 \\ X_1 &\xrightarrow{\beta_1} X_2 \xrightarrow{\beta_2} Y \end{aligned} \tag{6.5}$$

As with our later experiments, let us assume that the dependencies are linear, to keep things simple. We have indicated with the path coefficients β the true (population level) strengths of the dependencies between variables in these graphs. Let us also assume that the true values of these path coefficients are all equal to one, *i.e.*, $\beta_1 = \beta_2 = 1$ for both graphs.

In the context of regression, we may wish to estimate $\mathbb{E}[Y|\mathbf{X}]$. This expectation is, itself, dependent on a model of the conditional distribution of $Y|\mathbf{X}$. For both graphs above, our regression implies a conditional density $Y|X_1, X_2$, and this has different implications for each of the two graphs above, and the implications can be understood via the *d*-separation rules.

For the first graph in Eq. 6.5, the regression entailing the conditional density for $Y|X_1, X_2$ does nothing to interfere with the conditional independencies encoded by the original graph. Namely, $X_1 \perp\!\!\!\perp X_2$ regardless of whether we are conditioning on the X variables as part of the regression or not. Indeed, if we were to undertake a linear regression, or fit the data using a random forest, we would expect the associated coefficients/importances to be approximately equal, reflecting the fact that the true dependencies between these variables ($\beta_1 = \beta_2 = 1$) are also equal. In this case, the coefficients from a linear regression will be unbiased estimates of the true path coefficients, *i.e.*, $\hat{\beta}_1 \approx \beta_1$, and $\hat{\beta}_2 \approx \beta_2$.

On the other hand, in the second graph, the conditioning in our regression results in $X_1 \perp\!\!\!\perp Y|X_2$. In words, even though there is a clear dependency structure between X_1 and Y, by conditioning on X_2 this structure is 'broken', and X_1 and Y are rendered independent. Imagine the true dependencies are linear and that $\beta_1 = \beta_2 = 1$. This means the effect of both X_1 and X_2 on Y is also one (according to the multiplication of the path coefficients). Instead, the coefficients of

a linear regression $Y|X_1, X_2$ will estimate $\hat{\beta}_1 \approx 0 \not\approx (\beta_1\beta_2)$, which is misleading (although a natural consequence of the underlying structure), and conversely will correctly estimate $\hat{\beta}_2 \approx 1 \approx \beta_2$.

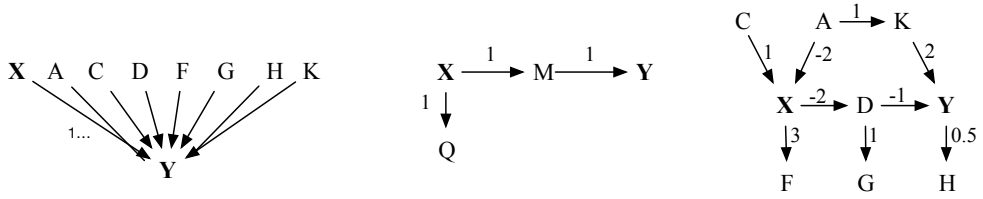
6.4 Methodology

Having reviewed some relevant background material, we take a moment to discuss the motivation for the following experiments.[2] On the face of it, given the nature of random forests and neural networks, there is little to concern us that a random forest would be necessarily prevented from leveraging any or all useful correlations between the predictors $\mathbf{X}$ and the outcome Y. Indeed, it seems like the opposite might be more likely: The fact that a random forest can arbitrarily partition the input space according to multiple bootstrapped decision trees, where the details of the partitioning are driven by a very general cost function (such as the squared error) perhaps encourages us to think that the algorithm can do whatever it wants to leverage any and all statistical associations in the data. Similarly, the fact that the neural network is a universal function approximator, and can expand the dimensionality of the predictors arbitrarily might lead us to believe that it has relatively free-reign or equal opportunity to use any and all useful variables. In turn, then, we might also expect explainability techniques (such as random forest importance measures, or Shapley values) to yield predictor importance levels which are relatively agnostic to the structure of the Data Generating Process (DGP) which led to the observations with which the models were trained.

However, and as we will see in the results section, in spite of the flexibility of random forests and neural networks, the methods are nonetheless sensitive to the interaction between the conditioning statements in the associated regression being undertaken, and the underlying structure of the DGP. As a result, the use of variable importance measures (including Shapley value techniques) does not help us to reliably identify predictively useful variables.

[2] Note that full code for the experiments is provided at `https://github.com/matthewvowels1/ML_structural_interactions`.

Figure 6.3: The causal structures of the two datasets used in the experiments.



Note. On the left, the graph is trivial - all predictors are independent of each other and cause the outcome Y. The path coefficients are all one. In the centre, the effect of X on Y is mediated by M, and Q represents a descendent of X. In a slight abuse on notation, in both graphs variables X and Y are highlighted in **bold**. On the right, the structure of the second dataset is more complex. This to emphasize that in practice, we might be particularly interested in the influence a particular predictor has on the outcome, when that predictor is just one in a system of many. In the left graph, the influence of X on Y is the same as all the others (and equal to one), whereas in the right graph, the influence of X on Y is equal to $-2 \times -1 = 2$ according to the multiplication of the path coefficients for the mediated path $X \to D \to Y$.

6.4.1 Data

We create three datasets, one with four variables (three predictors and one outcome variable), and two with nine variables (eight predictors and one outcome variable). The structures of these datasets are shown in Fig. 6.3. All datasets are generated according to linear functional relationships, and the corresponding path coefficients are denoted in the Figure. In a slight abuse of notation, variables X and Y are highlighted in **bold** in these graphs, because we assume them to be variables of interest for the sake of the experiments. For instance, X might refer to some kind of 'risk' variable, which we expect, as domain experts, to affect outcome Y. We use a sample size $N = 10,000$ to avoid estimation variability due to sample size.

It can be seen that the first dataset (Fig. 6.3, left) has a trivial, exogenous error structure, with independent predictors. The second dataset has a mediation structure (Fig. 6.3, centre) and has a system of equations (SCM) given by:

$$
\begin{aligned}
X \sim \mathcal{N}(0,1), \quad & U_M \sim \mathcal{N}(0,1), \\
M = X + U_M, \quad & U_Q \sim \mathcal{N}(0,1), \\
Q = X + U_Q, \quad & U_Y \sim \mathcal{N}(0,1), \\
Y = M + U_Y, . \quad &
\end{aligned}
\tag{6.6}
$$

Finally, the the third dataset (Fig. 6.3, right) is based on one from Peters et al. (J. Peters, Janzing, and Scholkopf, 2017) and Vowels (M. J. Vowels, 2021), and is more complex, containing direct effects, *e.g.*, $D \rightarrow Y$; mediated effects, *e.g.*, $X \rightarrow D \rightarrow Y$; 'backdoor' paths (Pearl, 2009), *e.g.*, $X \leftarrow A \rightarrow K \rightarrow Y$, where X is linked to Y via an indirect, non-causal path; and colliders, *e.g.*, $C \rightarrow X \leftarrow A$.

The system of equations (the SCM) representing this second dataset is:

$$
\begin{aligned}
C \sim \mathcal{N}(0,1), \quad & A \sim \mathcal{N}(0,0.8), \\
U_K \sim \mathcal{N}(0,0.1), \quad & K = A + U_K, \\
U_X \sim \mathcal{N}(0,0.2), \quad & X = C - 2A + U_X, \\
U_F \sim \mathcal{N}(0,0.8), \quad & F = 3X + U_F, \\
U_D \sim \mathcal{N}(0,0.5), \quad & D = -2X + U_D, \\
U_G \sim \mathcal{N}(0,0.5), \quad & G = D + U_G, \\
U_Y \sim \mathcal{N}(0,0.2), \quad & Y = 2K - D + U_Y, \\
U_H \sim \mathcal{N}(0,0.1), \quad & H = 0.5Y + U_H.
\end{aligned}
\tag{6.7}
$$

Here, $\sim \mathcal{N}(\mu, \sigma)$ denotes that observations for these variables are samples from a normal distribution with mean μ and standard deviation σ. Despite the increased complexity of this graph, in our opinion it is not so complex to be implausibly representative of real-world causal structures. The dataset is split 60/40 into train and test proportions. Given that previous work has highlighted the sensitivity of random forest importance measures to the variance of the

data, we standardized all data before use (C. Strobl et al., 2007). This also makes comparison between different explainability measures more comparable, particularly as the absolute values of the bivariate correlations are, by their definition, constrained to fall between 0 and 1. The bivariate correlations are shown in Table 6.1. It can be seen that all variables are highly (and statistically significantly) correlated with the outcome. This is intentional and provides a best case scenario for the explainability techniques - they are all important variables, and we wish to understand whether machine learning methods can help us identify them.

Table 6.1: Bivariate Pearson correlations and p-values, $R(p)$, for the right-hand DAG in Figure 6.3.

$r(p)$	X	D	A	K	C	F	G	H
Y	.92(.00)	-.94(.00)	-.60(.00)	-.59(.00)	.76(.00)	.91(.00)	-.93(.00)	1.00(.00)

6.4.2 Models / Algorithms

We provide results for bivariate correlations, Linear Regression (**LR**), Random Forest (**RF**), and MultiLayer Perceptron (**NN** - for Neural Network). It is generally known that the default parameters of random forests perform well across a range of applications, without the need for hyperparameter tuning (Probst, M. Wright, and Boulesteix, 2018), and as such, we use the default settings in the scikit-learn package (Pedregosa et al., 2011). Similarly, rather than undertaking an exhaustive hyperparameter search for the MLP, we stay close to the default parameters and verify that the test performance is comparable to that of the random forest. The dimensionality of the layers is set to be 100, the number of layers set to 2 (with one additional outcome layer), the activation is chosen to be ReLU, we use the Adam (Kingma and Ba, 2017) optimizer with an adaptive learning rate starting at 1×10^{-3}, and trained for 200 iterations. All implementations were written in Python 3.7.

6.4.3 Explainability Techniques

We provide the bivariate correlations between each of the predictors and the outcome, denoted '**bi-corrs**' in the results. For the linear regression we simply provide the coefficient values as measures of predictor importances, these are denoted '**LR-coefs**' in the results. Indeed, linear regression is straightforward to interpret in this regard. For the random forest we provide both

importances derived according to the built-in node impurity method in scikit-learn - denoted '**RF-imps**' in the results - as well as Shapley values using the SHAP (SHapley Additive exPlanations) 'Tree Explainer' package (Lundberg, G. Erion, et al., 2020) - denoted '**RF-Shap**' in the results. For the MLP we use the SHAP 'Kernel Explainer' package, denoted '**NN-Shap**' in the results. For both the Tree Explainer and the Kernel Explainer we use a train and test size of 1000 datapoints.

6.4.4 Trials and Results Presentation

In order to demonstrate the interaction between the structure of the second (structurally mode complex) dataset and the models, we undertake a number of analyses, each time removing different variables to understand the concomitant impact on the explanations. In each case, we provide a bar plot showing the relative importance of each variable for each method. In order to make the linear regression coefficients, the Shapley values, and random forest importances more visually comparable, we normalize them to have a range of zero to one (the bivariate correlations are left untouched). The Shapley results are derived to be the absolute values of the per-datapoint impact on model output, averaged over the datapoints. Finally, we provide mean squared errors for each of the algorithms / models.

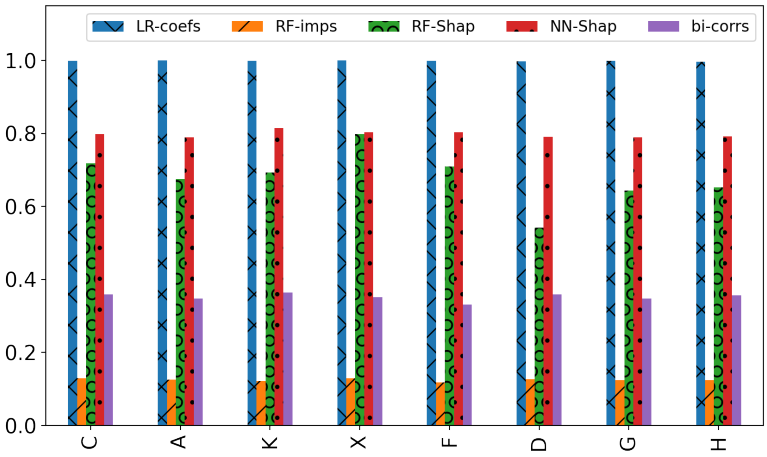
6.5 Results

In practice, we are unlikely to have access to the true graph structures as provided in Fig. 6.3. We may also be interested in the relationship (associational/predictive or causal) between two variables in particular, and we assume these to be X and Y, which are highlighted in bold in the graphs. In order to evaluate the sensitivity of explainability measures to the underlying structure, we can use a dataset for which we know *a priori* that there is a strong association between X and Y. For the left graph, the causal effect is equal to that of all other variables (one), for the right graph the effect is comparable to the other variables, and equal to $-2 \times -1 = 2$ according to the multiplication of the coefficients on the mediated path $X \rightarrow D \rightarrow Y$. We therefore also know that these variables should ideally be denoted to be of importance by the explainability

techniques. Indeed, if the explainability techniques cannot highlight the presence of a strong association (such as that between X and Y), we might easily discount otherwise key variables as being unrelated/unimportant.

Let us begin by checking that ML algorithms and explainability techniques are not fooled by trivial/idealistic structures. In Fig. 6.4 we show the results for the simple structure depicted on the left of Fig. 6.3. We know from the construction of this dataset that all variables have equal importance, because the causal effect of each variable on the outcome is equal to one. An evaluation of the importances in Fig. 6.4 proves to be reassuring because, indeed, regardless of which method we choose, the importances are rated as equal. even though there are differences in the absolute levels of importance *between* methods, one can nonetheless see that these importances are approximately equal for all variables. To this extent, we have confirmed our expectations that when the structure is trivial (all variables independently causing the outcome), machine learning algorithms can be used to highlight variables of particular importance. There is nothing in the conditional independency structure skewing our assessment of variable importance.

Figure 6.4: Results for trivial DAG structure.

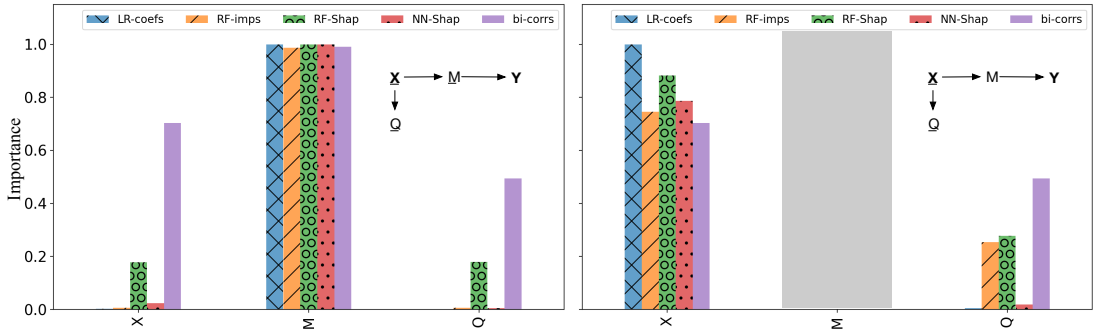


Note. Results for linear regression coefficients 'LR-coefs', random forest importances 'RF-imps', random forest Tree Explainer Shapley values 'RF-Shap', neural network Kernel Explainer Shapley values 'NN-Shap', and bivariate correlations 'bi-corrs' for the left-hand graph of Fig. 6.3. It can be seen that the importances derived from each method are equal across variables. Figure best viewed electronically and in colour.

The results for the second dataset with the mediation structure are shown in Fig. 6.5. For convenience, these plots feature the DAG in the plot whitespace, with underlined variables

highlight variables included as predictors in the models. On the left, we include X, M and Q as predictors of outcome Y. The results show that all explainability techniques place importance on the mediator M. In contrast, X is deemed to be unimportant with negligible 'LR-coefs', 'NN-Shap', and 'RF-imps'. Interestingly, some importance was assigned to X by the 'RF-Shap'. Results for Q were very similar to those for X. On the right, we exclude M from the set of predictors, thereby d-connecting X and Y. The consequence is that all the importance is assigned to X, thereby confirming that the inclusion of the mediator 'blocks' the path from X to Y. This simple example with only three predictors already illsutrates the sensitivity of the the importances to the underlying structure in the data.

Figure 6.5: Results for mediation DAG structure.



Note. Results for linear regression coefficients 'LR-coefs', random forest importances 'RF-imps', random forest Tree Explainer Shapley values 'RF-Shap', neural network Kernel Explainer Shapley values 'NN-Shap', and bivariate correlations 'bi-corrs' for the central graph of Fig. 6.3. It can be seen that the importances depend on the inclusion of M. Figure best viewed electronically and in colour.

The results for the third dataset with the more complex structure are shown in Fig. 6.6(i-iv), and these deserve a longer discussion. Once again, for convenience, these plots feature the DAG in the plot whitespace, with underlined variables highlight variables included as predictors in the models. Starting with plot (i), which includes all predictor variables, we see dramatic changes in the relative levels of importance between variables and across methods. For instance, the linear regression coefficients 'LR-coefs' on variables K and H are high, followed by D, and then all other coefficients are approximately zero. This particular result is easy to explain given knowledge of the true graph - the only paths which have not been blocked by other control variables in the linear model are $H \leftarrow Y$, $D \rightarrow Y$, and $K \rightarrow Y$. At least the linear regression is consistent in this regard, but remember that without this knowledge we would not be able to

use the coefficients to infer which variables are important.

Perhaps the next most reasonable set of importances are given by the neural network Shapley values 'NN-Shap'. Here, the top three most important variables are, as with the linear regression, K, D, and H. However, K's importance is doubtful. Both the random forest's importances 'RF-imps' and Shapley values 'RF-Shap' are very misleading - the only variable of note is H, with all others having very low importance.

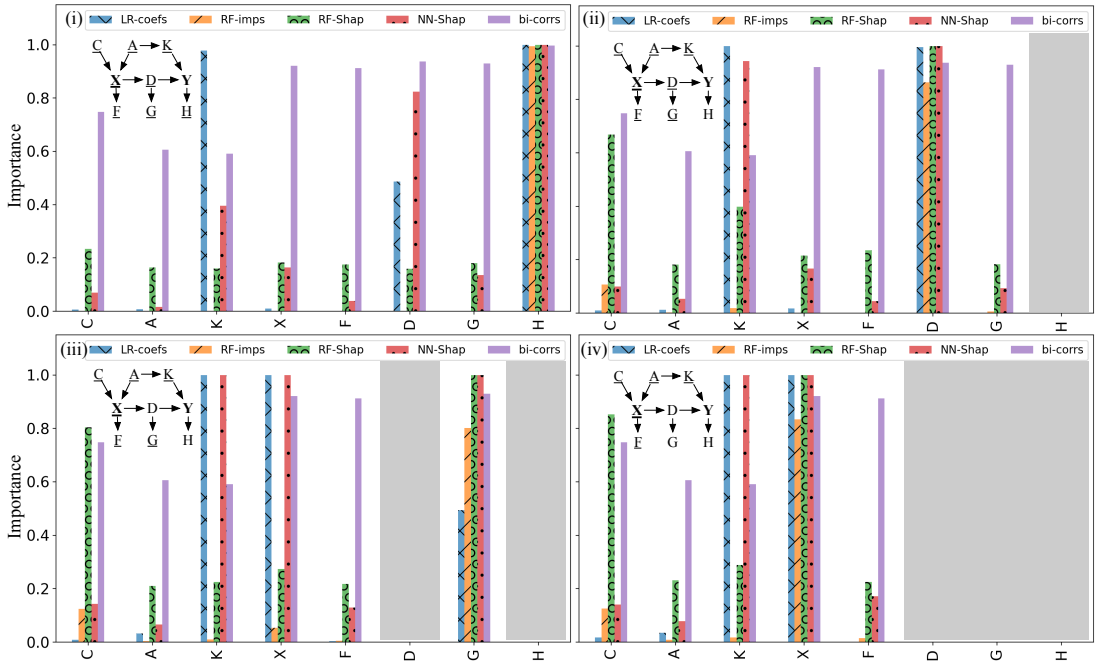
From the first plot alone, we see very strong interaction between the structure and the machine learning explainability results. If we were interested in understanding whether variable X is relevant to Y (which we know it certainly is, because unlike in practice, we have ground truth and simulated the data ourselves) we would have discounted it as unimportant. We could stop there - we have demonstrated that machine learning algorithms, despite their flexibility, are not able to overcoming the constraints deriving from the conditional independencies implied by the underlying graph. However, it is of interest to understand how these importances *change*, as variables are removed. In plot (ii) we remove H from the set of predictor variables. Linear regression again provides predictable results - the unblocked paths to the outcome are significant predictors, namely $K \rightarrow Y$ and $D \rightarrow Y$. It would still not be possible to reliably interpret these results without knowledge of the true graph. However, they are, at least, consistent with the graph, to the extent that we know there exist variables which 'block' the flow of statistical dependence. Again, the neural network provides results which are reasonably consistent with the linear regression, with K and D being highlighted as the most important. Unfortunately, the random forest importances and Shapley values are completely unpredictable: This time, variables C and D are most important.

In plot (iii) we have removed variables H and D. Variable X is now (finally) deemed to be an important variable by the linear regression and the NN. However, this is despite the fact that the descendent of the mediator G is still in the model. This may bias the estimate somewhat, but it is not enough to block the path completely between X and Y. As such, the fact that the linear regression coefficients and NN Shapley values indicate importance for X is still reasonable. They also indicate that G is important, which is also reasonable given the open path $G \leftarrow D \rightarrow Y$ without D included as a predictor. Once again, the random forest Shapley value

results are somewhat unexplainable, with C being denoted to be the most important variable, followed by G. The random forest importances only indicate that G is important.

The final plot (iv) removes H, D, and G. Now we expect unbiased estimates of the causal effect of X on Y by the linear regression, and this is indicated also by the high values for the coefficients on X. As before, K is also deemed important, as expected. The NN yields similar results, whereas, once again, the random forest importances and Shapley values are somewhat unpredictable.

Figure 6.6: Results for non-trivial DAG structure.



Note. Result plots (i-iv) for linear regression coefficients 'LR-coefs', random forest importances 'RF-imps', random forest Tree Explainer Shapley values 'RF-Shap', neural network Kernel Explainer Shapley values 'NN-Shap', and bivariate correlations 'bi-corrs' for the right-hand DAG in Fig. 6.3. The DAGs featured in the whitespace of each plot denote which variables are included as predictors according to whether the variable is underlined. It can be seen that the methods different widely in the relative importances of each variable depending on on the set of predictors as well as the algorithm used. Figure best viewed electronically and in colour.

6.6 Discussion

Linear regression and neural networks were severely influenced by the underlying structure, but the associated importances (linear regression coefficients and Shapley values, respectively) were at least consistent with our expectations given knowledge of the true graph. In this sense, the explainability techniques worked as expected. In contrast, random forests were also severely affected by the underlying structure but in ways which were quite unpredictable even given knowledge of this structure. Based on these experiments, we could recommend neural networks over random forests for identifying important predictors. However, this recommendation is somewhat moot, because even though neural networks were affected by the underlying structure in ways which were predictable, they nonetheless would not be useful in guiding research without prior knowledge of said structure. Indeed, in practice none of these methods would yield interpretable results without knowledge of the underlying graph, and so are not recommended as part of an initial exploration for the development of a theory.

The results confirmed that variables which are deemed to be unimportant or unpredictive of the outcome by the explainability techniques, may actually be highly predictive to the extent that the exclusion of a single variable can completely shift the spread of results. Estimations of predictive importance that are meaningless to the extent that variables which are deemed to be unimportant may yet still be important are not helpful to researchers. One might argue that we can use theory to strongly inform which variables are included, and whether some may be mediators. However, one of the original motivations for such an approach is to explore the data and to help discover structure in the data (Yarkoni and Westfall, 2017) *when it is not already known*. In recent work, researchers used large numbers of variables with these techniques - Joel, Eastwick, and E.J. Finkel (2017) explored datasets with upwards of 100 variables - and we suspect it would be most unlikely for a researcher to be able to confidently account for the underlying structure of such a dataset, at least in the domain of psychology.

