



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2010

**ADAPTIVELY WEIGHTED
MAXIMUM LIKELIHOOD ESTIMATION
OF DISCRETE DISTRIBUTIONS**

Michael AMIGUET

Michael AMIGUET, 2010, ADAPTIVELY WEIGHTED MAXIMUM LIKELIHOOD ESTIMATION OF DISCRETE DISTRIBUTIONS

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive.
<http://serval.unil.ch>

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Institut universitaire de médecine sociale et préventive

**ADAPTIVELY WEIGHTED
MAXIMUM LIKELIHOOD ESTIMATION
OF DISCRETE DISTRIBUTIONS**

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Michael AMIGUET

Physicien diplômé de l'Université de Lausanne

Jury

Prof. Luc Tappy, Président
Prof. Alfio Marazzi, Directeur de thèse
Prof. Alberto Holly, expert
Prof. Valentin Rousson, expert
Dr. Eva Cantoni, experte

Lausanne 2010

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

<i>Président</i>	Monsieur Prof.	Luc Tappy
<i>Directeur de thèse</i>	Monsieur Prof.	Alfio Marazzi
<i>Experts</i>	Monsieur Prof.	Alberto Holly
	Monsieur Prof.	Valentin Rousson
	Madame Dr	Eva Cantoni

le Conseil de Faculté autorise l'impression de la thèse de

Monsieur Michael Amiguet

physicien diplômé de l'Université de Lausanne

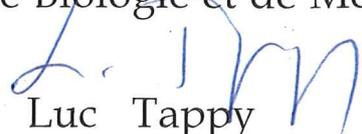
intitulée

**Adaptively Weighted Maximum Likelihood Estimation
of Discrete Distributions**

Lausanne, le 28 janvier 2011

pour Le Doyen
de la Faculté de Biologie et de Médecine

Prof. Luc Tappy



Contents

1	Introduction	5
2	The negative binomial model	13
2.1	Distribution	13
2.2	Maximum likelihood	14
2.3	Notations	16
3	The initial estimator	17
3.1	Notations and general considerations	17
3.2	Median, median absolute deviation	18
3.3	S-estimators of location and scale	20
3.4	Minimum disparity estimators	22
3.4.1	Breakdown point	23
3.4.2	Bias under contamination	27
3.4.3	Trade-off between robustness and efficiency	30
3.4.4	Cressie-Read disparities with $\lambda \leq -1$	35
4	The final estimator	39
4.1	Outlier rejection rules	39
4.1.1	Adaptive cut-off	40
4.1.2	Adaptive weights	42
4.2	The final estimator	44

5	Breakdown point	49
5.1	Asymptotic breakdown point	50
5.2	Finite sample breakdown point	52
6	Asymptotic behavior	55
6.1	Influence function	55
6.2	Asymptotic bias under contamination	58
6.2.1	Maximum asymptotic bias	59
6.2.2	Asymptotic bias for fixed ϵ	61
7	Empirical results	67
7.1	Simulations at the model	69
7.2	Simulations at contaminated models	72
8	Illustration with real data	75
8.1	Chemical mutagenicity data	75
8.2	Lengths of hospital stays	79
9	Computation	83
10	Conclusion and perspectives	85
10.1	Perspectives	88
A	Identifiability in location-scale families	91
B	Proof of Theorem 5	93
C	Bdp in Cressie-Read family with $\lambda \leq -1$	95
D	Simulation results	101

Chapter 1

Introduction

The purpose of the work presented in this dissertation is to develop a robust and efficient parametric estimation method for univariate discrete distributions. Discrete data arise in various research fields, typically when the observations are count data. In biological research for example, one is often concerned with plant or animal counts obtained for each of a set of equal units of space or time. In other experiments, one can be interested in the number of animals carrying a mutation after being exposed to a certain dose of a chemical (see example 8.1 in chapter 8). Another situation where the data are discrete is the analysis of the length of hospital stays, measured in days (see example 8.2 in chapter 8).