By consequence, rather than recommending that psychologists utilize *predictive approaches* to 'help gain a deeper understanding of the general structure of one's data' (Yarkoni and Westfall, 2017) (which can, as we have shown, greatly mislead us as to the relevance of certain variables),

we provide two recommendations for researchers who are, perhaps, at the early, exploratory stages of a research project, and who are seeking a means to identify important variables. Namely, we recommend mutual information as a means to identify statistical dependence between variables, without the need for assumptions about the functional form, and without needing to constrain the analysis to parametric distributions. M.I. is a measure of how much information one variable contains about another (Cover and Thomas, 2006; Kraskov, Stogbauer, and Grassberger, 2004; G. V. Steeg and Galstyan, 2012; G. Steeg and Galstyan, 2013; Gao, G. Steeg, and Galstyan, 2015; Kinney and Atwal, 2014; M. J. Vowels, 2021).

Secondly, we recommend researchers engage with techniques from the domains of causal discovery, in order to provide a means to highlight variables which have statistical relevance and to contextualise such variables within an initial estimate of the causal structure. For overviews of causal discovery methods, interested readers are directed to M. Vowels, N. Camgoz, and Bowden (2022), Heinze-Deml, Maathuis, and Meinshausen (2018), C. Glymour, K. Zhang, and Spirtes (2019), and Spirtes and K. Zhang (2016). As with any data-driven approach, particularly those which implicate causality, we, like Dawid and his reference to Bourdieu, warn researchers to beware of ”sliding from the model of reality to the reality of the model” (Dawid, 2008; Bourdieu, 1977). Descriptions of both mutual information and causal discovery can be found in the supplementary material to this chapter.

We would also like to emphasize that rather than practitioners being generally discouraged from using machine learning techniques as a consequence of this work, we instead highlight the potential for machine learning techniques to mitigate the need for unreasonable assumptions about the functional form. Indeed, in some regards it is reassuring that machine learning algorithms, in spite of their ‘black-box’ reputation, are nonetheless constrained according to the rules of regression and the conditional independency structure of the data. Furthermore, if we are able to integrate machine learning algorithms into analyses which adequately account for the underlying structure, we can benefit from the power of the machine learning algorithms without the associated problems demonstrated in this work (M. J. van der Laan and Starmans, 2014; Kennedy, 2020; Yoon, J. Jordan, and van der Schaar, 2018; M. J. Vowels, N. Camgoz, and Bowden, 2021; Wu and Fukumizu, 2022; W. Zhang, L. Liu, and J. Li, 2021).

6.7 Conclusion

The use of machine learning with explainability was motivated by a need to explore data, particularly when our existing theories are still in the development stage, and/or when we wish to understand their predictive validity (M. J. Vowels, 2021; Yarkoni and Westfall, 2017). The idea is essentially that the identification of predictive variables can help guide our theory development process, and this idea is already guiding current research in psychology (Joel, Eastwick, and E.J. Finkel, 2017; L. M. Vowels, M. J. Vowels, and K.P. Mark, 2021a). Unfortunately, in this work we have shown that flexible, powerful machine learning algorithms are not agnostic to the underlying conditional independency structure of the DGP which yielded the observations, and that concomitant estimations of variable importance are arbitrarily skewed by the choice of algorithm as well as the underlying (unknown) causal structure in the data.

We emphasise two points: Firstly, results depended heavily on the underlying structure in the DGP, and on which variables are included as predictors in the model. It is important to note that, in the absence of any dependency structure, *i.e.*, there exist direct causal paths between all variables and the outcome, as in the graph on the left hand side of Figure 6.3, importance measures *can* be used as a proxy for association. Assuming the algorithm is behaving without bias towards any particular variable, the associated importances should (assuming stable performance) be proportional to the amount of shared information between each predictor and the outcome. In contrast, in datasets containing variables between which there exist important structural relationships - that is, both correlational/predictive and causal - machine learning techniques can 'miss' key predictive variables, at least insofar as the explanations deem them to be unimportant. As we explained, this is consequence of the interaction between otherwise flexible machine learning algorithms, the task of regression conditional on some set of covariates, and the underlying structure in the data.

Secondly, a distinction must be clearly made between what explainability techniques can tell us about what the algorithm is doing (these explanations thereby remaining local to the model), and what they can (not) tell us about the presence or absence of correlations or associations in the real-world. As we have seen from the empirical evaluations, the nature and convergence of the

algorithm itself may be somewhat unpredictable, even if the explainability techniques function as a reliable means to identify variables which are important *for the algorithm* in predicting the outcome.

In summary, we would question the utility of measures of predictive importance and explainability techniques to psychologists wishing to explore the data to guide their research. Indeed, how useful is it for the development of a theory to know that variable X is useful for predicting variable Y in arbitrary algorithm f, if the estimation of usefulness is specifically tied to the algorithm and to choice of other predictors? The bottom line is that one cannot 'outrun causality in machine learning', and that despite of the powerful function approximation capabilities of machine learning algorithms, they cannot be used to reliably explore the data for theoretically relevant predictive and/or causal variables.

6.8 Supplementary Material

6.8.1 MultiLayer Perceptrons

The Multilayer Perceptron (MLP) is a generalization of the classic perceptron, developed by Rosenblatt (1958), and inspired by a simple model for a neuron in the human brain. Top-level diagrams for the perceptron and the MLP are shown in Figure 6.7. The perceptron can be viewed as a form of generalised linear model, where a weighted sum of the inputs and a bias offset is passed to a non-linear function (such as the sigmoid function). In fact, if the sigmoid is used as the activation function, the calculation is equivalent to that for a prediction from a logistic regression model parameterised by weights W and bias term b.

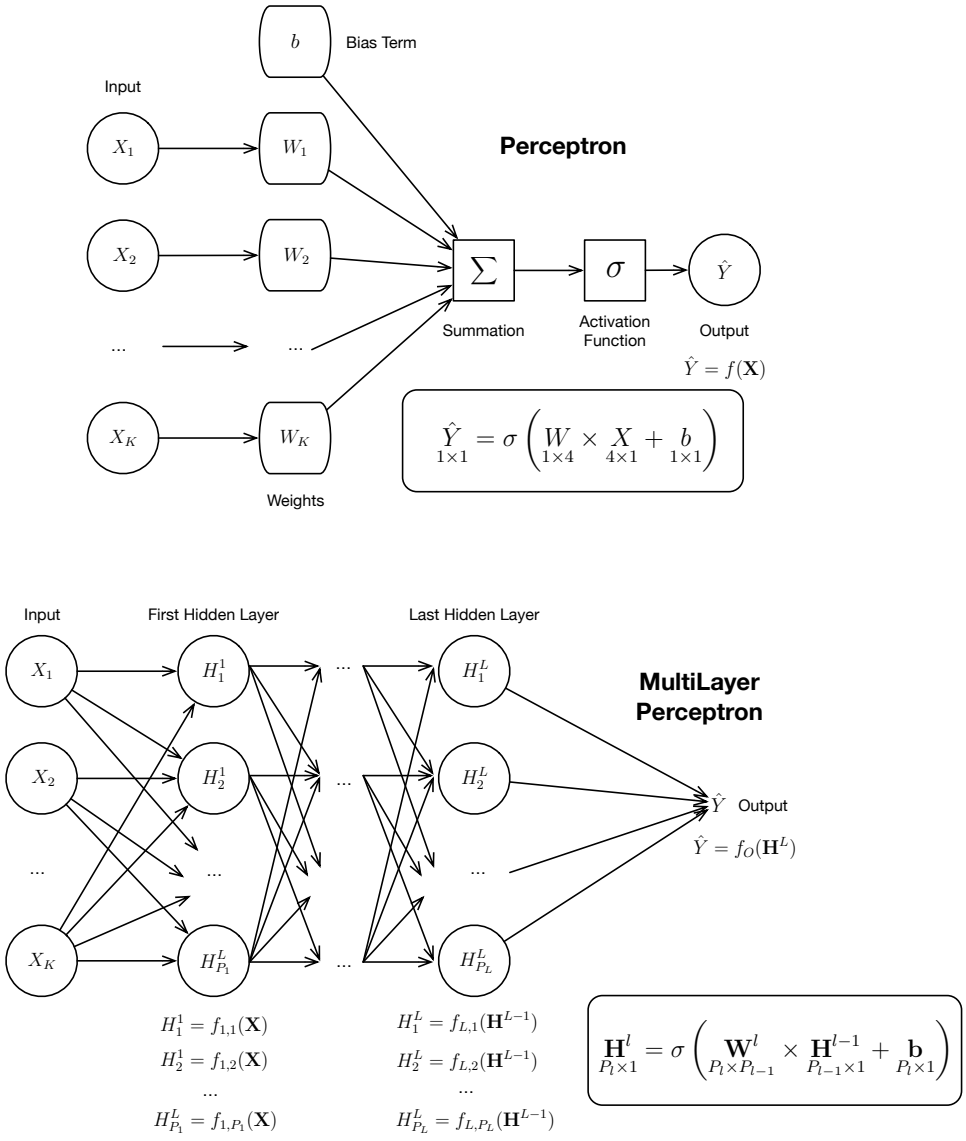
If this logistic regression model is stacked both 'horizontally' and 'vertically', we arrive at an MLP. In other words, instead of performing just one prediction from a logistic regression model we perform P_1 of them, where P_1 is the number of *neurons* in the first *hidden layer* of the MLP, we form the first vector of neuron values $\mathbf{H}^1$. Then, in turn, for each of the values in the vector of $\mathbf{H}^1$ we perform P_2 logistic regression prediction calculations, we arrive at the vector of neuron values $\mathbf{H}^2$. This process can be repeated L times, where L is the number of hidden

layers in the MLP, until we arrive at the last hidden layer, where we perform, one final logistic regression prediction. As such, the MLP is a set of recursive linear operations with a non-linear activation function between each subsequent operation. The number of neurons at each layer determines its *width*, which can vary across the layers, and the number of layers determines its *depth*. The computational interconnectedness thereby resembles a *network*, and these models are also forms of artificial neural network. Once the depth increases beyond, say, 30-40 layers, they tend to be called *deep* neural networks.

When the MLP is initialized, the values for the parameters (the weights and biases) are usually randomly sampled from a Gaussian with a mean of zero and a variance scaled according to the dimensions of the MLP architecture (I. Goodfellow, Bengio, and Courville, 2016; Glorot and Bengio, 2010). As such, the MLP begins by making nonsense predictions. In much the same way as a logistic regression needs to be 'fit' to the data, the MLP also requires a fitting or training process. The most common way to train these networks is via a form of *gradient descent*. The process of gradient descent involves deriving (usually automatically via a symbolic differentiation process) the derivatives of a target *objective function* or *loss function* which characterizes the predictive performance of the network, with respect to the corresponding parameters. The objective function can be, for example, the mean squared error in predictions from the network, and the derivatives thereby characterize how much the performance of the network changes with respect to a change in the parameters. The goal is then to find the parameters which maximise the performance in terms of the objective function. In the case of the mean squared error objective function, a network which performs well is a network which *minimizes* the mean squared error of its predictions with respect to the ground truth.

To briefly explain the process of gradient descent, we consider how it might be undertaken firstly in the case of the simple perceptron. Assuming the task of the network is regression, and that we wish the minimize the squared error, our objective function for a single prediction from the perceptron can be defined as $L = (\hat{Y} - Y)^2$. Here, $\hat{Y}$ is a prediction from the perceptron, and Y is the ground truth for that particular prediction. In turn, assuming the identity function as our activation (the process generalises to any differentiable activation function), the output for the network can be expressed as $\hat{Y} = WX + b$. The symbolic differentiation process then

Figure 6.7: Top-level diagrams for the perceptron and the multilayer perceptron.



Note. In the diagrams for both the perceptron and the MLPm X is a single input sample from a K-dimensional dataset, σ is a non-linear activation function (such as the sigmoid),$\hat{Y}$ is a scalar outcome prediction (*e.g.* for a regression task). In the perceptron model, b is a scalar bias term, and the weights W are a K-dimensional vector of scalar valued weights. The diagram for the MLP is a generalisation of the one for the perceptron. For each column or 'layer' of circles or 'neurons' after the input layer, each neuron represents a scalar value which is the outcome of a weighted linear combination with a bias added and an activation function, equivalent to the computation for $\hat{Y}$ in the perceptron. The MLP therefore represents a set of stacked perceptrons. l indexes the hidden layer number, and p_l indexes the number of neurons in hidden layer l (maximum P_l).

finds the partial derivatives of this objective function with respect to the parameters:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial W} \tag{6.8}$$

and

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial b} \tag{6.9}$$

Thanks to the chain rule, we are thus able to recursively decompose the derivative of the objective function with respect to the parameters one layer at a time. Each of the individual terms represents straightforward calculus. For example:

$$\frac{\partial L}{\partial \hat{Y}} = -2Y + 2\hat{Y} = 2(\hat{Y} - Y) \tag{6.10}$$

Once we have a symbolic representation for $\partial L / \partial \theta$ where θ comprises the weights and bias parameters, we can evaluate this derivative using the data and our predictions. The value of this derivative provides us with a proxy for how far our parameters are from being able to make perfect predictions. This is because, if the value of $\partial L / \partial \theta^* = 0$, we assume we have found the basin of lowest error (*i.e.*, at this particular θ^*, our objective function is at a minimum). This minimum is searched for following a simple iterative update rule:

$$\theta_{t+1} = \theta_t - \alpha \left(\frac{\partial L}{\partial \theta_t} \right) \tag{6.11}$$

In words, the next (and hopefully improved) values of the parameters, is equal to the previous values of the parameters plus some amount α of the derivative of the loss function with respect to the parameters, evaluated using the values in the dataset. There exist many hundreds of variations of this process (*e.g.*, ADAM Kingma and Ba, 2017, RMSProp Tieleman and Hinton, 2012), but they are all based on the principles of gradient descent.

The MLP can be generalised to yield multidimensional outputs, perform regression tasks, classification tasks, and hybrids between the two. The flexibility of the architectural design is partly

why these models have seen such wide success over such a diverse range of tasks, including computer vision and natural language processing (I. Goodfellow, Bengio, and Courville, 2016).

6.8.2 Explainability Methods

Whilst simple models, such as multiple linear regression models, have straightforward and simple explanations, others, such as random forests (Breiman, 2001a) and neural networks (I. Goodfellow, Bengio, and Courville, 2016) are much more complex and defy trivial explanations. There exist a range of options for explaining such complex models, including (1) gain based approaches, which include the traditional Gini and impurity based approaches in random forests, (2) split count based approaches, which evaluate how many times a particular feature or variable is used to split a decision process, and (3) permutation based approaches, which permute values of a feature of variable and assess the corresponding impact on the model (Lundberg and S.-I. Lee, 2017). Unfortunately, the first two methods do not yield consistent results, which means that "a model can change such that it relies more on a given feature, yet the importance estimate assigned to that feature decreases." (Lundberg and S.-I. Lee, 2017). Furthermore, none of these methods (including the permutation based methods which may otherwise be consistent) provide individualized results.

In contrast to these other (inconsistent and/or non-individualized) methods, SHAP is a unified framework for explanation which is both consistent and provides explanations for individual predictions from the model. It derives from the seminal game theoretic work of Lloyd Shapley (Shapley, 1953). The framework conceives of predictors as collaborating agents seeking to maximize a common goal (i.e., the regressor performance). The approach involves systematically evaluating changes in model performance in response to including or restricting the influence from different combinations of predictors. It is an additive method, which means that the explanation model g it provides is a linear function of binary variables z':

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \quad \text{(Lundberg and S.-I. Lee, 2017)} \tag{6.12}$$

Here, z' are binary indicators which represent where a variable was observed or unknown, and

ϕ are the variables' attribution values (*i.e.*, the degree to which the model attributes weight to the associated variable/feature). SHAP is designed such that the sum of the feature attributes ϕ is equal to the value of the statistical model's output ('statistical model' is what I call the model f that we wish to explain). Lundberg and S.-I. Lee (2017) explain that in order that the method can handle missing variables or features (and thereby attribute importance to features when they are not included), they need to define a mapping h_x which maps between the pattern of missingness determined by z' (where $z'_i = 0$ when the associated feature is missing), and the statistical model's input space. This mapping can then be used to evaluate $f(h_x(z'))$, where f is the statistical model, which enables us to calculate the consequence of including or not including the associated feature.

Then, the authors define S as a subset of the features which are present (*i.e.*, for these features, $z' = 1$), as well as $f_x(S) = f(h_x(z') = \mathbb{E}[f(x)|\mathbb{E}[f(x)|x_S]$. As such, $\mathbb{E}[f(x)|x_S]$ is the expected value of the statistical model $f(x)$, knowing/conditional on the subset of the inputs S. For N as the set all input variables/features, and following the attribution process of Shapley's game theoretic work, the attribution for input variable i is defined by Lundberg and S.-I. Lee (2017) as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S) \right] \tag{6.13}$$

Authors demonstrate that this approach is, at least theoretically, the only possible approach which fulfils an important set of requirements for explanation, including consistency and individualized explanations. Indeed, the SHAP TreeExplainer function from the SHAP software implementation provides estimations of the per-datapoint, per-predictor impact on model output, as well as the average predictor impacts for tree-based methods like random forests.

6.8.3 Mutual Information

To discuss mutual information, it is necessary to first define entropy (in terms of which mutual information is usually expressed). Entropy describes the degree of surprise, uncertainty, or information associated with a distribution and is computed as $H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i)$

where N is the number of datapoints in the sample distribution, x_i is a single datapoint in this distribution, and $p(x_i)$ is that datapoint's corresponding probability. If the entropy is low, it means the distribution is more certain and therefore also easier to predict. Taking a fair coin as an example, $p(x = \text{heads}) = p(x = \text{tails}) = 0.5$. The entropy of this distribution is $H(X) = -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1$. The units of entropy here are in bits. The fact that our fair coin has 1 bit of information should therefore seem quite reasonable - there are two equally possible outcomes and therefore one bit's worth of information. Furthermore, because the coin is unbiased, we are unable to predict the outcome any better than by randomly guessing. On the other hand, let's say we have a highly biased coin whereby $p(x = \text{heads}) = 0.99$ and $p(x = \text{tails}) = 0.01$. In this case $H(X) = -(0.99 \log_2 0.99 + 0.01 \log_2 0.01) = 0.08$. The second example had a much lower entropy because we are likely to observe heads, and this makes samples from the distribution more predictable. As such, there is less new or surprising information associated with samples from this distribution, than there was for the case where there was an equal chance of a head or a tail.

Mutual information is an information theoretic measure of the degree to which information associated with one variable X (or set of variables) is shared by another variable Y (or set of variables). It is a more general form of correlation which is a measure of statistical association that assumes linear forms of dependence. It can be expressed in terms of entropy as follows:

$$I(X;Y) = H(X) - H(X|Y), \quad \text{(Cover and Thomas, 2006)} \tag{6.14}$$

As the entropy is a measure of uncertainty, mutual information tells us to what extent knowledge of Y reduces the uncertainty in X. Using the definition of conditional entropy, mutual information can be expressed as:

$$I(X;Y) = \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \tag{6.15}$$

In this formulation, one can observe that when the two variables X and Y are completely independent of each other, the joint distribution $P(X,Y)$ will be equal to the product of their

marginals $P(X)P(Y)$, which results in the latter term $\log((P(X,Y)/P(X)P(Y)) = 0$, and, by consequence, mutual information $I(X;Y) = 0$.

Whilst these definitions concern discrete variables, there exist extensions to continuous variables. One must then decide whether parametric models suffice for the representations of the probability density functions. Of course, if parametric models are sufficient, there exist closed-form analytical representations of the definitions above. If not, then one must perform density estimation to approximate these functions.

6.8.4 Causal Discovery

Causal discovery is the process of exploiting properties of the joint distribution to identify causal/structural links in the data. These properties include conditional independencies (also known as constraint-based methods), distributional asymmetries, score-based, intervention based, and shadow-manifold based methods (M. Vowels, N. Camgoz, and Bowden, 2022). For the purposes of this explanation we consider the first two, but interested readers are directed to M. Vowels, N. Camgoz, and Bowden (2022) for more information.

Conditional Independencies: Consider the graph:

$$X \rightarrow Y \leftarrow Z.$$

This graph implies a set of conditional independencies, specifically:

$$X \perp\!\!\!\perp Z|\emptyset$$

$$X \not\perp\!\!\!\perp Y|\emptyset,$$

$$Z \not\perp\!\!\!\perp Y|\emptyset,$$

$$X \not\perp\!\!\!\perp Z|Y.$$

It is possible to test for these, in the linear case, by measuring partial correlations between them.

For instance, straightforward multiple linear regressions can be used for each of the conditional independencies above as follows:

$$\mathbb{E}[Z|X],$$

$$\mathbb{E}[Y|X],$$

$$\mathbb{E}[Y|Z],$$

$$\mathbb{E}[X|Z,Y].$$

Assuming these multiple regressions yield coefficients, one can use these coefficients to establish the existence or non-existence of a statistical dependence between the relevant variables. This approach can then be generalised the non-linear, non-parametric circumstances using, for example, mutual information (described above).

The challenge with this approach is that there exist a class of different structures called the Markov Equivalence Class, which all underpin joint distributions with the same conditional independencies. For instance, there is nothing in the conditional independencies which distinguish the following graphs:

$$X \rightarrow Y \rightarrow Z$$

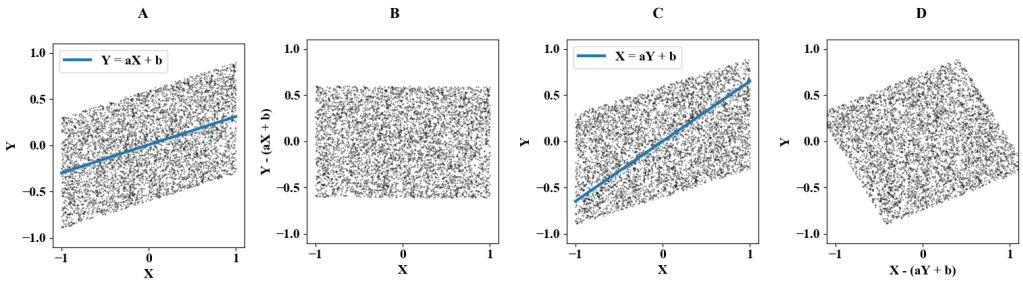
$$X \leftarrow Y \rightarrow Z$$

$$X \leftarrow Y \leftarrow Z$$

As such, collider structures (such as the one in the first example in this description, above) are helpful in orienting edges in larger structures.

Distributional Asymmetries: Figure 6.8 illustrates the possibility to recover the causal direction between two variables under the assumption of noise additivity. Regression Y on X yields a different association between residuals and covariates than the inverse (regressing X on Y). This concept generalises to a broader class of models, and may be detectable using machine

Figure 6.8: Illustration of how distributional asymmetries can be used to identify pairwise causal directionality.



Note. The true structural relationship is $Y = X + U_Y$ and $X = U_X$ where U_X and U_Y are uniform noise sources. (A) shows the regression line when regressing Y onto X, and (B) shows the corresponding residuals plotted against X. (C) shows the regression line when regressing X onto Y and (D) shows the corresponding residuals plotted against Y. Together, these demonstrate that under the assumption of linear functional form and non-Gaussian noise, the true structural direction is identifiable as the one for which X is independent of the residuals, as indicated in (B). Example adapted from J. Peters, Janzing, and Scholkopf, 2017 and M. Vowels, N. Camgoz, and Bowden, 2022.

learning methods, as in work by J. Mooij et al. (2016). The fact that this is (sometimes) possible comes as a consequence of the independence of mechanisms assumption in causal processes:

"Assuming that the true structural direction is $X \to Y$, the concept of independent mechanisms holds that $P(X)$ contains no information about $P(Y|X)$, and vice versa. A common illustrative example J. Peters, Janzing, and Scholkopf, 2017 involves measurements of temperature Y at weather stations of different altitudes X. Regardless of the distribution of weather station altitudes $P(X)$, the mechanisms linking altitude to temperature (*e.g.* the law determining the relationship between the temperature and pressure of a gas) exist independently, and changing the temperature around a weather station does not increase its altitude." (M. Vowels, N. Camgoz, and Bowden, 2022)

CHAPTER 7

Typical Yet Unlikely: Using Information Theoretic Approaches to Re-Characterize Normality

The content of this chapter is drawn from the following publication:

Vowels, M.J., Under Review, Typical Yet Unlikely: Using Information Theoretic Approaches to Identify Outliers which Lie Close to the Mean.

Abstract: Normality, in the colloquial sense, has historically been considered an aspirational trait, synonymous with harmony and ideality. The arithmetic average has often been used to characterize normality, and is often used as a blunt way to characterize samples and outliers. Prior commentaries in the fields of psychology and social science have highlighted the need for caution when reducing complex phenomena to a single mean value. However, to the best of our knowledge, none have described and explained why the mean provides such a poor characterization of normality, particularly in the context of multi-dimensionality and outlier detection. We demonstrate that even for datasets with a relatively low number of dimensions, data start to exhibit a number of peculiarities which become progressively severe as the number of dimensions increases. We show that normality can be better characterized with 'typicality', an information theoretic concept relating to entropy. An application of typicality to both synthetic and real-world data reveals that in multi-dimensional space, to be normal (or close to the mean)

is actually to be highly atypical. This motivates us to consider one example application of outlier detection, and we demonstrate typicality for outlier detection as a viable method which is consistent with this updated definition. In contrast, whilst the popular Mahalanobis based outlier detection method can be used to identify points far from the mean, it fails to identify those which are too close. Typicality can be used to achieve both, and performs well regardless of the dimensionality of the problem.

7.1 Introduction

In a well known United States Air Force (USAF) experiment seeking to identify 'the average man', Gilbert Daniels found that out of 4,063 men, not a single one fell within 30% of the arithmetic sample averages for each of ten physical dimensions simultaneously (which included attributes such as stature, sleeve length, thigh circumference, and so on) (Daniels, 1952). Rather than this being an unlikely fluke of the sample, averages, rather than representing the most 'normal' attributes in a sample, are actually highly abnormal in the context of multi-dimensional data. Indeed, even though averages may provide seemingly useful baselines for comparison, it is important and perhaps surprising to note that the chance of finding an individual with multiple traits falling close to the average is vanishingly small, particularly as the number of traits increases.

The arithmetic average has been used to represent normality (vis-a-vis abnormality, in the informal/colloquial/non-statistical sense), and is often used both productively and unproductively as a blunt way to characterize samples and outliers. Prior commentary has highlighted the pitfalls associated with the use of the mean as a summary statistic (Speelman and McGann, 2013); the limitations in relation to its applicability and usefulness of parametric representations (such as the Gaussian) when dealing with real-world phenomena (Micceri, 1989; Modis, 2007); and the societal context (Misztal, 2002; Comte, 1976), surrounding the potentially harmful perception of normality as a "figure of perfection to which we may progress" (Hacking, 1990, p. 168). Whilst these commentaries are valuable and important in developing an awareness for what it means to use averages to characterize humankind, they do not provide us with an alternative. They also do not discuss some of the more technical aspects of normality in the context of

multiple dimensions and outlier detection, or explain *why* normality, when characterized by the arithmetic average, is so difficult to attain in principle.[1]

In this paper, we discuss averages in the context of multi-dimensional data, and explain why it is that being normal is so abnormal. We touch on some of the peculiarities of multi-dimensional spaces, such as how a high-dimensional sphere has close to zero volume, and how high-dimensional random variables cluster in a thin annulus a finite distance away from the mean. Whilst not the primary focus of this work, we also consider the relevance of these phenomena to outlier detection, and suggest that outliers should not only be considered to include datapoints which lie far from the mean, but also those points close to the mean.

Using information theoretic concepts, we propose an alternative way of characterizing normality and detecting outliers, namely through the concept of 'typicality'. We demonstrate the peculiarities as well as the proposed concepts both on idealistic simulated data, as well as data from the 'Politics and Views' LISS panel survey (Scherpenzeel and Das, 2010).[2]

Finally, we compare the outlier detection performance of typicality with the most common alternative (based on the Mahalanobis distance) and demonstrate it to be a viable alternative. More broadly, we argue that if the average value in a multivariate setting is unlikely, then outlier detection techniques should be able to identify it as such. This means updating our working conceptualization of outliers to include not only points which lie far from the mean (as most outlier detection methods do) but also those points which lie too close to the mean, particularly as the dimensionality of the dataset increases.

7.2 Background

The notion of the mean of a Gaussian, or indeed its finite-sample estimate in the form of the arithmetic average, as representing a 'normal person' still holds strong relevance in society and research today. Quetelet did much to popularise the idea (Quetelet, 1835; Caponi, 2013), having devised the much used but also much criticized Body Mass Index for characterizing a person's

[1] One exception includes work by Kroc and Astivia (2021) for the determination of scale cutoffs.

[2] An anonmyized github repository with the Python code used for all analyses, simulations, and plots can be found at `https://anonymous.4open.science/r/Typicality-F87B`

weight in relation to their height. The average is used as a way to parameterize, aggregate, and compare distributions, as well as to establish bounds for purposes of defining outliers and pathology vis-à-vis 'normality' in individuals. Quetelet's perspective was also shared by Comte, who considered normality to be synonymous with harmony and perfection (Misztal, 2002). Even though it is important to recognize the societal and moral implications of such views, this paper is concerned with the technical aspects of multivariate distributions; in particular, those aspects which help us understand why averages can be such a poor characterization of 'normality', and what we should use as an alternative, particularly for the identification of outliers in data.

In the past, researchers and commentators (including well-known figures such as Foucault) have levied a number of critiques at the use of averages in psychology (Speelman and McGann, 2013; Myers, 2013; Foucault, 1984; Wetherall, 1996). Part of the problem is the over-imposition of the Gaussian distribution on empirical data. The Gaussian has only two parameters, and even if the full probability density function is given, only two pieces of information are required to specify it - the mean (which we treat as equivalent to the arithmetic average) and the variance. Even in univariate cases, the mean can be reductionist, draining the data of nuance and complexity. Many of the developments in statistical methodology have sought to increase the expressivity of statistical models and analyses in order to account for the inherent complexity in psychological phenomena. For example, the family of longitudinal daily diary methods (Bolger and Laurenceau, 2013), as well as hierarchical models (Raudenbush and Bryk, 2002) can be used to capture different levels of variability associated with the data generating process. Alternatively, other methods have sought to leverage techniques from the engineering sciences, such as spectral analysis, in order to model dynamic fluctuations and shared synchrony between partners over time (M. J. Vowels, K. Mark, et al., 2018; Gottman, 1979). Machine learning methods provide powerful, data-adaptive function approximation methods for 'letting the data speak' (M. J. van der Laan and S. Rose, 2011) as well as for testing the predictive validity of psychological theories (M. J. Vowels, 2021; Yarkoni and Westfall, 2017), and in the world of big data, comprehensive meta-analyses allow us to paint complete pictures of the gardens of forking paths (Gelman and Loken, 2013; Orben and Przybylski, 2019).

Multi-dimensional data exhibit a number of peculiar attributes which concern the use of averages.

Assuming one conceives of a 'normal person' as having qualities similar to those of a 'typical person', we find that the arithmetic average diverges from this conception rather quickly, as the number of dimensions increases. The peculiar attributes start to become apparent in surprisingly low-dimensional contexts (as few as four variables), and become increasingly extreme as dimensionality increases. Understanding these attributes is particularly important because the dimensionality of datasets and analyses is increasing along with the popularity of machine learning. For instance, a machine learning approach identifying important predictors of relationship satisfaction incorporated upwards of 189 variables (Joel, Eastwick, Allison, et al., 2020), and similar research looking at sexual desire used around 100 (L. M. Vowels, M. J. Vowels, and K.P. Mark, 2021b; Joel, Eastwick, and E.J. Finkel, 2017). Assuming that high-dimensional datasets will continue to be of interest to psychologists, researchers ought to be aware of some of the less intuitive but notable characteristics of such data.

As we will discuss, one domain for which the mean can be especially problematic in multiple-dimensional datasets is outlier detection. In general, outlier detection methods concern themselves with the distance that points lie from the mean. Even methods designed to explore distances from the median are motivated by considerations/difficulties with estimation, and are otherwise based on the assumption that the expected value (or the estimate thereof) provides an object against which to compare datapoints (Leys, Delacre, et al., 2019). Unfortunately, and as Daniel's discovered for the USAF, values close to the mean become increasingly unlikely as the number of dimensions increases, making the mean an inappropriate reference for classifying outliers. As we describe later, one *can* successfully summarise a set of datapoints in multiple dimensions in terms of their *typicality*. We later evaluate the performance of a well-known multivariate outlier method (based on the Mahalanobis distance) in terms of its capacity to identify values far from the empirical average as outliers, and compare it against our proposed measure of typicality.

7.3 Divergence from the Mean

This section is concerned with demonstrating some of the un-intuitive aspects of data in higher dimensions. We begin by showing that, as dimensionality increases, the 'distance' that a

datapoint is from the mean/average increases at a rate of $\sqrt{D}$ where D is the number of dimensions. We then provide a discussion of the ramifications. Finally. we briefly present an alternative geometric view that leads us to the same conclusions.

Notation: In terms of notation, we denote a datapoint for an individual i as $\mathbf{x}_i$ where $i = \{1, 2, ..., N\}$. The total number of individual datapoints is N, and the bold font indicates that the datapoint is a vector (*i.e.* it is multivariate). A single dimension d from individual i's datapoint is given as x_{id}, where $d \in \mathbb{Z}^{+D}$, where D is the total number of dimensions, and where we use the subscript i or d according to the relevant context.

7.3.1 Gaussian Vectors in High Dimensions

Let us begin in familiar territory - for a multivariate distribution with independently and identically distributed (*i.i.d*) Gaussian variables, the probability density function for each dimension may be expressed as:

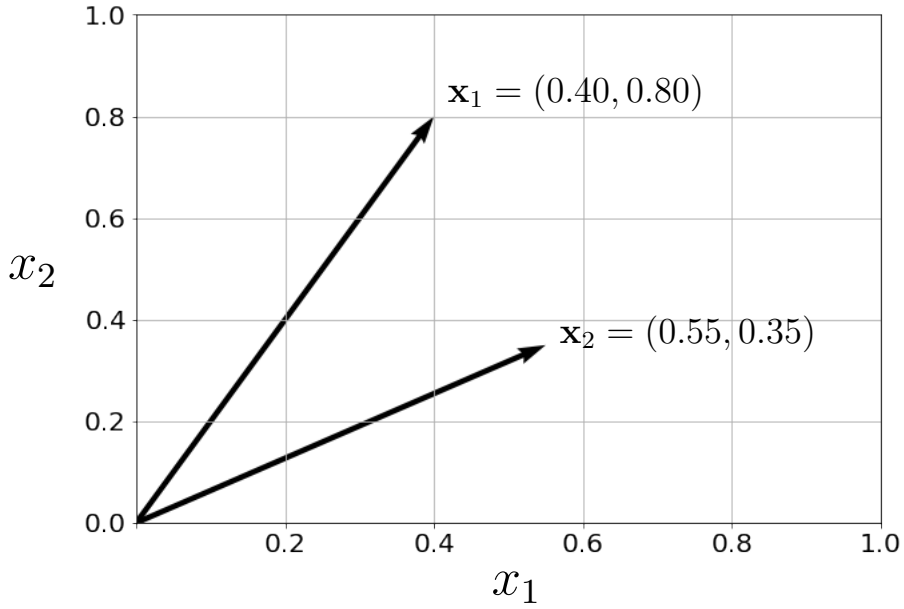
$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma_d^2}} e^{-\frac{(x_i - \mu_d)^2}{2\sigma_d^2}} \tag{7.1}$$

Each multivariate datapoint $\mathbf{x}$ may be considered as a vector in this D-dimensional space. An example of two datapoints drawn from a two-dimensional/bivariate version of this distribution (*i.e.*, $D = 2$), is shown in Figure 7.1. In this figure, the values of these two random samples are $\mathbf{x}_1 = (0.4, 0.8)$ and $\mathbf{x}_2 = (0.55, 0.35)$. Assuming that these datapoints are drawn from a distribution with a mean of 0 and a variance of 1 for all dimensions (*i.e.*, $\mathcal{N}(\mu_d = 0, \sigma_d^2 = 1) \ \forall \ d$), then we can compute the distance these datapoints fall from the mean $\boldsymbol{\mu} = \mathbf{0}$ using the squared Euclidean distance (see Eq. 7.2),

$$||\mathbf{x}||_2^2 = \sum_d^D x_d^2. \tag{7.2}$$

Here, we use the subscript d to index the dimension of the multidimensional datapoint $\mathbf{x}$. For the two example vectors in Figure 7.1, taking the square root of the values derived using Eq. 7.2,

Figure 7.1: Two-Dimensional Vector Space



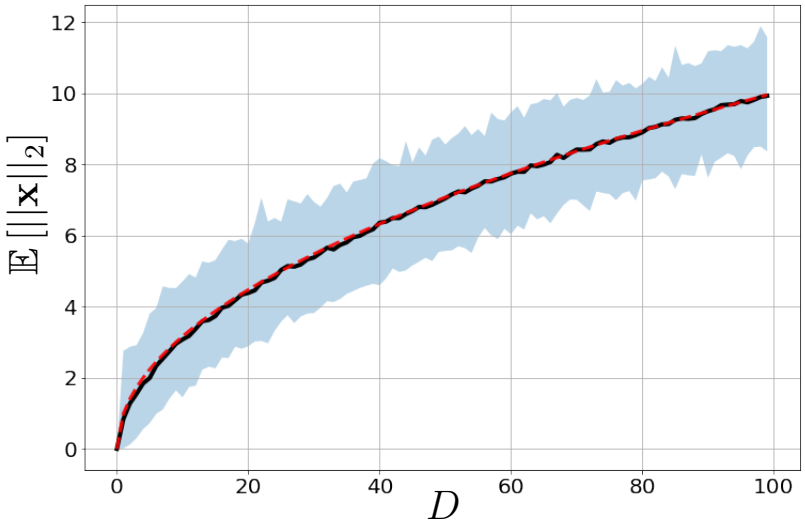
Note. Two samples in two-dimensional space, with their corresponding coordinates.

the distances are $||\mathbf{x}_1||_2 = 0.8$ and $||\mathbf{x}_2||_2 = 0.3$. Importantly, note that the *squared* Euclidean norm closely resembles the expression for sample variance (Eq. 7.3):

$$\overline{\text{Var}}(x_d) = \frac{1}{N}\sum_{i=1}^{N} x_i^2. \tag{7.3}$$

In other words, the variance of a sample is closely related to the distance that each sample is expected to fall from the mean. Note that, when computing the variance we sum across datapoints i, rather than dimensions d. Secondly, and more importantly, the variance contains a normalization term N^{-1}, whereas the expression for the norm does not. Consequently, the expected squared distance of each datapoint from the mean will grow with increasing dimensionality. In this example, we know that the variance of our distribution $\sigma_d^2 = 1$ for both dimensions, and as such, it is trivial to show that each individual dimension d will have an expected length equal to one. Without the normalization term (*i.e.*, D^{-1}), this means that the expected squared length of the vectors grows in proportion to the number of dimensions. Alternatively, taking a square root, we can say that the expected length of the vectors increases proportional to the square-root of the dimensionality of the distribution. More concretely:

Figure 7.2: Expected distances of vectors from the mean in high dimensions.



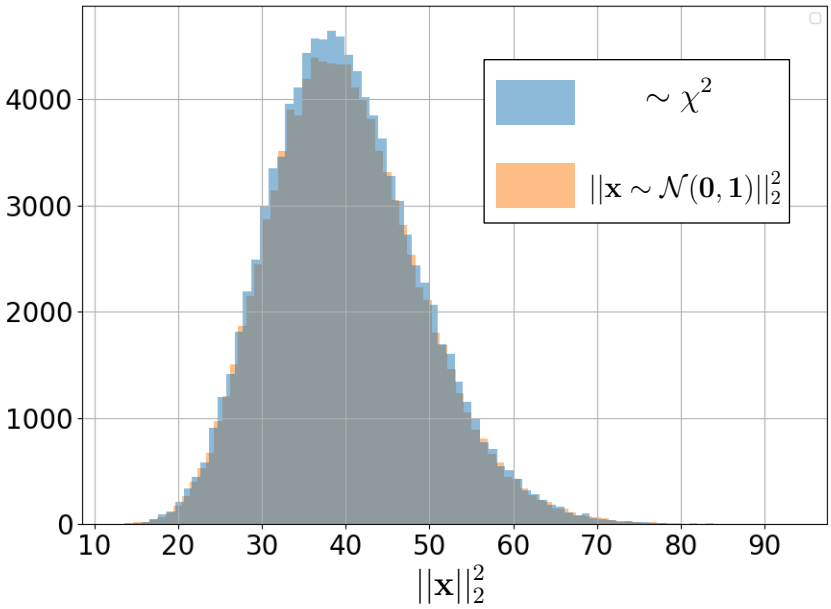
Note. The red dashed curve is simple $\sqrt{D}$, whilst the black curve is a simulated estimate of the expected lengths, calculated over 200 datapoints, for increasing dimensionality D. The blue interval represents the 1-99% percentiles.

$\mathbb{E}\left[||\mathbf{x}||_2\right] \propto \sqrt{D}$ (Vershynin, 2019).

This can of course also be verified in simulation, and Figure 7.2 shows both the analytical as well as sample estimates for the average length of the vectors as the number of dimensions increases. The intervals are defined by the 1st and 99th percentiles. Each approximation to the expectation is taken over a sample size of 200 datapoints. The dashed red curve depicts the $\sqrt{D}$ relationship, and the black simulated curve is a direct (albeit noisier, owing to the fact that this curve is simulated) overlay. This should start to remind us of Daniel's experience when working for the USAF - he found that out of 4,063 people, not a single one of them fell within 30% of the mean over ten variables. Indeed, if any had done, we should consider labelling them as outliers, in spite of the fact that most existing outlier detection methods are only sensitive to points which lie *far* from the mean.

The implications of this are important to understand. Whilst we know that each variable x_d has an expected value of zero and a variance of one, the expected length of a whole datapoint (all dimensions) grows in proportion to the square root of the number of variables. Dieleman (2020) summarised this informally when they observed that "if we sample lots of vectors from a

Figure 7.3: Histograms of χ^2 and sums of squares



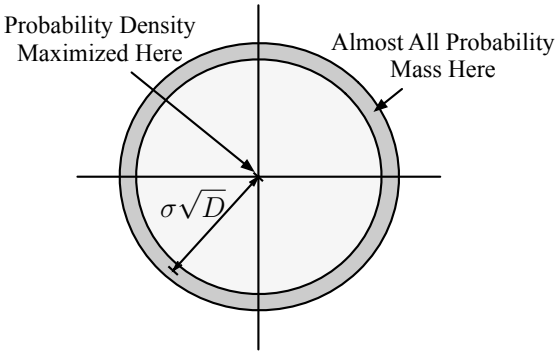
Note. For $D = 40$ these histograms show the distributions of 10,000 datapoints sampled from a χ^2 distribution (red) and the sums of squared distances $||\mathbf{x}||_2^2$.

100-dimensional standard Gaussian, and measure their radii, we will find that just over 84% of them are between 9 and 11, and more than 99% are between 8 and 12. Only about 0.2% have a radius smaller than 8!" In other words, the expected location of a datapoint in D-dimensional space moves further and further away from the mean $\boldsymbol{\mu} = \mathbf{0}$ as the dimensionality increases.

It can also be shown that such high-dimensional Gaussian random variables are distributed uniformly on the (high-dimensional) sphere with a radius of $\sqrt{D}$, and grouped in a thin annulus (Stein, 2020; Vershynin, 2019).[3] The uniformity tells us the direction of these vectors (*i.e.*, their location on the surface of this high-dimensional sphere) is arbitrary, and the squared distances, or radii, are Chi-squared distributed (it is well known that the Chi-squared distribution is the distribution of the sum of squares of D independent and identically distributed Gaussian variables). The distribution of distances (vis-à-vis the squared distances) is therefore Chi-distributed. Figure 7.3 compares samples from a Chi-squared distribution against the distribution of 10,000 squared vector lengths. Altogether, this means that in high-dimensions, (a) it is

[3] See also Gaussian Annulus Theorem (Blum, Hopcroft, and Kannan, 2020).

Figure 7.4: High-Dimensional Gaussian.

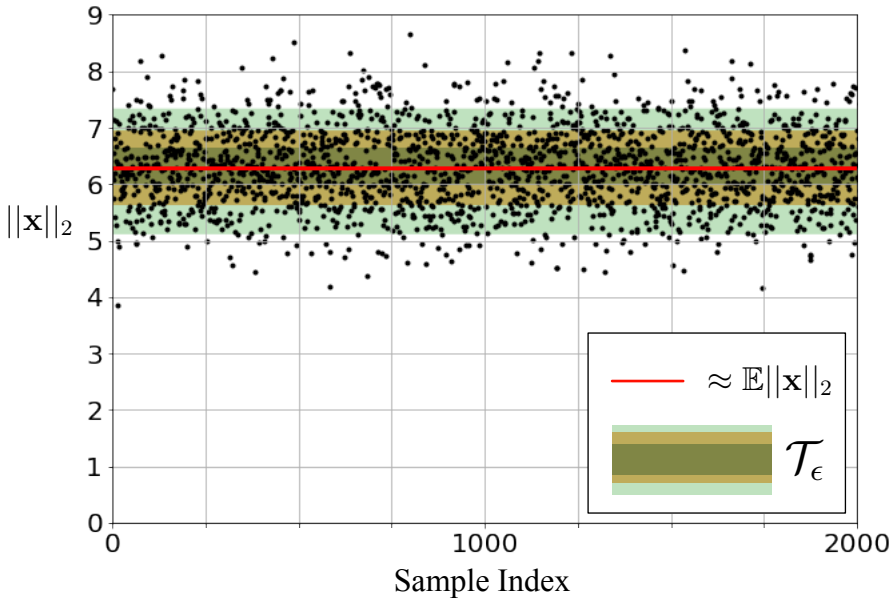


Note. The plot illustrates how, in high-dimensions, the probability mass is located in a thin annulus at a distance $\sigma\sqrt{D}$ from the average (in the text, we assume $\sigma = 1$), despite the mean representing the location which maximizes the probability density. Adapted from (MacKay, 1992).

unlikely to find datapoints anywhere close to the average (even though the region close to the mean represents the one with the highest likelihood, the probability is nonetheless negligible), (b) randomly sampled vectors are unlikely to be correlated (of course, in expectation the correlation will be zero because the dimensions of the Gaussian from which they were sampled are independent), and (c) randomly sampled vectors have lengths that are close to the expected length which increases at a rate $\sqrt{D}$. As such, the datapoints tend to cluster in a subspace which lies at a fixed radius from the mean (we will later refer to this subspace as the *typical set*). This is summarized graphically in Figure 7.4.

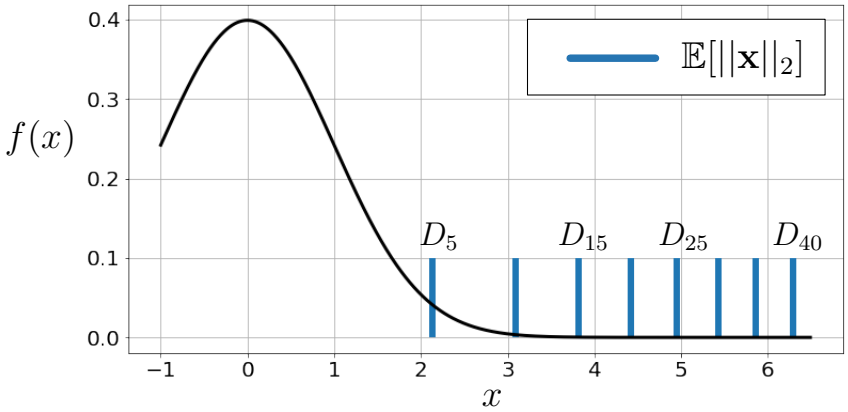
It is important that researchers understand that while the mean of such a high-dimensional Gaussian represents the value which minimizes the sums of squared distances (and is therefore the estimate which maximises the likelihood), most of the probability mass is actually *not* located at this point. As such, even though a set of values close to the mean represents the most likely in terms of its probability of occurrence, the magnitude of this probability is negligible, and most points fall in a space around $\sqrt{D}$ away from the mean. Figure 7.5 depicts the lengths of 2000 vectors sampled from a 40-dimensional Gaussian - they are nowhere close to the origin. Another way to visualize this is to plot the locations of the expected lengths for different dimensionalities on top of the curve for $\mathcal{N}(0, 1)$, and this is shown in Figure 7.6. In terms of the implications for psychological data - datasets which involve high numbers of variables are likely to comprise individuals who are similar only insofar as they appear to be equally ‘abnormal’, at

Figure 7.5: Vectors in High Dimensions and Typical Sets



Note. A scatter plot showing the lengths of 2,000 vectors sampled from a 40-dimensional Gaussian. Red line shows the average vector length, and the green intervals depict the size of the typical set for different values of ϵ. Note that the mean (0,0) is nowhere near the distribution of norms or the typical set.

Figure 7.6: Expected Lengths in Relation to the Standard Normal.