Robust estimation of discrete distributions has received some attention in the literature, principally in the framework of minimum disparity estimation. This type of procedures estimates the parameters of a distribution by minimizing a certain disparity between the observed distribution and the model. This method is particularly suited to the discrete framework, which offers the possibility of direct comparison of the observed and the expected

frequencies at each of the sample space elements.¹ A pioneering work by Beran (1977) showed that by using minimum Hellinger distance estimators one could obtain robustness properties together with full asymptotic efficiency (first order efficiency). His approach to robustness contrasted with the M-estimation approach, where the robustness is attained at some sacrifice of asymptotic efficiency (Hampel, Ronchetti, Rousseeuw, and Stahel, 1986). Mathematically, this must occur if one adheres to the notion that the influence function carries most of the critical information about the robustness of a procedure. If one insists, for example, that it be bounded, then it will generally not equal the influence function of the fully efficient maximum likelihood estimator (MLE). However, many authors (e.g. Beran (1977) and Lindsay (1994)) have discussed the limitations of the influence function approach in measuring the robustness of minimum disparity estimators (MDEs). We shall see in section 3.4 that a whole class of MDEs has very attractive robustness properties, both in terms of breakdown point and of contamination bias, while having the same influence function as the MLE.

Further investigation in the line of Beran (1977) came from Tamura and Boos (1986) and Simpson (1987), who provided an appealing justification of the robustness of the minimum Hellinger distance estimator. Cressie and Read (1984) introduced a family of divergences, indexed by a single index λ , which includes many important density-based divergences such as Pearson's and Neyman's chi-squares, Hellinger distance, Kullback-Leibler divergence, as well as the likelihood disparity, whose minimization yields the maximum likelihood estimator. Lindsay (1994) introduced a larger class of disparities, which contains the Cressie-Read disparities, and proposed some new alternatives to the members of the Cressie-Read family. In addition, he studied extensively the efficiency and robustness properties of MDEs in great gen-

¹In the continuous case, the implementation of such methods is made more complicated by the fact that one has to calculate a disparity between a discrete distribution (the empirical distribution) and a continuous one (the model).

erality. Notably, he showed that in the discrete setting, all MDEs are first order efficient. He showed that a trade-off between efficiency and robustness nevertheless existed in this context, taking place between resistance to outliers and second order efficiency as defined by Rao (1961), and he provided a criterion to control this trade-off. Low second order efficiency estimators can be substantially poor compared to the maximum likelihood estimator when the sample size is small. Thus, due to the trade-off, some of the most robust members of the Cressie-Read family can have quite low efficiencies in small samples. The poor performances of these highly robust estimators in small samples have also been noted by Harris and Basu (1994); Basu and Sarkar (1994); Basu et al. (1996), and the presence of potentially important bias in small samples has been recognized by Basu, Basu, and Chaudhuri (1997).

Apart from the MDE approach, some authors (e.g. Cadigan and Chen (2001); Marazzi and Yohai (2010)) developed M-estimating methods applicable to discrete distribution estimation. These estimators are generally not asymptotically fully efficient. They have a robustness-efficiency trade-off, regulated by a scalar parameter, which can be set by fixing a desired asymptotic efficiency.

Our approach to the problem is to build a two-phase estimation procedure, of the type introduced by Marazzi and Ruffieux (1999), Gervini and Yohai (2002) and Marazzi and Yohai (2004). These authors propose to start with a very robust - but not necessarily efficient - estimator (the initial estimator), and to use it to identify the outliers. Then, the outliers are either removed or given low weights, and a weighted maximum likelihood estimator is computed. (This second (final) estimator is defined in such a way as to be consistent when the data follow the model.) Generally, the final estimator keeps the breakdown point of the initial one, while being more efficient. The weights can even be defined in an adaptive way, so that asymptotically, when the data follow the model, no observations are removed or downweighted. This generally gives rise to a first order efficient final estimator.