Note. This plot shows the location of the expected lengths of vectors of different dimensionality in relation to the standard normal in one dimension. It can be seen that even at $D = 5$, the expected length is over two standard deviations from the mean.

least insofar as a univariate characterization of normality (e.g. the mean across the dimensions) is a poor one when used across multiple dimensions. Indeed, if an individual *does* possess characteristics close to the mean or the mode across multiple dimensions, they could reasonably be considered to be *outliers*. We will consider outlier detection more closely in a later section.

7.4 Typicality: An Information Theoretic Way to Characterize 'Normality'

In the previous section, we described how randomly sampled vectors in high-dimensional space tend to be located at a radius of length $\sqrt{D}$ away from the mean, and tend to be uncorrelated. This makes points close to the mean across multiple dimensions poor examples of 'normality'. In this section we introduce the concept of *typicality* from information theory, as a means to categorize whether a particular sample or a particular set of samples is/are 'normal' or 'abnormal' (and therefore also whether the points should be considered to be outliers).

7.4.1 Asymptotic Equipartition Property and Entropy

A few concepts should first be introduced. Once again, let us start with something familiar: The well celebrated Law of Large Numbers (LLN). LLN states that the expected value of independent and identically distributed random variables is close to the empirical approximation to this expected value for large sample sizes. More formally:

$$\mathbb{E}\left[x\right] \approx \frac{1}{N}\sum_{i=1}^{N} x_i = \hat{\mu} \text{ for sufficiently large } N \tag{7.4}$$

There exists an analogue of this law in information theory, known as the Asymptotic Equipartition Property (AEP). AEP itself is described in terms of entropy H, which is a useful quantity in its own right.[4] Entropy describes the degree of surprise, uncertainty, or information associated with a distribution and is computed as $-\sum_{i=1}^{N} p(x_i)\log p(x_i)$ where N is the number of datapoints in the sample distribution, x_i is a single datapoint in this distribution, and $p(x_i)$ is that datapoint's corresponding probability. If the entropy is low, it means the distribution is more certain and therefore also easier to predict.

Taking a fair coin as an example, $p(x = \text{heads}) = p(x = \text{tails}) = 0.5$. The entropy of this distribution is $H = -(0.5\log_2 0.5 + 0.5\log_2 0.5) = 1$. Recall from above that entropy

[4] We temporarily consider the discrete random variable case for this example, but note that the intuition holds for continuous distributions as well.

describes the amount of information content - the units of entropy here are in bits. The fact that our fair coin has 1 bit of information should therefore seem quite reasonable - there are two equally possible outcomes and therefore one bit's worth of information. Furthermore, because the coin is unbiased, we are unable to predict the outcome any better than by randomly guessing. On the other hand, let's say we have a highly biased coin whereby $p(x = \text{heads}) = 0.99$ and $p(x = \text{tails}) = 0.01$. In this case $H = -(0.99 \log_2 0.99 + 0.01 \log_2 0.01) = 0.08$. The second example had a much lower entropy because we are likely to observe heads, and this makes samples from the distribution more predictable. As such, there is less new or surprising information associated with samples from this distribution, than there was for the case where there was an equal chance of a head or a tail.

AEP states that (Cover and Thomas, 2006):

$$\frac{-1}{N} \log p(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) \approx H(\mathbf{x}) \text{ for sufficiently large } N \tag{7.5}$$

which, for *i.i.d.* samples (and/or variables), yields:

$$\frac{-1}{N} \sum_{i=1}^{N} \log p(\mathbf{x}_i) \approx H(\mathbf{x}) \text{ for sufficiently large } N \tag{7.6}$$

In words, the negative log of the joint probability tends towards the entropy of the distribution. Note that the form given in 7.6 is very similar to the form for the expectation in 7.4: whilst the expected value of the distribution tends towards the mean, the log joint probability tends towards the entropy of the distribution. AEP therefore gives us an alternative way to characterize normality: but instead of doing so using the arithmetic mean, we do so in terms of entropy. Now, rather than comparing the value of a new sample against the mean or expected value of a distribution, we can now consider the probability of observing that sample and its relation to the entropy of the distribution.

7.4.2 Defining the Typical Set

We are now ready to define the typical set. Rather than comparing datapoints against the mean, we can compare them against the entropy of the distribution $H(\mathbf{x})$. For a chosen threshold ϵ, datapoints may be considered typical according to (Dieleman, 2020; Cover and Thomas, 2006; MacKay, 2018):

$$\mathcal{T} = \{\mathbf{x} : 2^{-(H+\epsilon)} \leq p(\mathbf{x}) \leq 2^{-(H-\epsilon)}\} \tag{7.7}$$

In words, the typical set $\mathcal{T}$ comprises datapoints $\mathbf{x}$ which fall within the bounds defined either side of the entropy of the distribution. Datapoints which have a negative log likelihood close (where close is defined according to the magnitude of ϵ) to the entropy of the distribution are thereby defined as typical. The quantity given in Eq. 7.7 can be computed for Gaussian data using the analytical forms for entropy H for the univariate and multivariate Gaussian provided as Supplementary, and the probability density function for a univariate or multivariate Gaussian for $p(\mathbf{x})$. This is undertaken for the outlier detection simulation below.

Recall the thin annulus containing most of the probability mass, illustrated in Figure 7.4; this annulus comprises the typical set. Note that, because this annulus contains most of our probability mass, the set quickly incorporates all datapoints as ϵ is increased (Cover and Thomas, 2006). Note that this typical set (at least for 'modest' values of ϵ) does not contain the mean because, as an annulus, it cannot contain it by design (the mean falls at the centre of a circle who's radius defines the radius of the annulus).

7.4.3 Establishing Typicality in Practice

Even though it is arguable as to whether the Gaussian should be used less ubiquitously for modeling data distributions than it currently is (Micceri, 1989), one of the strong advantages of the Gaussian is its mathematical tractability. This tractability enables us to calculate (as opposed to estimate) quantities exactly, simply by substituting parameter values into the equations (assuming these parameters have themselves not been estimated). Thus, moving from a comparison

of dataset values against the average or expected value to a consideration for typicality does not necessitate the abandonment of convenient analytic solutions. A derivation of the (differential) entropy for a Gaussian distribution has been provided in supplementary material, and is given in Eq. 7.8.

$$H(f) = \frac{1}{2}\log_2(2\pi e\sigma^2) \tag{7.8}$$

Note that the mean does not feature in Eq. 7.8 - this make it clear that the uncertainty or information content of a distribution is independent of its location (*i.e.*, the mean) in vector space.[5] As well as being useful in categorising datapoints as typical or atypical (or, alternatively, inliers and outliers) in practice, Eq. 7.8 can also be used to understand the relationship between ϵ and the fraction of the total probability mass that falls inside the typical set. Returning to Figure 7.5 which shows the lengths of 2,000 vectors sampled from a 40-dimensional Gaussian, we can see that as ϵ increases, we gradually expand the interval to cover a greater and greater proportion of the empirical distribution. Note also that the mean, which in this plot has a location (0,0) is a long way from any of the points is not part of (and, by definition, cannot be part of) the typical set.

7.5 An Example with Real-World Data

To demonstrate that these effects do not only apply to idealistic simulations, we use the LISS longitudinal panel data, which is open access (Scherpenzeel and Das, 2010). Specifically, we use Likert-style response data from wave 1 of the Politics and Values survey, collected between 2007 and 2008, which includes questions relating to levels of satisfaction and confidence in science, healthcare, the economy, democracy etc. Given that no inference was required for these data, a simple approach was taken to clean it: all non-Likert style data were removed, leaving 58 variables, and text based responses which represented the extremes of the scale we replaced with integers (*e.g.*, 'no confidence at all' is replaced with a 0). For the sake of demonstration, all

[5] Note that entropy is closely related to the score function (the derivative of the log likelihood) as well as Fisher information, which is the variance of the score.

missing values were mean-imputed[6] (this may not be a wise choice in practice), and the data were standardized so that all variables were mean zero with a standard deviation of one. In total there were 6,811 respondents.

Figure 7.7 depicts the bivariate correlations for each pair of variables in the data. It can be seen that there exist many non-zero correlations, which makes these data useful in understanding the generality of our expositions above (which were undertaken with uncorrelated variables). Qualitatively, some variables were highly non-Gaussian, which again helps us understand the generality of the effects in multi-dimensional data. Figure 7.8 shows how the expected lengths of the vectors in the LISS panel data change as an increasing number of dimensions are used. To generate this plot, we randomly selected D variables 1000 times, where D range from three up to the total number of variables (58). For each of the 1,000 repeats, we computed the Euclidean distances of each vector in the dataset across these D variables, and then computed their average. Once the 1,000 repeats were complete, we compute the average across these repeats to obtain an approximation to the expectation of vector lengths in D dimensions. Finally, we overlaid a plot of $\sqrt{D}$ to ascertain how close the empirically estimated vector lengths are, compared with the expected lengths for a multivariate Gaussian. We also plot the 1-99% intervals, which are found to be quite wide, owing to the mix of lowly and highly correlated variables in conjunction with possibly non-Gaussianity.

These results demonstrate that even for correlated, potentially non-Gaussian, real-world data, the peculiar behaviour of multi-dimensional data discussed in this paper still occur. For the LISS data, the expected lengths were slightly lower than for samples from a 'clean' multivariate Gaussian, and this is likely to be due to the correlations present in the data.[7]

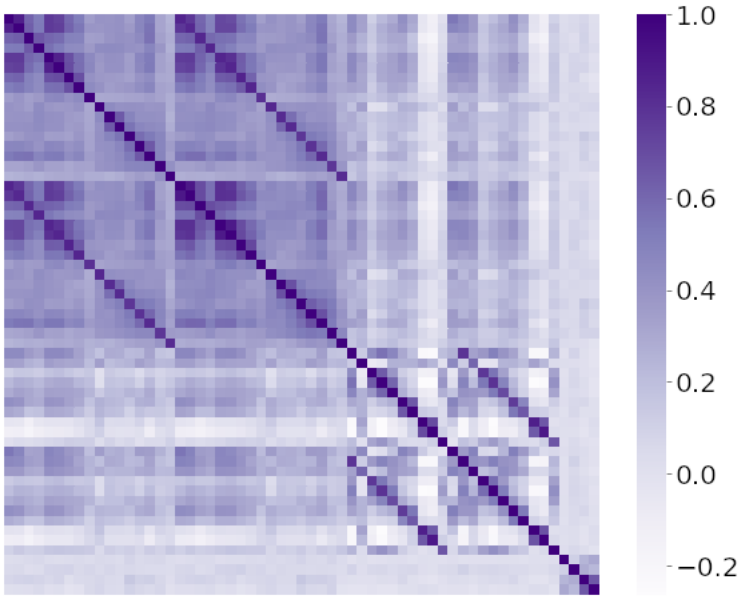
7.6 Moving Forward with Multivariate Outlier Detection

Grubbs defined outliers as samples which "deviate markedly from other members of the sample in which it occurs" (Grubbs, 1969). This definition is useful to us here, because it is not

[6] Across all included variables the amount of mean-imputation, on average, was 7.9%. Note that such imputation makes the demonstration more conservative, because it forces values to be equal to the mean for the respective dimension.

[7] For further discussion relating to this point, see Kroc and Astivia (2021).

Figure 7.7: LISS Panel Data Correlation Matrix

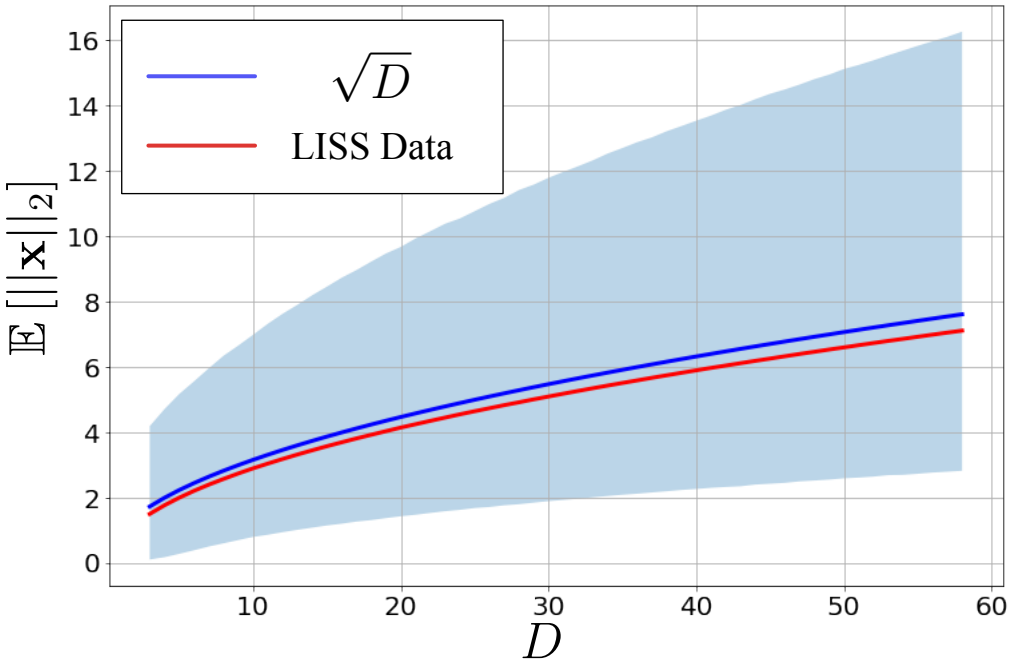


Note. Depicts the bivariate correlations for the LISS panel data (Scherpenzeel and Das, 2010).

expressed in terms of distance from the mean, but in broad/general terms. Indeed, as we have already discussed, in as few as four dimensions, points near the mean become increasingly unlikely. This suggests that outlier methods should not only identify points which are too far from the mean, but also those which are too close.

Two related definitions of outliers which were noted by Leys, Delacre, et al. (2019) are: "Data values that are unusually large or small compared to the other values of the same construct", and "Data points with large residual values." The first is quite similar to Grubbs' definition, identifying values as unusually large or small (*i.e.*, deviating markedly) with respect to other values of the same construct (*i.e.*, with respect to the other members of the sample in which they occur). The second defines them with respect to the residuals of a statistical model. In other words, they are values which lead to large discrepancies between true and predicted values. Note that both of these definitions bear the consequences for our work - whether we are comparing datapoints against the rest of the sample, or comparing them against the predictions from a

Figure 7.8: LISS Panel Data Vector Lengths



Note. The lengths for vectors from the LISS panel data (red), for increasing D, as well as the expected lengths for a multi-variate Gaussian (blue). The LISS panel data curve includes 1-99% percentile intervals (Scherpenzeel and Das, 2010).

statistical model designed to estimate an expected value (which is by far the most common case in psychology and social science), the relevance of these definitions to our discussion remains the same.

It is also, perhaps, of interest to note that our definition of outliers makes not value judgement about whether outliers are good or bad. Indeed, depending on the application and our research questions, outliers may represent 'golden' samples. Consider a manufacturer interested in fabricating the perfect mechanical prototype. Each sample may have its own unique blemishes, and our target may represent the perfect average across all (high-dimensional) opportunities for such blemishes. In such a case, the average represents the golden target for our manufacturer, and identifying it necessitates outlier detection methods which understand that values across high-dimensions close to the mean should be considered to be (in this case, desirable) outliers, in much the same way as samples which deviate because they are too far from the mean may also be outliers for opposite reasons.

Leys, Delacre, et al. (2019) provide a useful summary of options for both univariate and multivariate outlier detection, as well as a discussion about the consequences of outlier management decisions. Whilst their work provide an excellent introduction to multivariate outlier detection and good practice, they do not discuss the strange behaviour of the mean in multiple dimensions, nor the impact of this behaviour on multivariate outlier detection methods which are unable to detect outliers which lie close to the mean

We note, as other researchers have (Leys, Delacre, et al., 2019), that the most common method used for multidimensional/multivariate outlier detection in the domain of psychology is the Mahalanobis distance (Mahalanobis, 1930). For a description of the Mahalanobis distance and its application, readers are directed to work by X. Li et al. (2019) and Leys, Klein, et al. (2018). Briefly, the method assesses the distance of a point from the centroid (*i.e.*, the mean) of a cloud of points in (possibly correlated) multidimensional space. The researchers note that in order to compute the distance from putative outliers to the mean, it is first necessary to estimate the mean and covariance whilst including those points in the estimation (Leys, Ley, et al., 2013; Leys, Klein, et al., 2018). This process is somewhat problematic because if outliers are included in the calculation being used to compute the mean and covariance, the estimation of these quantities will themselves be biased towards these outliers, thereby reducing the chances of correctly identifying the outliers. A solution is proposed which is called the 'robust' Mahalanobis distance (X. Li et al., 2019; Leys, Klein, et al., 2018), which leverages what is known as the Minimum Covariance Determinant (MCD), and estimates the centroid / mean by selecting an estimate of the mean from a set of estimates derived from different subsets of the dataset.

Unfortunately, despite the Mahalanobis distance and its robust variant being the most commonly used multidimensional/multivariate outlier detection techniques in psychology, it suffers from the same problems as any multidimensional method based on distances from the centroid/mean. By consequence it would certainly not flag someone average in all dimensions as an outlier, even though statistically they would represent an extremely unusual individual (it would not help Daniels with his project, for example). It is therefore important that researchers qualify their definition of the outlying set to explicitly admit points which may fall too close to the mean.

When using the Mahalanobis distance, one can make decisions about the set of outliers $\mathcal{O}$ using the following expression:

$$\mathcal{O} = \{\mathbf{x} : M(\mathbf{x}) > c\}, \tag{7.9}$$

where $M(\mathbf{x}) = \sqrt{(\mathbf{x} - \hat{\boldsymbol{\mu}})^T \mathbf{S}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}})}$ and is the estimated Mahalanobis distance (in units of standard deviation) for the multivariate datapoint under consideration $\mathbf{x}$, and c is the threshold for classifying a point as an outlier. In the expression for $M()$, $\hat{\boldsymbol{\mu}}$ is the estimate of the mean of the distribution, $\mathbf{S}$ is the estimated covariance matrix. One of the benefits of the Mahalanobis based methods is that one can use them to threshold the data based on units of standard deviations. Thinking in terms of standard deviations is not unusual and therefore the process of selecting outliers in these terms thus leads to intuitive selection thresholds. In contrast, we see in Eq. 7.7 that the threshold for determining whether a datapoint falls within the typical set $\mathcal{T}$ depends on ϵ, which is not related to the standard deviation, but rather to a distance away from the entropy.

We have already seen how typicality has the added advantage of classifying datapoints which lie too *close* to the mean. In Figure 7.9 we show that, in low-dimensional settings, typicality can be used to make approximately the same classification of outliers as the Mahalanobis distance to the extent that some datapoints which lie *far* from the mean should still be classified as outliers. Of course, in practice a balance must be struck between the value of ϵ in Eq. 7.7, in the same way that c in Eq. 7.9 must be decided.

Specifically, for Figure 7.9, we generated 125 points from a bivariate Gaussian with a covariance of 0.5, and then added a set of equally spaced outlier points ranging from negative four to positive four on the y-axis (indicated with horizontal dashes). As such, not all these points are expected to be identified as outliers, because some of their values lie well within the tails of the distribution. They do, however, enable us to compare at which point they are identified as outliers by the two detection methods under comparison. Note that the subsequent estimation is done after the creation of the complete dataset (including the outliers) using all the empirical values. Using the robust MCD estimator mentioned above, we computed both the Mahalanobis distance (in units of standard deviation), and colored each point according to this distance. For

typicality, we followed the estimation of entropy for the multivariate Gaussian which also takes in an estimate for the covariance (see the Supplementary Material for the relationship between the covariance matrix and the entropy of a Gaussian), for which we again used the MCD method. The use of MCD for typicality arguably arguably makes our typicality estimator 'robust' for the same reason that it is considered to make the Mahalanobis distance estimation robust. The threshold for the Mahalanobis distance was set to three standard deviations, whilst the value for the typicality threshold was set to five. In practice, researchers may, of course, need to suitably select and justify these values.

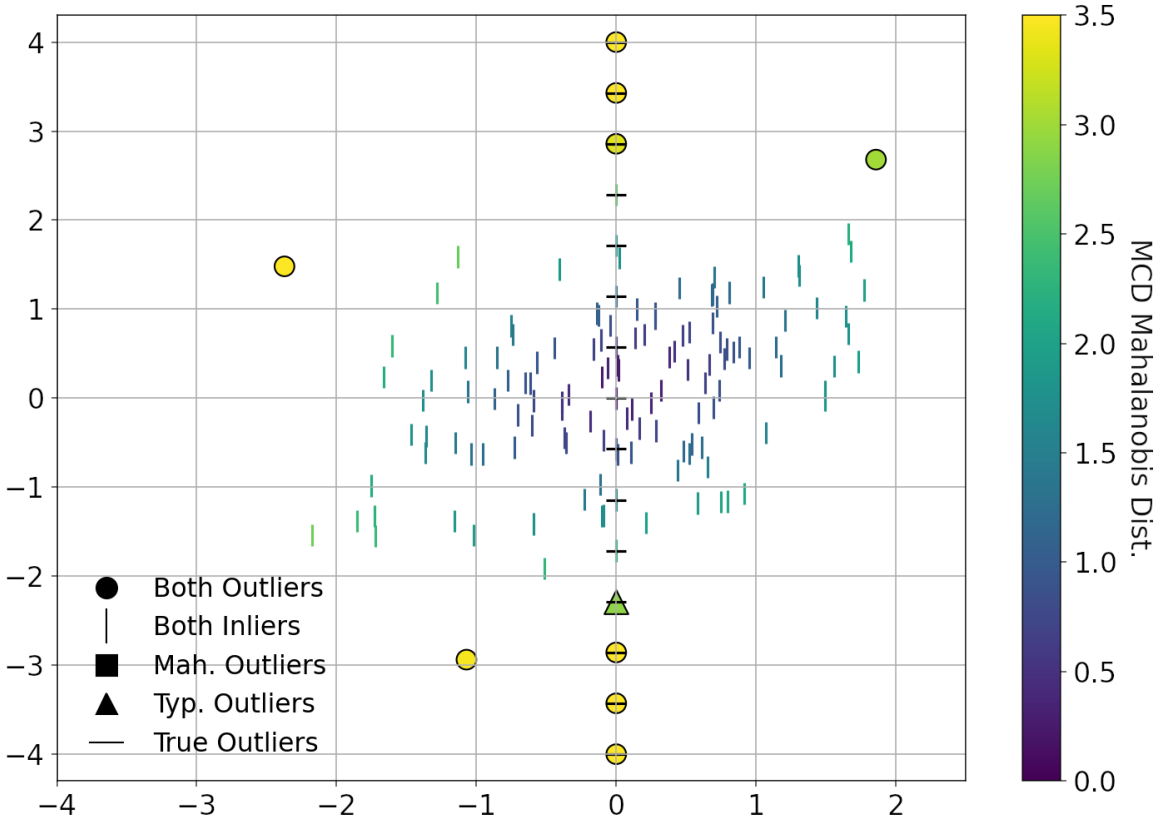
The scatter-plot marker shapes are set according to whether the outliers were classified as such by both methods (circles), just the Mahalanobis method (squares), or just the typicality method (triangles). If neither method classifies a point as an outlier, the points are set to vertical dashes (*i.e.*, 'inliers'). Note that there are no points which are classified as outliers by the Mahalanobis method which are not also classified as outliers by the typicality method. The inverse is not quite true, with one additional point (indicated with the triangle marker) being classified as an outlier by the typicality method.[8] Figure 7.9 therefore indicates that, in low-dimensions, Mahalanobis distance performance similarly to typicality as an outlier detection method.

In Figure 7.10, we undertake the same task, but this time in a 20-dimensional space. The figure shows the lengths of each of 1400 points (the lengths are used for visualisation purposes) drawn from a 20-dimensional, isotropic Gaussian. Fifteen of these points are manually set to fall very close to the mean / expected value of zero, and these are the simulated outliers we wish to identify. Now, in contrast to the example above, we see a large difference between the outliers identified using the two methods. Typicality successfully identifies all 15 true outliers as outliers, whereas MCD fails to identify any of them. Conversely, some points which lie far from the mean (but which have a low probability of occurrence relative to the entropy of the distribution) are identified by both MCD and typicality, although it is possible that by tweaking the thresholds one could achieve greater overlap between the classification of these points by the two methods.

In summary, typicality does not only have a role in detecting outliers in high-dimensional scenarios (where the outliers may include values close to the expected value), but can perform

[8] Although this classification is technically correct, this point lies on the limit of the cloud of true inliers, and so in practice it would not be clear whether this would represent a useful outlier classification or not.

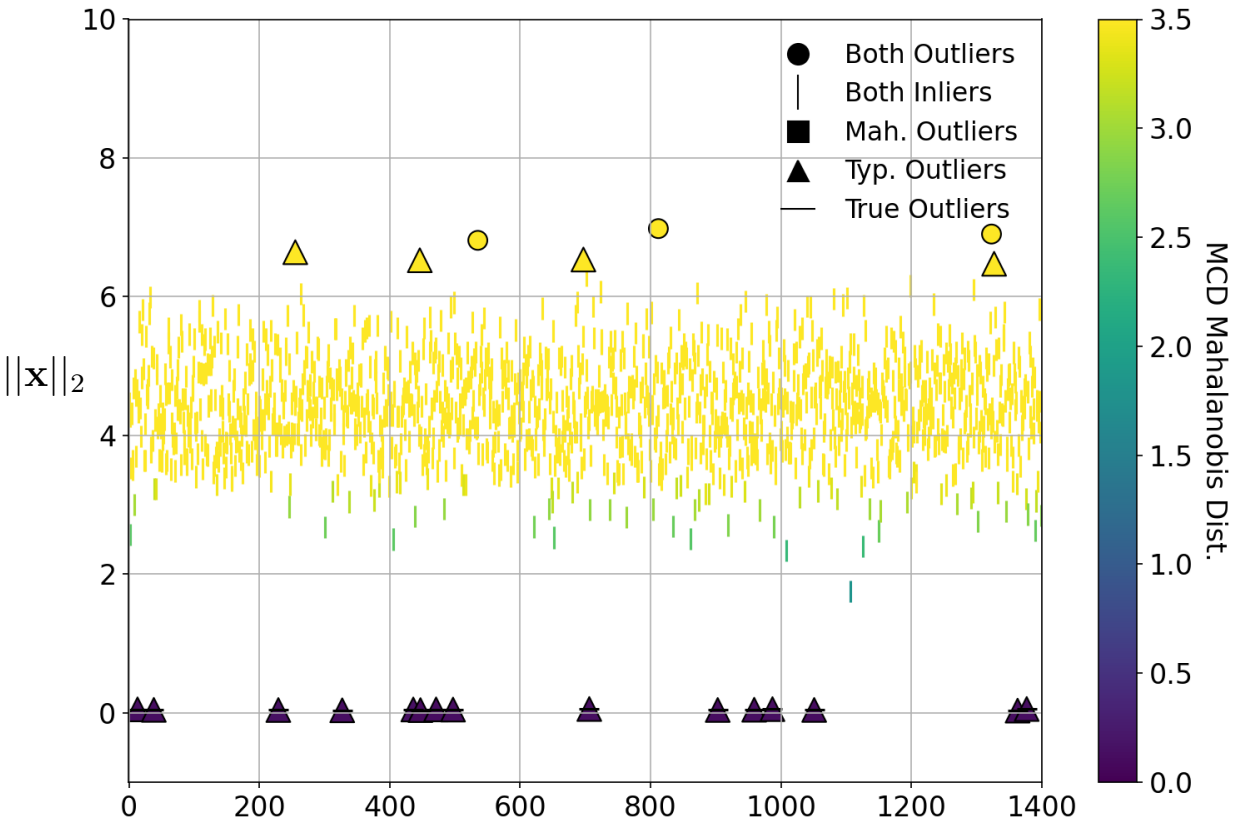
Figure 7.9: Outlier detection comparison.



Note. Comparison of Mahalanobis distance and typicality for outlier detection. The outliers are generated as a vertical set of equally spaced points (indicated with horizontal dashes) ranging from negative four to positive four on the y-axis, superimposed on a set of 125 points (indicated with vertical dashes) drawn from a bivariate Gaussian with a covariance of 0.5. The points identified to be outliers by both methods are indicated in circles, whilst those indicated to be outliers by the the Mahalanobis or typicality methods separately are indicated by squares or triangles, respectively. The color of the points represents the Mahalanobis distance in units of standard deviation. The estimation of the covariance matrices for both methods used the robust Minimum Covariance Determinant (MCD) method.

similarly to how current approaches (such as MCD) do in low-dimensional scenarios, which otherwise fail in high-dimensions. We thus recommend practitioners consider using typicality as a valid outlier detection approach under both low- and high-dimensional conditions, and especially in high-dimensions. To this extent, researchers a encouraged to consult various commentaries on the usage of outlier detection methods, such as the one by Leys, Delacre, et al. (2019) which provides general recommendations for practice (including pre-registration). It is notable that prior commentary does not include a discussion about the limitations of

Figure 7.10: Outlier detection comparison.



Note. Comparison of Mahalanobis distance and typicality for outlier detection in 20 dimensional space. The outliers are generated as a set of 15 points close to the expected value of 0, superimposed on a set of 1400 points drawn from a 20 dimensional isotropic Gaussian. For visualation purposes, this plot shows the lengths of each point (the x-axis is simply the index of the point in the dataset). The points identified to be outliers by both methods are indicated in circles, whilst those indicated to be outliers by the the Mahalanobis or typicality methods separately are indicated by squares or triangles, respectively. The color of the points represents the Mahalanobis distance in units of standard deviation. The estimation of the covariance matrices for both methods used the robust Minimum Covariance Determinant (MCD) method.

Mahalanobis based methods for outlier detection once the number of dimensions increases, which serves as a reminder of how important it is that researchers explore typicality. Finally, we recommend updating the working conceptualisation of outliers to include those points which, in high-dimensions (but as few as 4-10 dimensions) fall too close to the mean.

7.7 Conclusion

The arithmetic mean has been used both productively and unproductively as a blunt way to characterize samples and populations. It has been used to pathologize deviations from 'normality', where normality has been said to represent a harmonious ideal. Through our exploration of multi-dimensional space, we have shown that the mean, far from representing normality, actually represents abnormality, in so far as encountering a datapoint close to the mean in datasets comprising more than a handful of dimensions becomes incredibly unlikely, even with a large number of datapoints.

In contrast with the arithmetic average, the information theoretic quantity known as 'typicality' provides a way to establish normality (or rather, whether a datapoint is typical or atypical), which is particularly useful in high-dimensional regimes. Given that researchers in psychology and social science frequently deal with multivariate datasets, and that the peculiarities associated with multi-dimensional spaces start occurring in relatively low dimensions (as few as four), it is important that researchers have some awareness of the concepts presented in this paper.

Clearly, the motivations behind the characterizations of points as either normal or abnormal overlap strongly with those behind outlier detection. The discussion also provides us with a good justification for updating our working definition of 'outlier' to include points which lie unusually close to the mean. Unlike popular multivariate outlier detection techniques such as the Mahalanobis distance, which characterize outliers as points which lie *far* from the expected value of the distribution, typicality additionally offers a means to detect those which are *close*. Whilst such additional benefits of typicality based methods become more evident as the dimensionality of the dataset increases (where traditional methods like Mahalanobis distance fail) we showed that typicality also performs as one would hope/expect in low dimensions. To show this, we finished with an evaluation of typicality for bivariate outlier detection using a 'robust' version of entropy using the Minimum Covariance Determinant estimation technique, and verified via simulation that in low-dimensions it works well as an alternative to the popular Mahalanobis distance. Researchers are encouraged to consult the supplementary material, which includes code for computing and applying typicality based methods (including those for outlier

detection).

7.8 Supplementary: Differential Entropy of a Gaussian

Following Cover and Thomas, 2006, the *differential entropy* (in bits) is defined as:

$$H(f) = -\mathbb{E}[\log(f(x))] = -\int_{-\infty}^{+\infty} f(x) \log_e f(x) dx \tag{7.10}$$

The probability density function of the normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7.11}$$

Substituting the expression for $f(x)$ into $h(f)$:

$$H(f) = -\int_{-\infty}^{+\infty} f(x) \log_e \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right] \tag{7.12}$$

$$H(f) = -\int_{-\infty}^{+\infty} f(x) \log_2 e \left(\log_e \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log_e e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) \tag{7.13}$$

$$H(f) = -\int_{-\infty}^{+\infty} f(x) \log_2 e \left(-\log_e(\sqrt{2\pi\sigma^2}) - \frac{(x-\mu)^2}{2\sigma^2} \right) \tag{7.14}$$

$$H(f) = \log_2 e \log_e \sqrt{2\pi\sigma^2} \int_{-\infty}^{+\infty} f(x) dx + \log_2 e \int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{2\sigma^2} f(x) dx \tag{7.15}$$

Note that:

$$\int_{-\infty}^{+\infty} f(x) dx = 1 \tag{7.16}$$

and recall that:

$$\int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx = \mathbb{E}[(x-\mu)^2] = \text{Var}(x) = \sigma^2 \tag{7.17}$$

Therefore:

$$H(f) = \log_2 \sqrt{2\pi\sigma^2} + \frac{\log_2 e}{2\sigma^2}\sigma^2 \tag{7.18}$$

And finally:

$$H(f) = \frac{1}{2}\log_2(2\pi e\sigma^2) \tag{7.19}$$

For a D-dimensional Gaussian, the derivation for the entropy is as follows:

$$H(f) = -\mathbb{E}[\log(f(\mathbf{x}))] = -\int_{-\infty}^{+\infty} f(\mathbf{x}) \log_e f(\mathbf{x}) d\mathbf{x}, \tag{7.20}$$

where the bold font indicates multidimensionality.

$$H(f) = -\mathbb{E}[\log[(2\pi)^{-D/2}|\mathbf{S}|^{-0.5}\text{exp}(-0.5(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu}))]], \tag{7.21}$$

where $\mathbf{S}$ is the covariance matrix, T indicates the transpose, and $|.|$ indicates the determinant.

$$H(f) = 0.5D\log(2\pi) + 0.5\log|\mathbf{S}| + 0.5\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})] \tag{7.22}$$

$$= 0.5D(1+\log(2\pi)) + 0.5\log(\mathbf{S}) \tag{7.23}$$

This last expression can then be expressed in bits by multiplying by $\log_2 e$. Note that this derivation follows the approach provided by Gundersen (2020) and uses a number of 'tricks' relating to the trace operator

CHAPTER 8

Discussion, Limitations, and Future Work

Despite the attention that the replication crisis has drawn to itself, and the increased pressure to preregister studies, undertake power analyses, and collect data from a sufficiently large number of participants, I have argued that replication represents only one challenge out of many which face researchers in psychology and social science. Independently of whether data have been collected for a sufficient number of participants, if the statistical models we use are either functionally or structurally/causally misspecified, or both, the inference we undertake using these data can be arbitrarily biased. In the introduction I also discussed the high inherent complexity of psychological phenomena and the potential ramifications for research. Confronting this complexity can lead to an impasse - but I argued an awareness of it can help us identify the opportunities to improve research, and remain skeptical and cautious in our search for 'the truth'.

In Chapter 2, I discussed the nature of the misspecification problems at length, and provided some recommendations for researchers to (1) consider machine learning approaches for alleviating problems with the assumption of linearity; (2) to collaborate with researchers outside of psychology (for instance, experts in machine learning, statistics, engineering, causality); (3) to be more transparent and specific about whether they are taking a predictive or causal approach to research; (4) that researchers be concise and not overambitious in the specification of their research questions and hypotheses.

In Chapters 3 and 4, I provided application examples for the recommendations made in Chapter 2.

Specifically, in Chapter 3 I provided an example of machine learning and machine learning explainability techniques applied to the task of predicting perceived partner support from relational and individual variables. Whilst a psychological interpretation of these results is beyond the scope of this thesis (which is method based), the demonstration provides evidence for the utility of machine learning and explainability in exploring the data for possibly important associations. Then, in Chapter 4, I provided an example of causal discovery, machine learning, and causal inference techniques (specifically targeted learning) in combination, to demonstrate how we can use these tools to further knowledge in the domain of attachment styles and mental health. Again, a psychological interpretation of the results is beyond the scope of this thesis, although the relevance and utility of these techniques has been demonstrated in this work.

In Chapter 5, I introduced some techniques from the literature on Directed Acyclic Graphs and Probabilistic Graphical Models to identify ways in which the mathematical, graphical representation of a theory can be simplified without affecting the reliability of the associated estimation process (*e.g.*, for estimating a particular causal effect size). The motivation for these techniques stems from (1) an acknowledgement for the inherent complexity of psychological theories - if there exist opportunities to simplify the models we ought to take them; (2) the need to prioritise data collection for variables which are crucial for the subsequent estimation process; and (3) to encourage transparency when translating our otherwise often verbal theories into unambiguous mathematical objects. I also provided code to simplify graphs automatically, given a particular research question.

Given the burgeoning popularity of machine learning based methods for prediction, and also given the recommendation that researchers take causality seriously, Chapter 6 provided a set of simulation results to illustrate that one cannot 'outrun' causality with the use of powerful algorithms. In other words, it is necessary to always consider the underlying causal structure of Data Generating Process, and how this structure manifests statistically in the data. One should therefore consider the phenomenon from both functional *and* structural perspectives together. The interaction between the prediction task and the causal structure can result in important causes of an outcome as being seemingly unimportant for prediction - it all depends on what is included in the model and, again, what the causal structure of the phenomenon is.

The last contribution is presented in Chapter 7. In this Chapter, I approached the complexity challenge in psychology from the perspective of outlier detection / normality / typicality. Specifically, I explore the ways in which data tend to behave as the number of variables we collect for an individual increases. As I demonstrated, unintuitive statistical behaviours start occurring with as few as four dimensions, and the average/expected value become increasingly unsuitable as a characterization of the distribution as this number increases. I show that whilst the expected value represents the location of highest probability density, there exists barely any real probability *mass* at this location (*i.e.*, very few datapoints fall close to the expected value). I identify typicality as a valuable information theoretic measure for quantifying 'normality' (in the sense that a datapoint is normal or abnormal), and demonstrate its performance against a popular alternative which otherwise fails in high dimensions

Throughout the work involved in developing the ideas presented in this thesis, I have had the opportunity to apply some of the proposals made to 'real-world' psychological applications - in the Declaration Section, I provide a list of such works (five accepted for publication, three under review at the time of writing). The statistical, causal, and machine learning approaches I discuss in this work have been adapted to a wide variety of problems relating to obesity, COVID, and mental health, partner support, sexual desire, and others. These projects help motivate and justify the real-world applicability of the proposals made herein.

8.1 Limitations, Reflections, and Further Work

Whilst the specific limitations for each contribution are discussed in the associated contribution Chapters, it is worth discussing the limitations associated with the work in this thesis. Firstly, it is worth emphasising that many of the problems we identified are indicated to exist *in general* in psychology, and of course are not problems that apply to all researchers in all domains of psychology and social science. Indeed, despite the recommendations made and the problems with current research highlighted in this thesis, the two application contributions (Chapters 3 and 4) both included sub-optimal research practice. For instance, both included compromises associated with the limitations of real-world data (*e.g.*, potential missingness issues, differences in measures used in otherwise combined samples, simple averaging of individual items to

compose the constructs). Such compromises highlight how difficult it is to undertake the 'perfect study'. Indeed, if one were unwilling to make such compromises, it would be very difficult to undertake any research at all. To this extent, all research is likely to be, and likely to always be, limited by practical constraints. On the one hand, I therefore understand that being too absolute about what constitutes good or bad practice can paralyse researchers when they face the reality of necessary compromise. On the other hand, we need to be aware of the problems in order to overcome them where possible. Analytically, at least, in combining state-of-the-art methods in causal discovery, machine learning, and causal inference, Chapter 4 represents a significant step forward.

Secondly, and on a related point, the conclusions of Chapter 6 seems to contradict the recommendation to utilize machine learning with explainability techniques made in Chapter 2. Part of this comes as a natural consequence of the evolution of understanding I personally underwent during the work undertaken for this thesis, which in turn reflects the self-correcting nature we hope is reflected in science in general (whether this occurs in reality a matter of debate, see McElreath, 2016; Ioannidis, 2012). Indeed, the degree to which machine learning methods, explainability techniques, and the underlying causal structure interact was not well known to be until undertaking the work presented in Chapter 6, and to this extent, it contains content at odds with that in Chapter 2. I hope, therefore, that by presenting the work as it was carried out, as well as reflecting openly upon it in this way, speaks more to the potential for science to self-correct, than it does to my initial over-optimism regarding exploratory machine learning.

Thirdly, and in a general sense, all of the contributions are more or less compatible with the dominating paradigm current in psychology and social science, and provide ways to ameliorate certain problems within this paradigm (for example, DAGs are a more general form of SEMs). Whilst on the one hand this might represent an advantage because it implies that the proposals can be readily assimilated into the field, it represents a limitation because some of the top-level challenges remain unaddressed. In particular, the complexity issue discussed in the Introduction and referred to as 'The Big Assumption', may preclude the utility such high-level models for many psychological phenomena. In addition, we cannot expect more advanced statistical methods (*e.g.*, data-driven machine learning algorithms) to fix issues with poor theories (see

also: Smaldino, 2019). Fundamentally, it may be necessary to take an altogether different approach to the modeling of such complex, dynamic phenomena.

One option which bypasses the problem of The Big Assumption is to learn from the approaches taken in the domains of machine learning and artificial intelligence, and to adopt state-of-the-art, computer vision, signal processing, and other highly adaptive real-time analytical techniques. The author has actually already been working on such projects (M. Vowels, 2020), from which the initial results are promising. Whilst such an approach might yield better predictive performance than overly reductionist alternatives, it makes the interpretation of such a model very challenging (although, once again, we are reminded that the human-interpretability of a phenomenon is not necessarily a given). In turn, this makes it difficult to use such models to design new interventions and to make new discoveries about the nature of the phenomenon under study. Future work therefore should also explore the potential of machine learning explainability techniques with such complex, multi-modal approaches (*i.e.*, combining audio, language, vision, etc.).

Whether or not the use of highly parameterized models from the domains of machine learning and artificial intelligence really represent the future of psychology, there exist many problems with the *status quo* which can be substantially ameliorated without necessitating such a paradigm shift. Issues with replication, the over-utilization of linear models, conflation of correlational and causal approaches to research, a general lack of statistical awareness and engagement with meta-research, vague and untestable theories and hypotheses, etc. - all these issues can be vastly improved before we necessarily need to tackle more fundamental questions about research philosophies in psychology and social science. Indeed, perhaps if the field were to move in a positive direction with respect to the existing problems, the answers to the fundamental questions about the appropriateness of the existing research paradigm (*i.e.*, that psychological phenomena can be modeled usefully and accurately using researcher-specified mathematical models) would emerge naturally as part of rigorous research practice. In other words - by improving current research practice, the answer to whether or not a fundamental paradigm shift is needed may naturally emerge. As it stands, the current problems have the potential to seriously hinder progress (in any direction) and inhibit the self-correcting nature of science.

Finally, on multiple occasions I use the term 'sufficiently' as a way of delineating whether or not our models are 'sufficiently correctly specified in terms of the structural and functional form, with respect to the true, underlying process (see, in particular, Chapter 2). One recalls the famous George Box quote 'all models are wrong and some are useful' (Box and Jenkins, 1976). Indeed, short of modeling the trajectories of all sub-atomic particles in the universe (if, indeed, this is enough), our models will always represent simplifications and abstractions of the true, underlying real-world processes. Thus, whether or not a model is 'sufficiently' correctly specified is a subjective point which depends on some (arbitrary) measure of usefulness for a particular task, for a particular group of researchers. For instance, if we are interested in generalizing the estimated efficacy of a drug from our empirical sample to a population, our model has to be sufficiently correct for our analytical process to yield practically meaningful results. If our estimates are wildly biased (as I argue that many in psychology are likely to be), then clearly our model is not sufficiently correctly specified. If, on the other hand, they are biased by some negligible quantity (where, again, negligible has to be defined according to the task and our priorities) we might say the model *is* sufficiently correctly specified. Understanding the degree to which our model is biased is part of the process in science, and as my model explains observations more accurately, we progress (hopefully) towards a more correctly specified model, with respect to our particular measures of accuracy.

8.2 Putting Into Practice

In this thesis I made a number of proposals for improving practice and demonstrated with two application papers that these recommendations are not just ideas but can be put into practice today. As I have noted, I am certainly not the first to highlight some of the problems in psychology and social science, and neither am I the first to propose solutions. Unfortunately, despite many years (in some cases, decades) of meta-research commentary, the field still exhibits many of the same problems, as evidenced by the persistence and tone in these commentaries over the years. Of course, some things have changed / are changing. For instance, pre-registration is more common than it used to be, and being recommended and requested by some journals and funding agencies (Kupferschmidt, 2018). On the other hand, we have already discussed the

reluctance of researchers to use and engage with causal methods (Grosz, Rohrer, and Thoemmes, 2020; Rohrer, 2018), how linear models are still by far the most common family of statistical models being used (Blanca, Alarcon, and Bono, 2018), etc. And this, in my view, has not changed substantially over the decades.

So, how to we know we are making progress and how do we know that the work in this thesis (for example) can/will make a difference? We might be pessimistic about the answer to this question if we also accept that, to some extent, change cannot be achieved if any single one of the recommendations is adopted, but rather requires systematic change across multiple dimensions together. The requirement for systematic change (as opposed to incremental progress following an accumulation of evidence; see Godfrey-Smith, 2003; Popper, 1959, for a discussion about, amongst other things, Karl Popper's alternative viewpoint on the rational progression of science) is reflected in discussions from other researchers. For instance, McElreath (2016, pp.442) discusses the nature of progress in science as being achieved almost in spite of its own mechanisms:

"How can we reconcile such messy history, and widespread contemporary failure, with obvious successes like General Relativity? Science is a population-level process of variation and selective retention. It does not operate on individual hypotheses, but rather on populations of hypotheses. It comprises a mix of dynamics that may, over long periods of time, reveal the clockwork of nature. But these same dynamics generate error. So it's entirely possible for most findings at any one point in time to be false but for science in the long term to still function. This is analogous to how natural selection can adapt a biological population to its environment, even though most individual variation in any one generation is maladaptive."

This, in turn, reminds one of Max Planck's argument, which was also shared by Thomas Kuhn, that "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die and a new generation grows up that is familiar with it ..." (Planck, 1949; Kuhn, 1962)

In this sense, the instigation of a domain-wide shift in the approach to research, methodology, modeling, and analysis, takes time, and might only be achievable through the education of the next generation(s) of psychologists and social science researchers. Instead of only teaching stu-

dents about traditional forms of analysis, we might more broadly expose students to alternative modes of thinking and modeling (perhaps even the lumped parameter modeling techniques from engineering). They may then understand that there exists rich and lively debate as to the best modeling approaches for a particular task, and that the approaches themselves may fall at the intersection of statistics, causality, engineering, philosophy of science, machine learning, and artificial intelligence. Part of this also involves openness about the challenges the field currently faces.

The fact that this debate exists can itself be a stimulating topic of research, and encourage students to find the niche which suits their own passions (the topics of statistics and methodology, particularly in the domains of psychology and social science, are far from solved). Indeed, such transparency is probably well advised, in face of the conspicuous problems the domains face, and given the concomitant risk of disillusionment that might ensue once a student discovers that many of the 'facts' taught on typical psychological syllabi may not have the solid empirical foundations they might otherwise assume.

In my own experience, psychology attracts students with a wide variety of, often non-technical / non-research-based, aspirations. Thus, part of the shift towards training psychology and social science students who are eclectic in their methodological skill sets, may also necessitate the offering of multiple 'study tracks' for those technically-minded students who wish to specialize and to become the next generation of academics in psychology and social science. Unfortunately, even within the current paradigm, many students and professors misunderstand basic concepts relating to fundamental and traditional statistics (Gigerenzer, 2004; Cassidy et al., 2019). Yet another question then naturally emerges regarding *who* the field can call upon to update and improve the situation, if, in general, neither the current students, nor the current professors, have the expertise to do so by themselves. I leave an investigation of this question to future work, but might tentatively reason that by gradually fostering an increased interest in statistics and methodology in students and researchers, we might also indirectly begin to attract an increasing number of researchers from other domains (such as engineering, statistics, etc.), thereby forming a strong and diverse pool of researchers with technical expertise.

8.3 Conclusion

I conclude by returning to the quotation at the beginning of this thesis: "To question the foundations of a discipline or a practice is not necessarily to deny its value, but rather to stimulate a judicious and balanced appraisal of its merits." (Ashcroft and ter Meulen, 2004). A number of strong limitations associated with the current research paradigm in psychology were highlighted and discussed at length. In spite of the difficulties these problems present, I am optimistic that if researchers acknowledge them, then research methodology and analysis in psychology and social science can begin to move in a positive direction. Indeed, the nature of at least some of the problems, such as the ubiquity of unsophisticated research methodologies and analytical methods, would seem to encourage an optimistic interpretation of the situation: That there presently exists a tremendous opportunity to innovate and modernise the current approach to research, simply by assimilating recent advances and developments from other domains such as engineering, machine learning, and statistics. Psychology and social science are complex domains, full of rich and nuanced phenomena. The phenomena deserve to be represented and studied using research methodologies which are flexible enough to reflect this complexity.