Under these lines, and with robustness as the main goal, it is tempting to start with some of the most robust minimum disparity estimators, in spite of their possible shortcomings, in order to end up with a highly robust and efficient final estimator. However, it appears that the trade-off in the MDEs between robustness on one hand and bias and efficiency on the other hand, is somewhat transferred to the final estimator. To better explore the situation, we considered as initial estimators a selection of MDEs covering a certain range of the robustness-efficiency trade-off. The result is that the most performing MDEs, those with the best balance between robustness and efficiency, give rise to the most performing final estimators, although the differences in performance are much smaller between the final estimators than between the corresponding initial ones. In nearly all investigated situations, the final estimator outperforms the initial one.

We could have considered starting with an M-estimator, but the MDE approach seems more natural in the discrete setting. Moreover, MDEs need not compromise asymptotic efficiency to acquire robustness. Finally, MDEs seem to be more outlier-resistant in terms of contamination bias than M-estimators.²

We first considered a procedure directly inspired by the methods of Marazzi and Yohai (2004), who define an adaptive cut-off point and remove all observations that are beyond the cut-off. The cut-off is adaptive in that, at the model, it tends to infinity, so that asymptotically no observations are suppressed. This procedure, which we refer to as *truncated maximum likelihood* (TML), gave some promising results, but it was outperformed by some MDEs presenting a particularly good balance in robustness and efficiency. In other words, when the mentioned MDEs are used as initial estimators, the final estimator has weaker performances than the initial one, both in the

²In section 8.2 we apply MDEs to lengths of hospital stays data previously analyzed by Marazzi and Yohai (2010) with an 80% efficient M-estimator, and it is visible in that example that the MDEs are less influenced by the presence of outliers than the M-estimator.

presence of contamination and at the uncontaminated model.

The key idea in the present work is a modified version of the method by Marazzi and Yohai (2004), where each sample space element is given an adaptive downweighting factor, independently of a cut-off point. The downweighting factors are adaptive in the sense that, at the model, they all converge to 1 in probability, and thus asymptotically no observation is downweighted. This method allows to downweight specifically the positions suffering contamination, without removing all larger positions at the same time, thus reducing the efficiency loss. At the same time this procedure allows to reduce the influence of outliers at any position, which was more difficult with the cut-off method. We call this method the *weighted maximum likelihood* (WML). The WML performs better than all the MDEs we used as initial estimates, including the ones which outperformed the TML. The WML is particularly natural in the discrete setting, yet it could be extended to the continuous case, for which a procedure is sketched in section 10.1.

While most of this thesis is formulated in terms of a general family of probability densities on the sample space $X = \{0, 1, 2, \dots\}$, a specific focus is put on the negative binomial (NB) family. NB is a flexible general framework to model discrete data. It is flexible in the sense that it allows for the modeling of over-dispersion, i.e. it can handle situations where the variance is greater than the mean, thus offering a wider scope than the widely used Poisson model, for which the mean is equal to the variance. More specifically, the NB is a generalization of Poisson, which it admits as a limiting case. Let us refer again to the biological research problem of modeling plant or animal counts obtained for each of a set of equal units of space or time. If the individuals are uniformly and independently distributed in space or time the distribution of the counts will be Poisson. Over-dispersion in the counts will arise (somewhat counter-intuitively) if the organisms are “clustered” (meaning that “it is easier for an individual to establish itself close to another individual than further from it” (Clapham, 1936)). If the individuals are

clustered in such a way that the numbers of individuals in the clusters are distributed independently with a logarithmic distribution, the distribution of the counts will be NB (Anscombe, 1950). There are several other ways in which a NB distribution can be obtained, see Anscombe (1950) for a detailed presentation.

While the general results in this thesis are valid in a wide scope of discrete models, some stronger results have been demonstrated in the framework of the negative binomial model. Also, at the time of writing, programs have been developed for the specific case of estimation of the negative binomial parameters, and all examples of application are taken from the NB family.