Bibliography

Aarts, A. A. et al. (2015). "Estimating the reproducibility of psychological science". In: *Science* 349.6251, pp. 943–950. DOI: 10.1126/science.aac4716.

Aas, K., M. Jullum, and A. Loland (2019). "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values". In: *arXiv:1903.10464*.

Achen, C.H. (1977). "Measuring representation: perils of the correlation coefficient". In: *American Journal of Political Science* 21.4, pp. 805–815. DOI: 10.2307/2110737.

Ainsworth, M.D.S. et al. (1978). *Patterns of attachment: A psychological study of the strange situation*. Hillsdale, NJ: Erlbaum.

Alaa, A.M. and M. van der Schaar (2019). "Demystifying black-box models with symbolic metamodels". In: *33rd Conference on Neural Information Processing Systems*.

Allen, S.F., M.A. Wetherell, and M.A. Smith (2017). "The Cohen–Hoberman inventory of physical symptoms: Factor structure, and preliminary tests of reliability and validity in the general population". In: *Psychology and Health* 32.5, pp. 567–587. DOI: 10.1080/08870446.2017.1290237.

Angrist, J.D. and A.B. Krueger (2001). "Instrumental variables and the search for identification: From supply and demand to natural experiments". In: *Journal of Economic Perspectives* 15.4, pp. 69–85. DOI: 10.1257/jep.15.4.69.

Anzalidi, K. and K. Shifren (2019). “Optimism, Pessimism, Coping, and Depression: A Study on Individuals With Parkinson’s Disease”. In: *The International Journal of Aging and Human Development* 88.3, pp. 231–249. DOI: 10.1177/0091415018763401.

Arjovsky, M. et al. (2020). “Invariant risk minimization”. In: *arXiv:1907.02893v3*.

Aron, A., E.N. Aron, et al. (1991). “Close relationships as including other in the self”. In: *Journal of Personality and Social Psychologyersonality and Social Psychology* 60.2, pp. 241–253. DOI: 10.1037/0022-3514.60.2.241.

Aron, A. and B. Fraley (1999). “Relationship closeness as including other in the self: Cognitive underpinnings and measures.” In: *Social Cognition* 17.2, pp. 140–160. DOI: 10.1521/soco.1999.17.2.140.

Ashcroft, R. and R. ter Meulen (2004). “Ethics, philosophy, and evidence based medicine”. In: *Journal of Medical Ethics* 30.2, p. 119. DOI: 10.1136/jme.2003.007286.

Asuero, A.G., A. Sayago, and A.G. Gonzalez (2006). “The correlation coefficient: An overview”. In: *Critical Reviews in Analytical Chemistry* 36.1. DOI: 10.1080/10408340500526766.

Baker, D.H. et al. (2020). “Power contours: optimising sample size and precision in experimental psychology and human neuroscience”. In: *Psychological Methods*.

Baltrusaitis, T. et al. (2018). “OpenFace 2.0: Facial Behavior Analysis Toolkit”. In: *13th IEEE International Conference on Automatic Face and Gesture Recognition*.

Bartholomew, K. (1990). “Avoidance of intimacy: An attachment perspective”. In: *Journal of Social and Personal Relationships* 7.2, pp. 147–178. DOI: 10.1177/0265407590072001.

Bartholomew, K. and L.M. Horowitz (1991). “Attachment styles among young adults: A test of a four-category model”. In: *Journal of Personality and Social Psychology* 61.2, pp. 226–244. DOI: 10.1037/0022-3514.61.2.226.

Begley, C.G. and L.M. Ellis (2012). "Raise standards for preclinical cancer research". In: *Nature* 483, pp. 531–533. DOI: 10.1038/483531a.

Belkin, M. et al. (2019). "Reconciling modern machine-learning practice and the classical bias-variance trade-off". In: *PNAS* 116.32, pp. 15849–15854. DOI: 10.1073/pnas.1903070116.

Biggiogera, J. et al. (2021). "BERT meets LIWC: Exploring state-of-the-art language models for predicting communication behavior in couples' conflict interactions". In: *Companion Publication of the 2021 International Conference on Multimodal Interaction* 385-389. DOI: 10.1145/3461615.3485423.

Bijnens, E.M. et al. (2020). "Residential green space and child intelligence and bahavior across urban, suburban, and rural areas in Belgium: A longitudinal birth cohort study of twins". In: *PLoS Medicine* 17.8. DOI: 10.1371/journal.pmed.1003213.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Blanca, M.J., R. Alarcon, and R. Bono (2018). "Current practices in data analysis procedures in psychology: what has changed?" In: *Frontiers in Psychology*. DOI: 10.3389/fpsyg.2018.02558.

Blossfeld, H.P. (2009). "Causal analysis in population studies". In: ed. by H. Engelhardt et al. Spinger Science and Business Media. Chap. Causation as a generative process. The elaboration of an idea for the social sciences and an application to an analysis of an interdependent dynamic social system.

Blum, A., J. Hopcroft, and R. Kannan (2020). *Foundations of Data Science*. Cambridge: Cambridge University Press.

Boag, E. and K.B. Carnelley (2016). "Attachment and prejudice: The mediating role of empathy". In: *British Journal of Social Psychology* 55, pp. 337–356. DOI: 10.1111/bjso.12132.

Body, R.L. et al. (2022). *The development and psychometric properties of LIWC-22*. URL: https://www.liwc.app/.

Boker, S. M. and M. J. Wenger (2007). *Data analytic techniques for dynamical sytems*. New Jersey: Lawrence Erlbaum Associates.

Bolger, N. and J. P. Laurenceau (2013). *Intensive Longitudinal Methods*. New York: The Guilford Press.

Bolger, N., K.S. Zee, et al. (2019). "Causal processes in psychology are heterogeneous". In: *Journal of Experimental Psychology General* 148.4, pp. 601–618. DOI: 10.1037/xge0000558.

Bollen, K.A. (2019). "Model implied instrumental variables (MIIVs): An alternative orientation to structural equation modeling". In: *Multivariate Behavioral Research* 54.1, pp. 31–46. DOI: 10.1080/00273171.2018.1483224.

Borsboom, D. et al. (2021). "Theory construction methodology: a practical framework for building theories in psychology". In: *Perspectives on Psychological Science* 16.4, pp. 756–766. DOI: 10.1177/1745691620969647.

Botella, J. and J.I. Duran (2019). "A meta-analytical answer to the crisis of confidence of psychology". In: *Anales de psicologioa* 35.2, pp. 350–356. DOI: 10.6018/analesps.35.2.345291.

Bothwell, L.E., J.A. Greene, and S.H. Podolsky (2016). "Assessing the gold standard - lessons from the history of RCTs". In: *N. Engl. J. Med.* 374, pp. 2175–81. DOI: 10.1056/NEJMms1604593.

Bottou, L. et al. (2013). "Counterfactual reasoning and learning systems: the example of computational advertising". In: *Journal of Machine Learning Research* 14.

Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge: Cambridge University Press.

Bowlby, J. (1969). *Attachment and Loss: Volume 1. Attachment*. 1st. Basic Books.

Box, G. E. P. and G. M. Jenkins (1976). *Time series analysis: Forecasting and control*. San Francisco, CA: Holden-Day.

Breiman, L. (2001a). "Random forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

– (2001b). "Statistical modeling: The two cultures". In: *Statistical Science* 16.3, pp. 199–215.

Brennan, K.A., C.L. Clark, and P.R. Shaver (1998). "Attachment theory and close relationships". In: ed. by W.S. Rholes and J.A. Simpson. Guilford Press. Chap. Self-report measurement of adult attachment: An integrative overview, pp. 46–76.

Brown, T.B. et al. (2022). "Language models are few-shot learners". In: *arXiv preprint* arXiv:2005.14165v4.

Brulin, J.G. et al. (2022). "Attachment in the time of COVID-19: Insecure attachment orientations are associated with defiance of authorities' guidelines during the pandemic". In: *Journal of Social and Personal Relationships* 39.8, pp. 2528–2548. DOI: 10.1177/026540752210826.

Bryan, C.J., E. Tipton, and D.S. Yeager (2021). "Behavioural science is unlikely to change the world without a heterogeneity revolution". In: *Nature Human Behavior* 5, pp. 980–989. DOI: 10.1038/s41562-021-01143-3.

Bühler, J.L. et al. (2019). "Does Michelangelo care about age? An adult life-span perspective on the Michelangelo phenomenon". In: *Journal of Social and Personal Relationships* 36.4, pp. 1392–1412. DOI: 10.1177/0265407518766698.

Buolamwini, J. and T. Gebru (2018). "Gender Shades: Intersectional accuracy disparities in commercial gender classification". In: *Proc. of Machine Learning Research* 81, pp. 1–15.

Burkova, V.N. et al. (2021). “Predictors of anxiety in the COVID-19 pandemic from a global perspective: Data from 23 countries”. In: *Sustainability* 13. DOI: 10.3390/su13074017.

Button, K.S. (2019). “Double-dipping revisited”. In: *Nature Neuroscience* 22, pp. 688–690. DOI: 10.1038/s41593-019-0398-z.

Canevello, A. and J. Crocker (2010). “Creating good relationships: Responsiveness, relationship quality, and interpersonal goals”. In: *Journal of Personality and Social Psychology* 99.1, pp. 78–106. DOI: 10.1037/a0018186.

Cao, Y. T. and H. Daume III (2019). “Toward gender-inclusive coreference resolution”. In: *arXiv:1910.13913v2*.

Cao, Z. et al. (2018). “OpenPose: Realtime multi-person 2D pose estimation using part affinity fields”. In: *arXiv:1812.08008v1*.

Caponi, S (2013). “Quetelet, the average man and medical knowledge”. In: *Hist Cienc Saude Manguinhos Hist Cienc Saude Manguinhos* 20.3, pp. 830–847. DOI: 10.1590/S0104-59702013005000011.

Carbajal, J. et al. (2021). “The impact of COVID-19 on first responders’ resilience and attachment”. In: *Journal of Human Behavior in the Social Environment*. DOI: 10.1080/10911359.2021.1962777.

Cassidy, S.A. et al. (2019). “Failing grade: 89 percent of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly”. In: *Advances in Methods and Practices in Psychological Science* 2.3. DOI: 10.1177/2515245919858072.

Chen, H. et al. (2020). “True to the model or true to the data?” In: *arXiv:2006.16234v1*.

Chen, T. and C. Guestrin (2016). “XGBoost: A scalable tree bosting system”. In: *KDD Conference for Knowledge Discovery and Data Mining*.

Ciechanowski, P. et al. (2004). "Influence of patient attachment style on self-care and outcomes in diabetes". In: *Psychosomatic Medicine* 66.5, pp. 720–728. DOI: 10.1097/01.psy.0000138125.59122.23.

Cinelli, C., A. Forney, and J. Pearl (2022). "A crash course in good and bad controls". In: *Sociological Methods and Research*. DOI: 10.1177/00491241221099552.

Claesen, A. et al. (2019). "Preregistration: Comparing dream to reality". In: *PsyArXiv*. DOI: 10.31234/osf.io/d8wex.

Collins, N.L. and B.C. Feeney (2004). "Working models of attachment shape perceptions of social support: Evidence from experimental and observational studies". In: *Journal of Personality and Social Psychology* 87.3, pp. 363–383. DOI: 10.1037/0022-3514.87.3.363.

Colquhoun, D. (2014). "An investigation of the false discovery rate and the misinterpretation of p-values". In: *Royal Society Open Science* 1.3. DOI: 10.1098/rsos.140216.

– (2017). "The reproducibility of research and the misinterpretation of p-values". In: *Royal Society Open Science* 4.12. DOI: 10.1098/rsos.171085.

– (2019). "The false positive risk: a proposal concerning what to do about p-values". In: *The American Statistician* 73. DOI: 10.1080/00031305.2018.1529622.

Comte, A. (1976). *The foundation of sociology*. Ed. by K. Thompson. London: Nelson.

Correll, J. et al. (2020). "Avoid Cohen's 'small', 'medium', and 'large' for power analysis". In: *Trends in Cognitive Sciences* 24.3. DOI: 10.1016/j.tics.2019.12.009.

Cover, T. M. and J. A. Thomas (2006). *Elements of information theory*. New York: John Wiley and Sons Inc.

Coyle, J.R. et al. (2020). “Targeted learning: Robust statistics for reproducible research”. In: *arXiv2006.07333*.

Crutzen, R. and G.J.Y Peters (2017). “Targeting next generations to change the common practice of underpowered research”. In: *Frontiers in Psychology* 8. DOI: 10.3389/fpsyg.2017.01184.

D‘Amour, A. (2019). “On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility and alternatives”. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics* 89.

Daniels, G.S. (1952). “The “average man”?” In: *Technical Note 53-7 Wright Air Development Center, USAF*.

Davis, M.H. (1980). “A multidimensional approach to individual differences in empathy”. In: *JSAS Catalog of Selected Documents in Psychology* 10.85.

Dawid, A.P. (2008). “Beware of the DAG!” In: *NeurIPS Workshop on Causality*.

Day, L.C. and E.A. Impett (2018). “Giving when it costs: How interdependent self-construal shapes willingness to sacrifice and satisfaction with sacrifice in romantic relationships”. In: *Journal of Social and Personal Relationships* 35.5, pp. 722–742. DOI: 10.1177/0265407517694965.

Deaton, A. and N. Cartwright (2018). “Understanding and misundertstanding randomized controlled trials”. In: *Social Science and Medicine* 210, pp. 2–21. DOI: 10.1016/j.socscimed.2017.12.005.

DeDeo, S. (2020). “When science is a game”. In: *arXiv:2006.05994v2*.

Derogatis, L.R. and M. Lopez (1983). *PAIS and PAIS-SR-administration scoring and procedures manual*. Baltimore: Clinical Psychometric.

Devlin, J. et al. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint* arXiv:1810.04805v2.

Diaz, I. and M.J. van der Laan (2013). "Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems". In: *The International Journal of Biostatistics* 9.2, pp. 149–160.

Dieleman, S. (2020). *Musings on typicality*. URL: https://benanne.github.io/2020/09/01/typicality.html.

Diener, E. et al. (49). "The satisfaction with life scale". In: *Journal of Personality Assessment* 1.71-75. DOI: 10.1207/s15327752jpa4901_13.

Dowd, B. E. (2011). "Separated at birth: statistcians, social scientists, and causality in health services research". In: *Health Research and Educational Trust* 46.2, pp. 397–420. DOI: 10.1111/j.1475-6773.2010.01203.x.

Drigotas, S.M. (2002). "The Michelangelo phenomenon and personal well-being". In: *Journal of Personality* 70.1, pp. 59–77. DOI: 10.1111/1467-6494.00178.

Drigotas, S.M. et al. (1999). "Close partner as sculptor of the ideal self: Behavioral affirmation and the Michelangelo phenomenon". In: *Journal of Personality and Social Psychology* 77.2, pp. 293–323. DOI: 10.1037/0022-3514.77.2.293.

Drucker, H. (1997). "Improving regressors using boosting techniques". In: *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 107–115. DOI: https://doi.org/10.5555/645526.657132.

Duda, R. O., P. E. Hart, and D. G. Stork (2001). *Pattern Classification*. New York: John Wiley and Sons Inc.

Ellery, M. (2018). *An introduction to space robotics*. 4th. Chichester, West Sussex: Spinger-Praxis.

Ernst, A. F. and C. J. Albers (2017). "Regression assumptions in clinical psychology research practice - a systematic review of common misconceptions". In: *PeerJ* 5. DOI: 10.7717/peerj.3323.

Eronen, M.I. (2020). "Causal discovery and the problem of psychological interventions". In: *New Ideas in Psychology* 59. DOI: doi:10.1016/j.newideapsych.2020.100785.

Eronen, M.I. and L.F. Bringmann (2021). "The theory crisis in psychology: how to move forward". In: *Perspectives on Psychological Science* 16.4. DOI: doi:10.1177/1745691620970586.

Eronen, M.I. and J.W. Romeijn (2020). "Philosophy of science and the formalization of psychological theory". In: *Theory and Psychology* 30.6, pp. 786–799. DOI: 10.1177/0959354320969876.

European Union (2016). "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR)". In: *Official Journal of the European Union* 59.

Fang, J. and M.H. Alderman (2000). "Serum uric acid and cardiovascular mortality: the NHANES I epidemiologic follow-up study". In: *JAMA* 283.18, pp. 2404–2410.

Feeney, B.C. (2004). "A secure base: Responsive support of goal strivings and exploration in adult intimate relationships". In: *Journal of Personality and Social Psychology* 87.5, pp. 6331–648. DOI: 10.1037/0022-3514.87.5.631.

Feeney, B.C. and N.L. Collins (2015). "A new look at social support: A theoretical perspective on thriving through relationships". In: *Personality and Social Psychology Review* 19.2, pp. 113–147. DOI: 10.1177/1088868314544222.

Feeney, B.C. and R.L. Thrush (2010). "Relationship influences on exploration in adulthood: The characteristics and function of a secure base". In: *Journal of Personality and Social Psychology* 98.1, pp. 57–76. DOI: 10.1037/a0016961.

Fiedler, K. (2017). "What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing". In: *Perspectives on Psychological Science* 12.1. DOI: 10.1177/1745691616654458.

Field, A. (2009). *Discovering statistics using SPSS*. 3rd. Los Angeles: Sage.

Finkel, E. (2020a). "Northwestern NSF1 (V1)". In: *UNC Dataverse*. DOI: 10.15139/S3/ENAX3L.

– (2020b). "Northwestern NSF2 (V1)". In: *UNC Dataverse*. DOI: 10.15139/S3/AKSZOS.

Fisher, A.J., J.D. Medaglia, and B.F. Jeronimus (2018). "Lack of group-to-individual generalizability is a threat to human subjects research". In: *PNAS* 115.27. DOI: 10.1073/pnas.1711978115.

Fitzsimons, G.M. and E.J. Finkel (2018). "Transactive-goal-dynamics theory: A discipline-wide perspective". In: *Current Directions in Psychological Science* 27.5, pp. 332–338. DOI: 10.1177/0963721417754199.

Flake, J. and E. Fried (2020). "Measurement schmeasurement: questionable measurement practices and how to avoid them". In: *Advances in Methods and Practices in Psychological Science*. DOI: 10.1177/2515245920952393.

Florian, V., M. Mikulincer, and I. Bucholtz (1995). "Effects of adult attachment style on the perception and search for social support". In: *The Journal of Psychology* 129.6, pp. 665–676. DOI: 10.1080/00223980.1995.9914937.

Foucault, M. (1984). *Madness and Civilization*. Ed. by P. Rabinow. London: Penguin Books.

Fraley, R.C. and P.R. Shaver (1998). "Airport separations: A naturalistic study of adult attachment dynamics in separating couples". In: *Journal of Personality and Social Psychology* 75.5, pp. 1198–1212. DOI: 10.1037/0022-3514.75.5.1198.

Fraley, R.C., N.G. Waller, and K.A. Brennan (2000). "An item response theory analysis of self-report measures of adult attachment." In: *Journal of Personality and Social Psychology* 78.2, pp. 350–365. DOI: 10.1037//0022-3514.78.2.350.

Freedman, D. (1985). "Cohort Analysis in Social Research." In: ed. by W.M. Mason and S.E. Fienberg. New York: Springer. Chap. Statistics and the Scientific Method.

Frieden, T.R. (2017). "Evidence for health decision making - beyond randomized, controlled trials". In: *N. Engl. J. Med.* 377, pp. 465–475. DOI: 10.1056/NEJMra1614394.

Fu, K.S., R.C. Gonzalez, and C.S.G. Lee (2018). *Robotics: Control, Sensing, Vision, and Intelligence*. Ed. by Herbert Freeman. New York: McGraw Hill Eductation.

Gao, S., G.V. Steeg, and A. Galstyan (2015). "Efficient estimation of mutual information for strongly dependent variables". In: *AISTATS*.

Gelman, A. (2014). "Correlation does not even imply correlation". In: *Statistical Modeling, Causal Inference, and Social Science (BLOG)*.

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gelman, A., J. Hill, and A. Vehtari (2021). *Regression and other stories*. Cambridge: Cambridge University Press.

Gelman, A. and E. Loken (2013). *The garden of forking paths: why multiple comparisons can be a problem even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time*. URL: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Gere, J. and U. Schimmack (2013). "When romantic partners' goals conflict: Effects on relationship quality and subjective well-being". In: *Journal of Happiness Studies* 14.1, pp. 37–49. DOI: 10.1007/s10902-011-9314-2.

Gernsbacher, M.A. (2019). "Three ways to make replication mainstream". In: *Behav. Brain. Sci.* 41. DOI: 10.1017/S0140525X1800064X.

Gigerenzer, G. (2004). "Mindless Statistics". In: *Journal of Socio-Economics* 33, pp. 587–606. DOI: 10.1016/j.socec.2004.09.033.

– (2018). "Statistical rituals: the replication delusion and how we got there". In: *Advances in Methods and Practices in Psychological Science* 1.2, pp. 198–218. DOI: 10.1177/2515245918771329.

Gische, C., S.G. West, and M.C. Voelkle (2020). "Forecasting causal effects of interventions versus predicting future outcomes". In: *Structural Equation Modeling: A Multidisciplinary Journal*. DOI: 10.1080/10705511.2020.1780598.

Glorot, X. and Y. Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.

Glymour, C. (1998). "What Went Wrong? Reflections on Science by Observation and the Bell Curve". In: *The University of Chicago Press on behalf of the Philosophy of Science Association* 65.1. DOI: 10.1086/392624.

– (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. MIT Press.

Glymour, C., K. Zhang, and P. Spirtes (2019). "Review of causal discovery methods based on graphical models". In: *Frontiers in Genetics* 10.

Godfrey-Smith, P. (2003). *Theory and reality: An introduction to the philosophy of science*. Chicago: University of Chicago Press.

Goldberg, L.R. (1999). “Personality Psychology in Europe”. In: ed. by I. Mervielde et al. Tilburg, The Netherlands: Tilburg University Press. Chap. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models, pp. 7–28.

Goldberg, L.R. et al. (2006). “The international personality item pool and the future of public-domain personality measures”. In: *Journal of Research in Personality* 40, pp. 84–96.

Goldstein, B.A., E.C. Polley, and F.B.S. Briggs (2011). “Random forests for genetic association studies”. In: *Stat. Appl. Genet. Mol. Biol.* 10.1, p. 32.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. Cambridge, Massachusetts: MIT Press.

Goodfellow, I. J. et al. (2014). “Generative Adversarial Nets”. In: *arXiv:1406.2661*.

Gottman, J. M. (1979). “Detecting Cyclicity in Social Interaction”. In: *Psychological Bulletin* 86.2, pp. 338–348. DOI: 10.1037/0033-2909.86.2.338.

Goudet, O. et al. (2019). “Cause effect pairs in machine learning”. In: ed. by I. Guyon, A. Statnikov, and B. Batu. Springer. Chap. Learning bivariate functional cusal model.

Greer, F. and J. Liu (2016). “Pinciples and methods of test construction: Standards and recent advances”. In: ed. by K. Schweizer and C. DiStefano. Hogrefe Publishing. Chap. Creating short forms and screening measures, pp. 272–287.

Grosz, M.P., J.M. Rohrer, and F. Thoemmes (2020). “The taboo against explicit causal inference in nonexperimental psychology”. In: *Perspectives on Psychological Science*, pp. 1–13. DOI: 10.1177/1745691620921521.

Grubbs, F.E. (1969). “Procedures for detecting outlying observations in samples”. In: *Technometrics* 11.1, pp. 1–21. DOI: 10.1080/00401706.1969.10490657.

Gultchin, L. et al. (2020). "Differentiable causal backdoor discovery". In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics* 108.

Gundersen, G. (2020). *Entropy of the Gaussian*. URL: https://gregorygundersen.com/blog/2020/09/01/gaussian-entropy/.

Hacking, I. (1990). *The taming of chance*. Cambridge: Cambridge University Press.

Hamilton, J. D. (1994). *Time Series Analysis*. New Jersey: Princeton University Press.

Hampel, F. R. (1974). "The influence curve and its role in robust estimation". In: *Journal of the American Statistical Association* 69.346, pp. 383–393.

Hardt, M., E. Price, and N. Srebro (2016). "Equality of opportunity in supervised learning". In: *arXiv:1610.02413v1*.

Harris, M.A. and U. Orth (2020). "The link between self-esteem and social relationships: A meta-analysis of longitudinal studies". In: *Journal of Personality and Social Psychology* 119.6, pp. 1459–1477. DOI: 10.1037/pspp0000265.

Haslbeck, J.M.B. et al. (2021). "Modeling psychopathology: From data models to formal theories". In: *Psychological Methods*. DOI: doi:10.1037/met0000303.

Haufe, S. et al. (2014). "On the interpretation of weight vectors of linear models in multivariate neuroimaging". In: *NeuroImage* 87.15, pp. 96–110. DOI: 10.1016/j.neuroimage.2013.10.067.

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. Delhi: Pearson/Prentice Hall.

Heesen, R. and L.K. Bright (2020). "Is peer review a good idea?" In: *The British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axz029.

Heinze-Deml, C., M.H. Maathuis, and N. Meinshausen (2018). "Causal structure learning". In: *Annual Review of Statistics and Its Application* 5.

Hernan, M. (2018a). “The c-word: scientific euphemisms do not improve causal inference from observational data”. In: *American Journal of Public Health* 108.5, pp. 625–626. DOI: 10.2105/AJPH.2018.304337.

– (2018b). “The c-word: the more we discuss it, the less dirty it sounds”. In: *American Journal of Public Health* 108.5, pp. 625–626. DOI: 10.2105/AJPH.2018.304392.

Hernan, M.A. (2022). “Causal analyses of existing databases: no power calculations required”. In: *Journal of Clinical Epidemiology* 144, pp. 203–205. DOI: 10.1016/j.jclinepi.2020.08.028.

Heyman, R.E. and A.M.S. Slep (2001). “The hazards of predicting divorce without crossvalidation”. In: *J Marriage Fam.* 63.2, pp. 473–479. DOI: 10.1111/j.1741-3737.2001.00473.x.

Higgins, E.T. et al. (2001). “Achievement orientations from subjective histories of success: Promotion pride versus prevention pride”. In: *European Journal of Social Psychology* 31.1, pp. 3–23. DOI: 10.1002/ejsp.27.

Hilpert, P. et al. (2019). “What Can Be Learned From Couple Research: Examining Emotional Co-Regulation Processes in Face-to-Face Interactions”. In: *Journal of Counseling Psychology*.

Hines, O. et al. (2021). “Demystifying statistical learning based on efficient influence functions”. In: *arXiv preprint* arXiv:2107.00681.

Hinz, A. et al. (2017). “Frequency of somatic symptoms in the general population: Normative values for the Patient Health Questionnaire-15 (PHQ-15).” In: *Journal of Psychosomatic Research* 96, pp. 27–31. DOI: 10.1016/j.jpsychores.2016.12.017.

Hornik, K., M. Stinchcombe, and H. White (1989). “Multilayer feedforward networks are universal approximators”. In: *Neural Networks* 2, pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

Howard, A. and J. Borenstein (2018). "The ugly truth about ourselves and our robot creations: the problem of bias and social inequity." In: *Science and engineering ethics* 24.5, pp. 1521–1536. DOI: 10.1007/s11948-017-9975-2.

Hoyle, R.H. and A.T. Panter (1995). "Structural equation modelling: COncepts, issues, and applications". In: ed. by R.H. Hoyle. Thousand Oaks, CA: SAGE Publications. Chap. Writing about structural equation models, pp. 158–176.

Huang, Y. and M. Valtorta (2006). "Pearl's calculus of intervention is complete". In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* arXiv:1206.6831, pp. 217–224. DOI: 10.5555/3020419.3020446.

Huckins, J.F. et al. (2020). "Mental health and behavior of college students during the early phases of the COVID-19 pandemic: Longitudinal smartphone and ecological momentary assessment study". In: *Journal of Medical Internet Research* 22.e20185, pp. 1–13. DOI: 10.2196/20185.

Hughes, M.E. et al. (2004). "A short scale for measuring loneliness in large surveys: Results from two population-based studies". In: *Research on Aging* 26.6, pp. 655–672. DOI: 10.1177/0164027504268574.

Hullman, J. et al. (2022). "The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning". In: *arXiv preprint* arXiv:2203.06498.

Hünermund, P. and E. Bareinboim (2021). "Causal inference and data fusion in econometrics". In: *arXiv preprint* arXiv:1912.09104v3.

Imbens, G.W. and D.B. Rubin (2015). *Causal inference for statistics, social, and biomedical sciences. An Introduction.* New York: Cambridge University Press.

Impett, E.A. and A.M. Gordon (2010). "Why do people sacrifice to approach rewards versus to avoid costs? Insights from attachment theory." In: *Personal Relationships* 17.2. DOI: 10.1111/j.1475-6811.2010.01277.x.

Ioannidis, J. P. A. (2012). "Why science is not necessarily self-correcting". In: *Perspectives on Psychological Science* 7.6, pp. 645–654. DOI: 10.1177/174569161246405.

Jakubiak, B.K. and B.C. Feeney (2016). "Daily goal progress is facilitated by spousal support and promotes psychological, physical, and relational well-being throughout adulthood". In: *Journal of Personality and Social Psychology* 111.3, pp. 317–340. DOI: 10.1037/pspi0000062.

Jakubiak, B.K., B.C. Feeney, and R.A. Ferrer (2020). "Benefits of daily support visibility versus invisibility across the adult life span." In: *Journal of Personality and Social Psychology* 118.5, pp. 1018–1043. DOI: 10.1037/pspi0000203.

Janitza, S. and R. Hornung (2018). "On the overestimation of random forest's out-of-bag-error". In: *PLoS One* 13.8. DOI: 10.1371/journal.pone.0201904.

Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge, Massachusetts: MIT Press.

Jiménez-Luna, J., F. Grisoni, and G. Schneider (2020). "Drug discovery with explainable artificial intelligence". In: *Nature Machine Intelligence* 2, pp. 573–584.

Joel, S., P.W. Eastwick, C.J. Allison, et al. (2020). "Machine learning uncovers the most robust self-report predictors of relationships quality across 43 longitudinal couples studies". In: *PNAS* 117.32, pp. 19061–71.

Joel, S., P.W. Eastwick, and E.J. Finkel (2017). "Is romantic desire predictable? Machine learning applied to initial romantic attraction". In: *Psychological Science* 28.10. DOI: 10.1177/0956797617714580.

Jones, D.S. and S.H. Podolsky (2015). "The history and fate of the gold standard". In: *Lancet* 385.1502-3. DOI: 10.1016/S0140-6736(15)60742-5.

Jonsson, K. et al. (2000). "Learning support vectors for face verification and recognition". In: *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*.

Jostmann, N.B., D. Lakens, and T.W. Schubert (2016). "A short history of the weight-importance effect and a recommendation for pre-testing: commentary on Ebersole et al. (2016)". In: *JESP* 67. DOI: 10.1016/j.jesp.2015.12.001.

Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing*. New Jersey: Pearson Prentice Hall.

Kalainathan, D. et al. (2020). "Structural agnostic modeling: Adversarial learning of causal graphs". In: *arXiv:1803.04929v3*.

Kassraian-Fard, P. et al. (2016). "Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example". In: *Frontiers in Psychology* 7.177. DOI: 10.3389/fpsyt.2016.00177.

Kelley, H.H. and J.W. Thibaut (1978). *Interpersonal relations: A theory of interdependence*. Wiley-Interscience.

Kennedy, E.H. (2020). "Optimal doubly robust estimation of hetereogeneous causal effects". In: *arXiv preprint* arXiv:2004.14497v2.

Khalilia, M., S. Chakraborty, and M. Popescu (2011). "Predicting disease risks from highly imbalanced data using random forest". In: *BMC Med. Info. Dec. Making* 11.51.

Kilbertus, N. et al. (2017). "Avoiding discrimination through causal reasoning". In: *31st Conference on Neural Information Processing Systems*.

King, G. (1986). "How not to lie with statistics: Avoiding common mistakes in quantitative political science". In: *American Journal of Political Science* 30.3, pp. 666–687.

Kingma, D. P. and J. L. Ba (2017). "Adam: a method for stochastic optimization". In: *arXiv:1412.6980v9*.

Kinney, J.B. and G.S. Atwal (2014). "Equitability, mutual information, and the maximal information coefficient". In: *PNAS* 111.9, pp. 3354–3359. DOI: 10.1073/pnas.1309933111.

Kline, R.B. (2005). *Principles and practice of structural equation modeling*. Guilford Press.

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, Massachusetts: MIT Press.

Kraskov, A., H. Stogbauer, and P. Grassberger (2004). "Estimating mutual information". In: *Physical Review E* 69. DOI: 10.1103/PhysRevE.69.066138.

Krasuska, M. et al. (2018). "The role of adult attachment orientation and coping in psychological adjustment to living with skin conditions." In: *British Journal of Dermatology* 178.6, pp. 1396–1403. DOI: 10.1111/bjd.16268.

Kreif, N. and K. DiazOrdaz (2019). "Machine learning in policy evaluation: new tools for causal inference". In: *arXiv:1903.00402v1*.

Kriegeskorte, N. et al. (2009). "Circular analysis in systems neuroscience: the danges of double dipping". In: *Nat. Neurosci* 12. DOI: 10.1038/nn.2303.

Krishnan, R. G., U. Shalit, and D. Sontag (2017). "Structured inference networks for nonlinear state space models". In: *Association for the Advancement of Artificial Intelligence*.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "ImageNet classification with deep convolutional neural networks". In: *Proceeding NIPS Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097–1105.

Kroc, E. and O.L.O. Astivia (2021). "The importance of thinking multivarietly when selecting subscale cutoff scores". In: *Educational and Psychological Measurement* online first. DOI: 10.1177/00131644211023569.

Kroenke, K., R.L. Spitzer, and J.B. Williams (2002). "The PHQ-15: validity of a new measure for evaluating the severity of somatic symptoms". In: *Psychosomatic Medicine* 64.2, pp. 258–266. DOI: 10.1097/00006842-200203000-00008.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kumar, E.I. et al. (2020). "Problems with Shapley-value-based explanations as feature importance measures". In: *Proceedings of the 37th International Conference on Machine Learning*.

Kumashiro, M., E.J. Finkel, and C.E. Rusbult (2002). "Self-respect and pro-relationship behavior in marital relationships." In: *Journal of Personality* 70.6, pp. 1009–1050. DOI: 10.1111/1467-6494.05030.

Kupferschmidt, K. (2018). "More and more scientists are preregistering their studies. Should you?" In: *Science: News*. DOI: 10.1126/science.aav4786.

Lakens, D. (2022). "Sample size justification". In: *Collabra: Psychology*. DOI: 10.1525/collabra.33267.

Lakens, D. and E.R.K. Evers (2014). "Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies". In: *Perspectives on Psychological Science* 9.3, pp. 278–292. DOI: 10.1177/1745691614528520.

Lakens, D., J. Hilgard, and J. Staaks (2016). "On the reproducibility of meta-analyses: six practice recommendations". In: *BMC Psychology* 4.24. DOI: 10.1186/s40359-016-0126-3.

Lantagne, A. and W. Furman (2017). "Romantic Relationship Development: The Interplay Between Age and Relationship Length". In: *Developmental Psychology* 53.9, pp. 1738–1749. DOI: 10.1037/dev0000363.

Launer, L.J. et al. (1994). “Body mass index, weight change, and risk of mobility disability in middle-aged and older women: the epidemiologic follow-up study of NHANES I”. In: *JAMA* 271.14, pp. 1093–1098.

Lavrakas, P.J. (2008). “Encyclopedia of survey research methods”. In: vol. 1. Thousand Oaks, CA: SAGE Publications. Chap. Respondent fatigue. DOI: 10.4135/9781412963947.

Leys, C., M. Delacre, et al. (2019). “How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration”. In: *International Review of Social Psychology* 32.1, pp. 1–10. DOI: 10.5334/irsp.289.

Leys, C., O. Klein, et al. (2018). “Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance”. In: *Journal of Experimental Social Psychology* 74, pp. 150–156. DOI: 10.1016/j.jesp.2017.09.011.

Leys, C., C. Ley, et al. (2013). “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the mean”. In: *Journal of Experimental Social Psychology* 49.4, pp. 764–766. DOI: 10.1016/j.jesp.2013.03.013.

Li, H. et al. (2022). “Evaluating the robustness of targeted maximum likelihood estimators via realistic simulations in nutrition intervention trials”. In: *Statistics in Medicine* 41.2. DOI: https://doi.org/10.1002/sim.9348.

Li, X. et al. (2019). “Outlier detection based on robust Mahalanobis distance and its application”. In: *Open Journal of Statistics* 9.1. DOI: 10.4236/ojs.2019.91002.

Lindsay, S. (2020). *Apology re Clark et al.* URL: https://onlineacademiccommunity.uvic.ca/lindsaylab/2020/06/26/apology-re-clark-et-al/.

Liu, H. et al. (2019). “Does gender matter? Towards fairness in dialogue systems.” In: *arXiv:1910.10486v1*.

Locatello, F. et al. (2019). “On the fairness of disentangled representations”. In: *arXiv:1905.13662v1*.

Lockwood, P., C.H. Jordan, and Z. Kunda (2002). "Motivation by positive or negative role models: Regulatory focus determines who will best inspire us". In: *Journal of Personality and Social Psychology* 4.854-864. DOI: 10.1037/0022-3514.83.4.854.

Loehlin, J.C. and A.A. Beaujean (2017). *Latent Variable Models: An introduction to factor, path, and structural equation analysis*. New York: Routledge Taylor and Francis.

Lohmann, G. et al. (2012). "Critical comments on dynamic causal modelling". In: *Neuroimage* 59.3, pp. 2322–9. DOI: 10.1016/j.neuroimage.2011.09.025.

Louizos, C., U. Shalit, et al. (2017). "Causal effect inference with deep latent-variable models". In: *31st Conference on Neural Information Processing Systems*.

Louizos, C., K. Swersky, et al. (2017). "The variational fair autoencoder". In: *arXiv:1511.00830*.

Lozano, E.B. and R.C. Fraley (2021). "Put your mask on first to help others: Attachment and sentinel behavior during the COVID-19 pandemic". In: *Personality and Individual Differences* 171. DOI: 10.1016/j.paid.2020.110487.

Lundberg, S.M., G. Erion, et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2, pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

Lundberg, S.M., G.G. Erion, and S-I. Lee (2017). "Consistent individualized feature attribution for tree ensembles". In: *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia*.

Lundberg, S.M. and S-I. Lee (2017). "A unified approach to interpreting model predictions". In: *31st Conference on Neural Information Processing Systems*.

Luque-Fernandez, M.A. et al. (2018). "Targeted maximum likelihood estimation for a binary treatment: A tutorial". In: *Statistics in Medicine* 37.16, pp. 2530–2546. DOI: 10.1002/sim.7628.

MacKay, D. J. C. (1992). “A practical Bayesian framework for backprop networks”. In: *Dawin College Address*.

– (2018). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Maclaren, O.J. and R. Nicholson (2020). “What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems”. In: *arXiv preprint* arXiv:1904.02826v4.

Mahalanobis, P.C. (1930). “On tests and measures of group divergence”. In: *Journal and Proceedings of Asiatic Society of Bengal* 26, pp. 541–588. DOI: http://hdl.handle.net/10263/1639.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos (2020). “The M4 competition: 100,000 time series and 61 forecasting methods”. In: *International Journal of Forecasting* 36.1, pp. 54–74. DOI: 10.1016/j.ijforecast.2019.04.014.

Marsman, M. et al. (2017). “A Bayesian bird’s eye view of ‘replications of important results in social psychology’”. In: *R. Soc. open sci.* 4. DOI: 10.1098/rsos.160426.

Martin, A.M.L., R.L. Paetzold, and W.S. Rholes (2010). “Adult attachment and exploration: Linking attachment style to motivation and perceptions of support in adult exploration”. In: *Basic and Applied Social Psychology* 32.2, pp. 196–205. DOI: 10.1080/01973531003738452.

Martin, G.N. and R.M. Clarke (2017). “Are psychology journals anti-replication? A snapshot of editorial practices”. In: *Frontiers in Psychology* 8. DOI: 10.3389/fpsyg.2017.00523.

Maruyama, G.M. (1998). *Basics of Structural Equation Modeling*. Thousand Oaks, CA: SAGE Publications.

Maxwell, S.E. (2004). “The persistence of underpowered studies in psychological research: Causes, consequences, and remedies”. In: *Psychological Methods* 9.2, pp. 147–163. DOI: 10.1037/1082-989X.9.2.147.

Mayo, D. (2013). "Some surprising facts about (the problem of) surprising facts". In: *Studies in the History and Philosophy of Science*. DOI: 10.1016/j.shpsa.2013.10.005.

Mazza, C. et al. (2021). "The COVID-19 outbreak and psychological distress in healthcare workers: The role of personality traits, attachment styles, and sociodemographic factors". In: *Sustainability* 13. DOI: 10.3390/su13094992.

McBride, O. et al. (2021). "Monitoring the psychological, social, and economic impact of the COVID-19 pandemic in the population: Context, design and conduct of the longitudinal COVID-19 psychological research consortium (C19PRC) study". In: *International Journal of Methods in Psychiatric Research* 30.1. DOI: 10.1002/mpr.1861.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. New York: Chapman and Hall. DOI: 10.1201/9781315372495.

McShane, B.B. et al. (2019). "Abandon statistical significance". In: *The American Statistician* 73, pp. 235–245. DOI: 10.1080/00031305.2018.1527253.

Meehl, P.E. (1990). "Why summaries of research on psychological theories are often uninterpretable". In: *Psychological Reports* 66, pp. 195–244. DOI: 10.2466/pr0.1990.66.1.195.

Meredith, P.J., J. Strong, and J.A. Feeney (2005). "Evidence of a relationship between adult attachment variables and appraisals of chronic pain". In: *Pain Research and Management* 10.4, pp. 191–200. DOI: 10.1155/2005/745650.

Micceri, T. (1989). "The unicorn, the normal curve, and other improbable creatures". In: *Psychological Bulletin* 105, pp. 156–166. DOI: 10.1037/0033-2909.105.1.156.

Michell, J. (2016). "Normal science, pathological science and psychometrics". In: *Theoretical Psychology* 10.5. DOI: 10.1177/0959354300105004.

Mikulincer, M. and P.R. Shaver (2007). “Handbook of motivation scienc”. In: ed. by James Y. Shah and W.L. Gardner. Guilford Press. Chap. Contributions of attachment theory and research on motivation science, pp. 201–216.

– (2009). “An attachment and behavioral systems perspective on social support”. In: *Journal of Social and Personal Relationships* 26.1, pp. 7–19. DOI: 10.1177/0265407509105518.

– (2016). “Attachment in adulthood: Structure, dynamics, and change”. In: 2nd. New York: Guilford Press. Chap. Attachment bases of psychopathology, pp. 395–442.

Mikulincer, M., P.R. Shaver, et al. (2005). “Attachment, caregiving, and altruism: Boosting attachment security increases compassion and helping”. In: *Journal of Personality and Social Psychology* 89.5, pp. 817–839. DOI: 10.1037/0022-3514.89.5.817.

Miller, G. (1956). “The magical number seven, plus or minus two: Some limits on our capacity for processing information”. In: *The Psychological Review* 63.2, pp. 81–97. DOI: 10.1037/h0043158.

Misztal, B.A. (2002). “Rethinking the concept of normality: The criticism of Comte’s theory of normal existence”. In: *Polish Sociological Review* 138, pp. 189–202.

Mitchell, M.W. (2011). “Bias of the random forest out-of-bag (OOB) error for certain input parameters”. In: *Open Journal of Statistics* 1.3. DOI: 10.4236/ojs.2011.13024.

Moccia, L. et al. (2020). “Affective temperament, attachment style, and the psychological impact of the COVID-19 outbreak: an early report on the Italian general population”. In: *Brain, Behavior, and Immunity* 87, pp. 75–79. DOI: 10.1016/j.bbi.2020.04.048..

Modis, T. (2007). “The normal, the natural, and the harmonic”. In: *Technological Forecasting and Social Change* 74, pp. 391–494. DOI: 10.1016/j.techfore.2006.07.003.

Mooij, J. M. et al. (2010). “Probabilistic latent variable models for distinguishing between cause and effect”. In: *NIPS*, pp. 1687–1695.

Mooij, J.M. et al. (2016). "Distinguishing cause from effect using observational data: methods and benchmarks". In: *Journal of Machine Learning Research* 17.32, pp. 1–102.

Morgan, S.L. and C. Winship (2015). *Counterfactuals and causal inference: Methods and principles for social research*. Cornwall: Cambridge University Press.

Moyer, D. et al. (2018). "Invariant representations without adversarial training". In: *NeurIPS*.

Mun, E-Y and F. Geng (2019). "Predicting post-experimenta fatigue among healthy young adults: random forest regression analysis". In: *Psychol Test Assess Model* 61.4. DOI: PMCID: PMC7007183.

Murphy, K. P (2012). *Machine Learning: A probabilistic Perspective*. Cambridge, Massachusetts: MIT Press.

Murphy, R. R. (2000). *Introduction to AI robotics*. Cambridge, Massachusetts: MIT Press.

Muthukrishna, M. and J. Henrich (2019a). "A problem in theory". In: *Nature Human Behavior* 3, pp. 221–229. DOI: 10.1038/s41562-018-0522-1.

– (2019b). "A problem in theory". In: *Nature Human Behavior* 3, pp. 221–229. DOI: 10.1038/s41562-018-0522-1.

Myers, S. (2013). "Normality in Analytic Psychology". In: *Behavioural Sciences (Basel, Switzerland)* 3.4, pp. 647–661. DOI: 10.3390/bs3040647.

Navarro, D.J. (2021). "If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology". In: *Perspectives on Psychological Science* 16.4. DOI: 10.1177/1745691620974769.

Nielsen, D.G. et al. (2020). "Estimation of Optimal Values for Lumped Elements in a Finite Element - Lumped Parameter Model of a Loudspeaker". In: *Journal of Computational Acoustics* 28.2. DOI: 10.1142/S2591728520500127.

Nozek, B.A. et al. (2018). “The preregistration revolution”. In: *Proceedings of the National Academy of Sciences* 115.11, pp. 2600–2606. DOI: doi:10.1073/Pnas.1708274114.

Nuijten, M.B. et al. (2016). “The prevalence of statistical reporting errors in psychology (1985-2013)”. In: *Behavior Research Methods* 48, pp. 1205–1226. DOI: 10.3758/s13428-015-0664-2.

Oberauer, K. and S. Lewandowsky (2019). “Addressing the theory crisis in psychology”. In: *Psychonomic Bulletin and Review* 26, pp. 1596–1618. DOI: 10.3758/s13423-019-01645-2.

Onwuegbuzie, A.J. and L.G. Daniel (1999). “Uses and misuses of the correlation coefficient”. In: *MSERA* 9.1, pp. 73–90.

Orben, A. and D. Lakens (2020). “Crud (re)defined”. In: *Advances in Methods and Practices in Psychological Science* 3.2, pp. 238–247. DOI: 10.1177/2515245920917961.

Orben, A. and A.K. Przybylski (2019). “The association between adolescent well-being and digital technology use”. In: *Nature Human Behavior* 3, pp. 173–182. DOI: 10.1038/s41562-018-0506-1.

Orlenko, A. and J.H. Moore (2021). “A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions”. In: *BioData Mining* 14.9.

Paulhus, D.L. (1984). “Two-component models of socially desirable responding”. In: *Journal of Personality and Social Psychology* 46.3, pp. 598–609. DOI: 10.1037/0022-3514.46.3.598.

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.

– (2012). “On a Class of Bias-Amplifying Variables that Endanger Effect Estimates”. In: *arXiv:1203.3503*.

– (2018). “Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright.” In: *Social Science and Medicine*.

Pearl, J., M. Glymour, and N.P. Jewell (2016). *Causal inference in statistics: A primer*. Wiley.

Pearl, J. and D. Mackenzie (2018). *The book of why*. Penguin Books.

Pedregosa, F. et al. (2011). “Scikit-learn: Machine learning in Python”. In: *JMLR* 12, pp. 2825–2830.

Peters, J., D. Janzing, and B. Scholkopf (2017). *Elements of Causal Inference*. Cambridge, Massachusetts: MIT Press.

Peters, O. and M.J. Werner (2017). “A recipe for irreproducible results”. In: *arXiv:1706.07773v1*.