This thesis is organized as follows: chapter 2 gives a review of the negative binomial model. Chapter 3 considers different candidates for the initial estimator, including two different pairs of location-scales estimators. This approach is attempted because the procedure proposed by Marazzi and Ruffieux (1999) started with location-scale estimators. However they considered continuous density estimation, and different shortcomings of this method in the discrete setting lead us to abandon it. The rest of the chapter concentrates on MDEs, for which we present some known results and establish some new ones. Also, we propose a new MDE which offers a good compromise between robustness and efficiency. Chapter 4 presents the outlier rejection methods and the final estimator. Chapter 5 establishes that the breakdown point (bdp) of the WML is at least as high as the bdp of the initial estimator. In chapter 6 we analyze the asymptotic behavior of the WML. We show that it has the same influence function as the MLE at the model, which strongly suggests that it is asymptotically fully efficient. We also explore its asymptotic bias under contamination, and compare it with the MDEs and the TML. In chapter 7 we give simulation results, both in contaminated and uncontaminated situation. Again, we compare the WML with the MDEs and the TML. Chapter 8 presents two examples of application of the WML

to real data. Chapter 9 concerns the computation of the estimates. Chapter 10 concludes this dissertation and presents a possible method for applying the WML to continuous data.

Chapter 2

The negative binomial model

2.1 Distribution

NB is a two parameter family of discrete probability densities, whose sample space is the set of non-negative integers (including 0). Various parametrizations are possible. We chose the following one: the probability that a variable Y , following a NB distribution with parameters m and α , takes the value y is

$$\text{NB}_{m,\alpha}(x) = \frac{\Gamma(y + \alpha^{-1})}{y! \Gamma(\alpha^{-1})} \left(\frac{\alpha m}{1 + \alpha m} \right)^y (1 + \alpha m)^{-1/\alpha}, \quad (2.1)$$

where $m, \alpha \in \mathbb{R}_+^*$ and $\Gamma(y) = \int_0^\infty t^{y-1} e^{-t} dt$ is the Gamma function.

The expected value of Y is $E(Y) = m$ and its variance is $\text{Var}(Y) = m + \alpha m^2$. Due to the form of the variance, the parameter α is called the dispersion parameter. The variance is always larger than the mean, so that the NB model is specifically suited for overdispersed data. From the expression for the variance, we see that if $\alpha = 0$ the variance becomes equal to the mean, like in the Poisson model. And indeed, letting $\alpha \rightarrow 0$ in (2.1) yields the Poisson distribution of mean m .

This parametrization is convenient firstly because m is the mean of the distribution, thus having an immediate interpretation. A parametrization

using m and α^{-1} is also possible, but the use of α is much more convenient for estimation, as noted by Ross and Preece (1985), Clark and Perry (1989) and Piegorsch (1990). Indeed, using α^{-1} is problematic for various methods of estimation, including the maximum likelihood, and the method-of-moments (Clark and Perry, 1989). In the case of maximum likelihood, we get infinite values of the estimate of α^{-1} as soon as the sample mean exceeds the sample variance, so that the expected value of the estimator is infinite (see next section).

2.2 Maximum likelihood

Since the estimation method we propose (the WML) is a modification of the maximum likelihood (MLE) estimator, we briefly present the MLE in the NB model. The log-likelihood corresponding to distribution (2.1) (up to a constant in (m, α)) is

$$l(y, m, \alpha) = \log \left[\frac{\Gamma(y + \alpha^{-1})}{\Gamma(\alpha^{-1})} \right] + y \log(\alpha m) - (y + \alpha^{-1}) \log(1 + \alpha m). \quad (2.2)$$

Following Lawless (1987) and Piegorsch (1990), we use the following property of the gamma function:

$$\Gamma(y + 1) = y\Gamma(y).$$

Thus we obtain

$$\frac{\Gamma(y + \alpha^{-1})}{\Gamma(\alpha^{-1})} = \begin{cases} 1 & \text{if } x_i = 0 \\ \prod_{\nu=0}^{y-1} (\nu + \alpha^{-1}) & \text{if } x_i > 0, \end{cases}$$

which inserted into (2.2) yields

$$l(y, m, \alpha) = Q(y, \alpha) + y \log(m) - (y + \alpha^{-1}) \log(1 + \alpha m), \quad (2.3)$$

where

$$Q(y, \alpha) = \begin{cases} 0 & \text{if } y = 0 \\ \sum_{\nu=0}^{y-1} \log(1 + \alpha\nu) & \text{if } y > 0. \end{cases}$$

The advantage of using the log-likelihood in form (2.3) is that it no longer contains gamma functions, which simplifies numerical computations (the gamma function grows extremely fast and would often reach the computational limit).

Differentiation of (2.3) with respect to m and α yields the following score functions:

$$s_m(y, m, \alpha) = \frac{y}{m} - \frac{1 + \alpha y}{1 + \alpha m}, \quad (2.4)$$

$$s_\alpha(y, m, \alpha) = \frac{\partial Q}{\partial \alpha}(x, \alpha) + \alpha^{-2} \log(1 + \alpha m) - \frac{m(y + \alpha^{-1})}{1 + \alpha m}, \quad (2.5)$$

where

$$\frac{\partial Q}{\partial \alpha}(y, \alpha) = \begin{cases} 0 & \text{if } y = 0 \\ \sum_{\nu=0}^{y-1} \frac{\nu}{1+\alpha\nu} & \text{if } y > 0. \end{cases}$$

Let us consider a sample $\{y_1, \dots, y_n\}$ of i.i.d. observations. Setting $\sum_{i=1}^n s_m(y_i, \hat{m}, \hat{\alpha}) = 0$ yields $\hat{m} = \bar{y}$.¹ Then, solving

$$\sum_{i=1}^n s_\alpha(y_i, \hat{m}, \hat{\alpha}) = 0 \quad (2.6)$$

for $\hat{\alpha}$ gives the maximum likelihood estimate of α .

It should be mentioned that equation (2.6) does not always have a positive solution. Anscombe (1950) identified that this happens when the sample mean is superior to the sample variance. However, he also established that in such cases the value of α which maximizes the sample likelihood at $m = \bar{y}$, $\sum_{i=1}^n l(y_i, \bar{y}, \alpha)$, over \mathbb{R}_+ , is 0, i.e. we get a Poisson distribution². Thus, the

¹The MLE for m is just the arithmetic mean, which is known to be very non-robust. We shall see in sections 6.2.2 and 7.2 that the MLE for α is also non-robust.

²Some authors (Ross and Preece, 1985; Piegorisch, 1990) suggest that negative values of estimates of α should be allowed, indicating underdispersion compared to Poisson. These values would arise precisely when the sample mean is superior to the sample variance. But then the density would be positive only for $x = 0, 1, \dots, i$, where i is the largest integer less than $-1/\alpha$, and thus the sample space would change (in some cases, one would get a “positive” binomial distribution). Here, we decide to restrict to positive values of α .

MLE of α is defined as the solution of (2.6) if it is positive, and zero otherwise (and it is now clear why the estimation in terms of α^{-1} is more problematic).

The MLEs for m and α are asymptotically independent, and Fisher's information matrix is given by

$$i(m, \alpha) = \begin{pmatrix} \frac{1}{m+\alpha m^2} & 0 \\ 0 & E(s_\alpha^2(Y, m, \alpha)) \end{pmatrix}.$$

2.3 Notations

To simplify notations, when no confusion is possible, the same symbol $\text{NB}_{m,\alpha}$ will be used to denote

- a random variable with a negative binomial distribution of parameters m and α ,
- the cumulative distribution function (cdf) of that variable
- the probability density function of that variable.