Petersen, M. et al. (2017). “Association of implementation of a universal testing and treatment intervention with HIV diagnosis, receipt of antiretroviral therapy, and viral suppression in East Africa”. In: *Journal of American Medical Association* 317.21, pp. 2196–2206. DOI: 10.1001/jama.2017.5705.

Pierce, M. et al. (2020). “Mental health before and during the COVID-19 pandemic: A longitudinal probability sample survey of the UK population”. In: *The Lancet Psychiatry* 7.10, pp. 883–892. DOI: 10.1016/S2215-0366(20)30308-4.

Pietronmonaco, P.R., B. Uchino, and D. Schetter (2013). “Close relationship processes and health: implications of attachment theory for health and disease”. In: *Health Psychology* 32.5, pp. 499–513. DOI: 10.1037/a0029349.

Planck, M. (1949). *Scientific autobiography and other papers*. Williams and Norgate.

Platt, J.C. (1999). “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods”. In: *Advances in Large Margin Classifiers*, pp. 61–74.

Plonsky, O. et al. (2016). “Psychological forest: predicting human behavior”. In: *SSRN Electronic Journal*. DOI: doi:10.2139/ssrn.2816450.

Popper, K. (1959). *The logic of scientific discovery*. Chicago: Basic Books.

Probst, P., M.N. Wright, and A.-L. Boulesteix (2018). “Hyperparameters and tuning strategies for random forest”. In: *Wires Data Mining and Knowledge Discovery*. DOI: 10.1002/widm.1301.

Quetelet, A. (1835). *Sur l’homme et le developpement de ses facultes*. Paris: Fayard.

Rabiner, L. R. and R. W. Schafer (1978). *Digital processing of speech signals*. Ed. by A. Oppenheim. Prentice-Hall.

Raghunathan, T.E. and J.E. Grizzle (1995). “A split questionnaire survey design”. In: *Journal of the American Statistical Association* 90.429, pp. 54–63. DOI: doi:10.2307/2291129.

Rahbar, H. et al. (2020). “The value of patient and tumor factors in predictive preoperative breast MRI outcomes”. In: *Radiology: imaging cancer* 2.4. DOI: 10.1148/rycan.2020190099.

Randall, A.K. et al. (2021). “Coping with global uncertainty: Perceptions of COVID-19 psychological distress, relationship quality, and dyadic coping for romantic partners across 27 countries”. In: *Journal of Social and Personal Relationships* 39.1. DOI: 10.1177/02654075211034236.

Raudenbush, S.W. and A.S. Bryk (2002). *Hierarchical Linear Models: Applications and data analysis methods*. SAGE Publications.

Reblin, M. and B.N. Uchino (2008). “Social and emotional support and its implication for health”. In: *Current Opinion in Psychiatry* 21.2, pp. 201–205. DOI: 10.1097/YCO.0b013e3282f3ad89.

Reis, H.T. (2007). "Steps toward the ripening of relationship science". In: *Personal Relationships* 14.1, pp. 1–23. DOI: 10.1111/j.1475-6811.2006.00139.x.

Reis, H.T., M.S. Clark, and J.G. Holmes (2004). "Handbook of closeness and intimacy". In: ed. by D.J. Mashek and A.P. Aron. Lawrence Eribaum Associates. Chap. Perceived partner responsiveness as an organizing construct in the study of intimacy and closeness, pp. 201–225.

Rempel, J.K., J.G. Holmes, and M.P. Zanna (1985). "Trust in close relationships". In: *Journal of Personality and Social Psychology* 49.1, pp. 95–112. DOI: 10.1037/0022-3514.49.1.95.

Reynolds, J.J. (2021). "Let's talk about stats: Revising our approach to teaching statistics in psychology". In: *Psychological Reports* 0.0. DOI: doi:10.1177/00332941211043447.

Righetti, F. and M. Kumashiro (2012). "Interpersonal goal support in achieving ideals and oughts: The role of dispositional regulatory focus". In: *Personality and Individual Differences* 53.5, pp. 650–654. DOI: 10.1016/J.PAID.2012.05.019.

Righetti, F., C. Rusbult, and C. Finkenauer (2010). "Regulatory focus and the Michelangelo Phenomenon: How close partners promote one another's ideal selves." In: *Journal of Experimental Social Psychology* 46.6, pp. 972–985. DOI: 10.1016/j.jesp.2010.06.001.

Robinaugh, D.J., J.M.B. Haslbeck, et al. (2019). *Advancing the network theory of metal disorders: A computational model of panic disorder*. preprint.

Robinaugh, D.J., R.H. Hoekstra, et al. (2020). "The network approach to psychopathology: a review of the literature 2008-2018 and an agenda for future research". In: *Psychological Medicine* 50.3, pp. 353–366. DOI: 10.1017/S0033291719003404.

Rohrer, J.M. (2018). "Thinking clearly about correlations and causation: Graphical causal models for observational data". In: *Advances in Methods and Practices in Psychological Science*. DOI: 10.1177/2515245917745629.

Ropovik, I. (2015). “A cautionary note on testing latent variable models”. In: *Frontiers in Psychology* 6. DOI: 10.3389/fpsyg.2015.01715.

Rose, A. (2010). “Are face-detection cameras racist?” In: *Time Business*.

Rosenblatt, F. (1958). “The perceptron: A probabilistic model for information storage and organization in the brain”. In: *Psychological Review* 65, pp. 386–408.

Rosseel, Y. (2012). “An R package for structural equation modeling”. In: *Journal of Statistical Software* 48.2, pp. 1–36. DOI: 10.18637/jss.v048.i02.

Rubin, D. B. (2005). “Causal inference using potential outcomes: Design, modeling, decisions.” In: *Journal of the American Statistical Association* 100.469, pp. 322–331. DOI: 10.1198/016214504000001880.

Rudgard, O. (2020). *Growing up near green space makes city children more intelligent and better-behaved*. URL: https://www.telegraph.co.uk/news/2020/08/24/living-near-green-space-makes-city-children-intelligent-better/.

Rudin, C. (2019). “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1, pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (323). “Learning representations by back-propagting errors.” In: *Nature* 6188.533-536.

– (1985). “Learning internal representations by error propagation”. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Ed. by D. E. Rumelhart and J. L. McClelland. Vol. 1. Bradford Books / MIT Press.

Rusbult, C.E., M. Kumashiro, M.K. Coolsen, et al. (2019). “Rusbult NSF Michelangelo Longitudinal study 2002-2004 (V1)”. In: *UNC Dataverse*. DOI: 10.15139/S3/GNBYN5.

Rusbult, C.E., M. Kumashiro, E.J. Finkel, et al. (2019). "Rusbult NSF Michelangelo Longitudinal study 2000-2001". In: *UNC Dataverse*. DOI: 10.15139/S3/RAOEBC.

Rusbult, C.E., J.M. Martz, and C.R. Agnew (1998). "The investment model scale: Measuring commitment level, satisfaction level, quality of alternatives, and investment size". In: *Personal Relationships* 5.4, pp. 357–387. DOI: 10.1111/j.1475-6811.1998.tb00177.x.

Rusbult, C.E. and P.A.M. Van Lange (2003). "Interdependence, interaction, and relationships." In: *Annual Review of Psychology* 54.1, pp. 351–375. DOI: 10.1146/annurev.psych.54.101601.145059.

Ryan, R.M. and E.L. Deci (2000). "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being". In: *The American Psychologist* 55.1, pp. 68–78. DOI: 10.1037/0003-066X.55.1.68.

Sassenberg, K. and L. Ditrich (2019). "Research in social psychology changed between 2011 and 2016: larger sample sizes, more self-report measures, and more online studies". In: *Advances in Methods and Practices in Psychological Science* 2.2, pp. 107–114. DOI: 10.1177/2515245919838781.

Scheel, A.M. (2022). "Why most psychological research findings are not even wrong". In: *Infant and Child Development* 31.1. DOI: 10.1002/icd.2295.

Scheel, A.M. et al. (in press). "Why hypothesis testers should spend less time testing hypotheses". In: *Perspectives on Psychological Science*.

Scherpenzeel, A.C. and M. Das (2010). "Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies". In: ed. by P. Ester M. Das and L. Kaczmirek. Boca Raton: Taylor and Francis. Chap. "True" Longitudinal and Probability-Based Internet Panels: Evidence from the Netherlands, pp. 77–104.

Schmidt, F.L. and I-S. Oh (2016). “The crisis of confidence in research findings in psychology: is lack of replication the real problem? Or is it something else?” In: *Archives of Scientific Psychology* 4, pp. 32–37. DOI: 10.1037/arc0000029.

Schnitzer, M.E. et al. (2014). “Effect of breastfeeding on gastrointestinal infection in infants: a targeted maximum likelihood approach for clustered longitudinal data”. In: *Annals of Applied Statistics* 8.2, pp. 703–725.

Scholkopf, B. (2019). “Causality for machine learning”. In: *arXiv:1911.10500v1*.

Sedlmeier, P. and G. Gigerenzer (1989). “Do studies of statistical power have an effect on the power of studies”. In: *Psychological Bulletin* 105.2, pp. 309–316. DOI: 10.1037/0033-2909.105.2.309.

Shalit, U., F. D. Johansson, and D. Sontag (2017). “Estimating individual treatment effect: generalization bounds and algorithms”. In: *ICML*.

Shannon, C.E. and W. Weaver (1949). *The mathematical theory of communication*. Urbana and Chicago: University of Illinois Press.

Shapley, L.S. (1953). “A value for n-person games”. In: *Contributions to the Theory of Games* 2.28, pp. 307–317.

Shevlin, M. et al. (2021). “Refuting the myth of a ‘tsunami’ of mental ill-health in populations affected by COVID-19: evidence that response to the pandemic is heterogeneous, not homogeneous”. In: *Psychological Medicine*. DOI: 10.1017/S0033291721001665.

Shi, C., D. M. Blei, and V. Veitch (2019). “Adapting neural networks for the estimation of treatment effects”. In: *33rd Conference on Neural Information Processing Systems*.

Shmueli, G. (2010). “To explain or to predict?” In: *Statistical Science* 25.3, pp. 289–310. DOI: doi:10.1214/10-STS330.

Shpitser, I., K. Mohan, and J. Pearl (2015). "Missing data as a causal and probabilistic problem". In: *AUAI*.

Shpitser, I. and J. Pearl (2008). "Complete identification methods for the causal hierarchy". In: *Journal of Machine Learning Research* 9, pp. 1941–1979. DOI: 10.5555/1390681.1442797.

Shrout, P.E. and J.L. Rodgers (2018). "Psychology, science, and knowledge construction: broadening perspectives from the replication crisis". In: *Annual Review of Psychology* 69, pp. 487–510. DOI: 10.1146/annurev-psych-122216-011845.

Siegerink, B. et al. (2016). "Causal inference in law: an epidemiological perspective". In: *European Journal of Risk Regulation* 7.1, pp. 175–186. DOI: 10.1017/S1867299X0000547X.

Smaldino, P. (2019). "Better methods can't make up for mediocre theory". In: *Nature* 575.7781. DOI: 10.1038/d41586-019-03350-5.

Smith, G.T., J.L. Combs, and C.M. Pearson (2012). "APA handbook of research methods in psychology, vol. 1. Foundations, planning, measures, and psychometrics". In: ed. by H. Cooper et al. American Psychological Association. Chap. Brief instruments and short forms, pp. 395–409. DOI: doi:10.1037/13619-021.

Speelman, C.P. and M. McGann (2013). "How mean is the mean?" In: *Frontiers in Psychology* 4.451. DOI: 10.3389/fpsyg.2013.00451.

Spellman, B.A. (2015). "A short (personal) future history of revolution 2.0". In: *Perspectives on Psychological Science* 10.6, pp. 886–899. DOI: 10.1177/1745691615609918.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, prediction, and search*. 2nd. Cambridge, Massachusetts: MIT Press.

Spirtes, P. and K. Zhang (2016). "Causal discovery and inference: concepts and recent methodological advances". In: *Applied Informatics* 3.3. DOI: 10.1186/s40535-016-0018-x.

Spitzer, R.L. et al. (2006). “A brief measure for assessing generalized anxiety disorder: The GAD-7”. In: *Archives of Internal Medicine* 166.10, pp. 1092–1097. DOI: 10.1001/archinte.166.10.109.

Spurk, D. and A.E. Abele (2011). “Who earns more and why? A multiple mediation model from personality to salary”. In: *Journal of Business and Psychology* 26, pp. 87–103. DOI: doi:10.1007/s10869-010-9184-3.

Steeg, G. V. and A. Galstyan (2012). “Information transfer in social media”. In: *WWW*. DOI: 10.1145/2187836.2187906.

– (2013). “Information-theoretic measures of influence based on content dynamics”. In: *WSDM*. DOI: 10.1145/2433396.2433400.

Stein, S. (2020). *Concentration properties of high-dimensional normal distributions*. URL: https://stefan-stein.github.io/posts/2020-03-07-concentration-properties-of-high-dimensional-normal-distributions/.

Stevens, J.R. (2017). “Replicability and reproducibility in comparitive psychology”. In: *Frontiers in Psychology* 8. DOI: 10.3389/fpsyg.2017.00862.

Stricker, J. and A. Günther (2019). “Scientific misconduct in psychology”. In: *Zeitschrift für Psychologie* 227. DOI: 10.1027/2151-2604/a000356.

Strobl, C. et al. (2007). “Bias in random forest variable importance measures: illustrations, sources and a solution”. In: *BMC Bioinformatics* 8.25.

Strobl, E.V. (2018). “A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias”. In: *arXiv:1805.02087v1*.

Sugihara, G. et al. (2012). “Detecting causality in complex ecosystems”. In: *Science* 338.

Sundararajan, M. and A. Najmi (2020). "The many Shapley values for model explanation". In: *arXiv:1908.08474v2*.

Suter, R., D. Miladinovic, and B. Scholkopf (2019). "Robustly disentangled causal mechanisms: validating deep representations for interventional robustness". In: *PMLR*.

Taleb, N. (2019). "Fooled by correlation: Common misinterpretations in social science". In: *Academia Online*.

Tangney, J.P., R.F. Baumeister, and A.L. Boone (2004). "High self-control predicts good adjustment, less pathology, better grades, and interpersonal success". In: *Journal of Personality* 72, pp. 271–324. DOI: 10.1111/j.0022-3506.2004.00263.x.

Tauchert, C., P. Buxmann, and J. Lambinus (2020). "Crowdsourcing data science: a qualitative analysis of organizations' usage of Kaggle competitions". In: *Collaboration for Data Science*.

Tenenbaum, J. B. and T.L. Griffiths (2002). "Theory-based causal inference". In: *Neural Information Processing Systems*.

Tian, J. and J. Pearl (2002). "A general identification condition for causal effects". In: *AAAI*.

Tieleman, T. and G. E. Hinton (2012). "Lecture 6.5 - RMSProp, COURSERA: Neural networks for machine learning." In.

Tong, C. (2019). "Statistical inference enables bad science; statistical thinking enables good science". In: *The American Statistician* 73, pp. 246–261. DOI: 10.1080/00031305.2018.1518264.

van Dalen, H.P. (2021). "How the publish-or-perish principle divides a science: the case of economists". In: *Scientometrics* 126, pp. 1675–1694. DOI: 10.1007/s11192-020-03786-x.

van der Laan, M. J. and S. Rose (2011). *Targeted Learning - Causal Inference for Observational and Experimental Data*. New York: Springer International.

– (2018). *Targeted Learning in Data Science*. Switzerland: Springer International.

van der Laan, M. J. and R. J. C. M. Starmans (2014). "Entering the era of data science: targeted learning and the integration of statistics and computational data analysis". In: *Advances in Statistics*.

van der Laan, M.J. (2015). *Statistics as a science, not an art: The way to survive in data science*. URL: https://magazine.amstat.org/blog/2015/02/01/statscience_feb2015/.

van der Laan, M.J., E.C. Polley, and A.E. Hubbard (2007). "Super Learner". In: *Statistical Applications of Genetics and Molecular Biology* 6.25. DOI: 10.2202/1544-6115.1309.

Van Lange, P.A.M. et al. (1997). "Willingness to sacrifice in close relationships". In: *Journal of Personality and Social Psychology* 72.6, pp. 1373–1395. DOI: 10.1037/0022-3514.72.6.1373.

van Rooij, I. and Blokpoel (2020). "Formalizing Verbal Theories". In: *Social Psychology* 51.5, pp. 285–298. DOI: 10.1027/1864-9335/a000428.

VanderWeele, T.J. (2019). "Principles of confounder selection". In: *European Journal of Epidemiology* 34.3, pp. 211–219. DOI: doi:10.1007/s10654-019-00494-6.

– (2020). "Causal inference and constructed measures: towards a new model of measurement for psychological constructs". In: *arXiv:2007.00520*.

Vankov, I., S. Bowers, and M.R. Munafo (2014). "On the persistence of low power in psychological science". In: *The Quarterly Journal of Experimental Psychology* 67.5, pp. 1037–1040. DOI: 10.1080/17470218.2014.885986.

Verhofstadt, L.L., A. Buysse, and W. Ickes (2007). "Social Support in Couples: An Examination of Gender Differences Using Self-report and Observational Methods". In: *Sex Roles* 57.3-4, pp. 267–282. DOI: 10.1007/s11199-007-9257-6.

Vershynin, R. (2019). *High-dimensional probability: An introduction with applications in data science*. Cambridge: Cambridge Series in Statistical and Probabilistics Mathematics.

Vindegaard, N. and M.E. Benros (2020). "COVID-19 pandemic and mental health consequences: Systematic review of the current evidence". In: *Brain, Behavior, and Immunity* 89, pp. 531–542. DOI: 10.1016/j.bbi.2020.05.048.

Vismara, L., L. Lucarelli, and C. Sechi (2022). "Attachment style and mental health during the later stages of COVID-19 pandemic: The mediation role of loneliness and COVID-19 anxiety". In: *BMC Psychology* 10.1. DOI: 10.1186/s40359-022-00767-y.

Vowels, L. M., M. J. Vowels, and K.P. Mark (2020). "Identifying the Most Important Predictors of Sexual Satisfaction using Interpretable Machine Learning". In: *under review*.

– (2021a). "Is Infidelity Predictable? Using Interpretable Machine Learning to Identify the Most Important Predictors of Infidelity". In: *Journal of Sex Research*.

– (2021b). "Uncovering the Most Important Factors for Predicting Sexual Desire using Interpretable Machine Learning". In: *Journal of Sexual Medicine*. DOI: 10.1016/j.jsxm.2021.04.010.

Vowels, L.M. and K.B. Carnelley (2022). "Partner Support and Goal Outcomes: A Multilevel Meta-Analysis and a Methodological Critique". In: *European Journal of Social Psychology* 52.4, pp. 679–694. DOI: 10.1002/ejsp.2846.

Vowels, L.M., K.B. Carnelley, and R.R.R. Francois-Walcott (2021). "Partner support and goal outcomes during COVID-19: A mixed methods study". In: *European Journal of Social Psychology* 51.2, pp. 393–408. DOI: 10.1002/ejsp.2745.

Vowels, L.M., K.B. Carnelley, and S.C.E. Stanton (2022). "Attachment anxiety predicts worse mental health outcomes during COVID-19: Evidence from two longitudinal studies". In: *Personality and Individual Differences* 185. DOI: https://doi.org/10.1016/j.paid.2021.111256..

Vowels, L.M., M.J. Vowels, and K.P. Mark (2021). "Uncovering the Most Important Factors for Predicting Sexual Desire using Interpretable Machine Learning". In: *Journal of Sexual Medicine*.

Vowels, M. J. (2021). "Misspecification and unreliable interpretations in psychology and social science". In: *Psychological Methods*. DOI: 10.1037/met0000429.

Vowels, M. J., N. C. Camgoz, and R. Bowden (2021). "VDSM: Unsupervised Video Disentanglement with State-Space Modeling and Deep Mixtures of Experts". In: *IEEE Conference on Computer Vision and Pattern Recognition*.

– (2021). "Targeted VAE: Structured inference and targeted learning for causal parameter estimation". In: *IEEE SMDS*.

Vowels, M. J., K. Mark, et al. (2018). "Using spectral and cross-spectral analysis to identify patterns and synchrony in couples' sexual desire". In: *PLoS One* 13.10, e0205330. DOI: 10.1371/journal.pone.0205330.

Vowels, M. J., L. M. Vowels, and N. Wood (2021). "Spectral and Cross-Spectral Analysis-a Tutorial for Psychologists and Social Scientists". In: *Psychological Methods*.

Vowels, M.J. (2020). "Towards an Estimation of Internal State Through Dense, Multi-Modal Representation Learning". In: *Face and Gesture Doctoral Consortium*. DOI: https://fg2020.sunai.uoc.edu/wp-content/uploads/2020/11/Matthew_Vowels.pdf.

– (2022). "Trying to outrun causality with machine learning: Limitations of model explainability techniques for identifying predictive variables". In: *arXiv preprint* arXiv:2202.09875.

Vowels, M.J., S. Akbari, et al. (2023). "A Free Lunch with Influence Functions? An Empirical Evaluation of Influence Functions for Average Treatment Effect Estimation". In: *Transactions on Machine Learning Research*. DOI: https://openreview.net/forum?id=dQxBRqCjLr.

Vowels, M.J., N.C. Camgoz, and R. Bowden (2021). "Shadow-mapping for unsupervised neural causal discovery". In: *IEEE Conference on Computer Vision and Pattern Recognition Causality in Vision Workshop*.

– (2022). "D'ya like DAGs? A survey on structure learning and causal discovery". In: *ACM Comput. Surv.* DOI: 10.1145/3527154.

Wachter, S., B. Mittelstadt, and C. Russell (2018). "Counterfactual explanations without opening the black box: automated decisions and the GDPR". In: *Harvard Journal of Law and Technology* 31.841.

Wagenmakers, E-J. et al. (2012). "An agenda for purely confirmatory research". In: *Perspectives on Psychological Science* 7.6, pp. 632–638. DOI: 10.1177/1745691612463078.

Wang, C. et al. (2020). "A longitudinal study on the mental health of general population during the COVID-19 epidemic in China". In: *Brain, Behavior, and Immunity* 87, pp. 40–48. DOI: 10.1016/j.bbi.2020.04.028..

Wang, Y. and D. M. Blei (2019). "The blessings of multiple causes". In: *arXiv:1805.06826v3*.

Wetherall, M. (1996). *Identities, Groups and Social Issues*. London: SAGE Publications.

Wolpert, D.H. and W.G. Macready (1997). "No free lunch theorems for optimization". In: *IEEE Transacions on Evolutionary Computation* 1.67. DOI: 10.1109/4235.585893.

Wood, J. et al. (2019). "Comparing different planned missingness designs in longitudinal studies". In: *Sankhya B* 81, pp. 226–250. DOI: doi:10.1007/s13571-018-0170-5.

Wright, J. (1998). “Finite element analysis as a loudspeaker design tool”. In: *Journal of the Audio Engineering Society* Paper ML-11. DOI: http://www.aes.org/e-lib/browse.cfm?elib=7992.

Wright, S. (1921). “Correlation and causation”. In: *Journal of Agriculture Research* 20, pp. 557–585.

– (1923). “The theory of path coefficients: a reply to Niles’ criticism”. In: *Genetics* 8, pp. 239–255.

Wu, P.A. and K. Fukumizu (2022). “Intact-VAE: Estimating treatment effects under unobserved confounding”. In: *ICLR*.

Yao, L. et al. (2018). “Representation learning for treatment effect estimation from observational data”. In: *32nd Conference on Neural Information Processing Systems (NeurIPS)*.

Yarkoni, T. (2019). “The generalizability crisis”. In: *PsyArXiv*. DOI: 10.31234/osf.io/jqw35.

Yarkoni, T. and J. Westfall (2017). “Choosing prediction over explanation in psychology: lessons from machine learning”. In: *Perspectives on Psychological Science*. DOI: 10.1177/1745691617693393.

Yoon, J., J. Jordan, and M. van der Schaar (2018). “GANITE: Estimation of individualized treatment effects using generative adversarial nets”. In: *ICLR*.

Zhang, W., L. Liu, and J. Li (2021). “Treatment effect estimation with disentangled latent factors”. In: *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*.

Zou, H. and T. Hastie (2005). “Regularization and variable selection via the elastic net”. In: *J. R. Statist. Soc.* 67.2, pp. 301–320.

Zuo, P.Y. et al. (2020). “A dyadic test of the association between trait self-control and romantic relationship satisfaction”. In: *Frontiers in Psychology* 11. DOI: 10.3389/fpsyg.2020.594476.