Chapter 3

The initial estimator

3.1 Notations and general considerations

Let us start with some notations and definitions. We shall consider the following setup: Let $X = \{0, 1, 2, \dots\}$ be the sample space of a family of probability densities $m_\beta(x)$, indexed by the parameter $\beta \in \Theta \subseteq \mathbb{R}^p$. We will assume that $m_\beta(x) > 0 \forall x \in X, \forall \beta \in \Theta$. Let $\mathbf{x} = \{x_1, \dots, x_n\}$ be an observed sample and define the function $d(x)$ to be equal to the proportion of observations which had value x .¹ Unless otherwise noted, the symbol “ \sum ” will denote summation on variable x over X .

Several candidates for the initial estimator have been considered. One possible approach, applicable to families with a bi-dimensional parameter, is to calculate robust location and scale measures of the observations (Marazzi and Ruffieux, 1999; Marazzi and Barbati, 2003); the initial estimates are then the parameters corresponding to the model whose functional forms of the measures are equal to the observed measures. One difficulty with this approach in the discrete distribution setting is that discrete distribution families

¹Sometimes, e.g. when considering asymptotic situations, $d(x)$ will be defined without an explicit reference to a sample. It should then be considered simply as a function on the sample space.

on a fixed sample space can never be location-scale families². For location-scale families the model parameters are uniquely determined by any pair of location and scale measures (see Appendix A for a proof). This property is not present in the discrete setting, and this method will often suffer an identification problem. Moreover, the problem of finding the initial estimates given the location and scale measures does not reduce to solving the problem for a standard member of the family.

Another approach, more popular in the discrete setting (see the Introduction), is to estimate the model parameters directly, by minimizing a disparity between the model and the observed distribution. Estimators of this type are called *minimum disparity estimators* (MDEs).

The first two estimators presented in this section follow the lines of the first approach described above, by first computing location and scale measures and then finding the corresponding model m_β . These two candidates are presented mainly because they are used in related estimation procedures in the continuous setting (Marazzi and Ruffieux, 1999; Marazzi and Barbati, 2003). They are however quickly discarded because of different shortcomings specific to the discrete setting.

The rest of this section concentrates on MDEs.

3.2 Median, median absolute deviation

A possibility is to start with the median and the median absolute deviation (to the median) of the sample, which are known to be very robust location and scale estimators (e.g. Maronna, Martin, and Yohai (2006)).

Let $\mathbf{x} = (x_1, \dots, x_n)$ be a sample of observations. The median and the

²If F is a family of random variables on Ω , with Ω a discrete subset of \mathbb{R} , $X \in F$ and $a \in \mathbb{R}$, then $aX \notin F$ unless $a = 1$.

median absolute deviation are given by, respectively,

$$\text{Med}(\mathbf{x}) = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)} + x_{((n+1)/2)}}{2} & \text{if } n \text{ is even} \end{cases} \quad (3.1)$$

$$\text{MAD}(\mathbf{x}) = \text{Med}(|\mathbf{x} - \text{Med}(\mathbf{x})|), \quad (3.2)$$

where $x_{(i)}$ is the i^{th} order statistic.

Consider a variable Y on the sample space $X = \{0, 1, 2, \dots\}$, distributed according to the model m_β with cdf $M_\beta(x)$, $x \in X$, where β is a bi-dimensional parameter $\beta^t = (\beta^1, \beta^2) \in \Theta \subseteq \mathbb{R}^2$. The median and the MAD of Y are given by the functionals

$$\text{Med}(m_\beta) = \min\{x : M_\beta(x) \geq 0.5\} \quad (3.3)$$

$$\text{MAD}(m_\beta) = \min\{y : G_\beta(x) \geq 0.5\}, \quad (3.4)$$

where G_β , the cdf of $Z = |Y - \text{Med}(M_\beta)|$, is given by

$$G_\beta(x) = M_\beta(\text{Med}(M_\beta) + x) - M_\beta(\text{Med}(M_\beta) - x) + P(Z = \text{Med}(M_\beta) - x).$$

The initial estimate β_1 of β is then defined as the solution of the system

$$\begin{cases} \text{Med}(\mathbf{x}) &= \text{Med}(m_{\beta_1}) \\ \text{MAD}(\mathbf{x}) &= \text{MAD}(m_{\beta_1}). \end{cases} \quad (3.5)$$

An immediate shortcoming of this method is that β_1 is not uniquely determined by system (3.5).

Indeed, as a consequence of definitions (3.3) and (3.4), the quantities $a = \text{Med}(m_{\beta_1})$ and $b = \text{MAD}(m_{\beta_1})$ are non-negative integers and thus the set A of all possible values of the vector (a, b) is countable. On the other hand, β^1 and β^2 are real numbers, so the set B of all possible values of the vector (β^1, β^2) is not countable. Consequently, no 1-1 application from A to B exists and the solution of (3.5) is in general an infinite subset of B .

One could try to solve this issue by using a trimmed mean and a trimmed (mean) absolute deviation instead of the median and the MAD, however

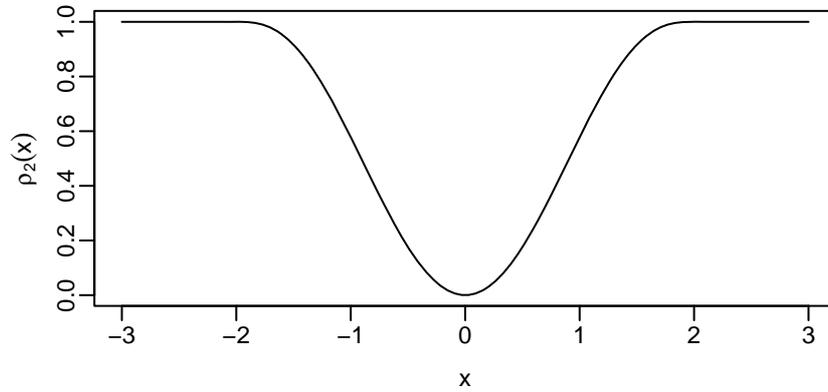


Figure 3.1: Tukey's biweight function with $k = 2$.

an identifiability problem will still be present, regardless of the trimming proportion, for similar reasons to the Med-MAD case.

Consequently, we do not consider this candidate any further.

3.3 S-estimators of location and scale

Another possible choice of location and scale measures is the S-estimator. Let again $\mathbf{x} = (x_1, \dots, x_n)$ be a sample of observations. Define the dispersion measure $S(M)$, for a given value of k , as the solution of

$$\frac{1}{n} \sum_{i=1}^n \rho_k \left(\frac{x_i - M}{S(M)} \right) = 0.5,$$

where

$$\rho_k(x) = \begin{cases} 1 - [1 - (x/k)^2]^3 & \text{if } |x| \leq k \\ 1 & \text{if } |x| > k \end{cases} \quad (3.6)$$

is Tukey's biweight function (see Figure 3.1).

The S-estimates of location and scale, respectively $\mu_S(\mathbf{x})$ and $\sigma_S(\mathbf{x})$, are defined as follows:

$$\begin{aligned} \mu_S(\mathbf{x}) &= \arg \min_M S(M) \\ \sigma_S(\mathbf{x}) &= S(\mu_S(\mathbf{x})). \end{aligned}$$

For a distribution F , define $\mu_S(F)$ and $\sigma_S(F)$ as the asymptotic values of, respectively, $\mu_S(\mathbf{x})$ and $\sigma_S(\mathbf{x})$, where \mathbf{x} is generated according to F .

The initial estimate β_1 of the bi-dimensional parameter β , is then defined by

$$\begin{cases} \mu_S(\mathbf{x}) &= \mu_S(m_{\beta_1}) \\ \sigma_S(\mathbf{x}) &= \sigma_S(m_{\beta_1}). \end{cases} \quad (3.7)$$

There are at least two drawbacks to this method. One is that the system (3.7) does not always have a solution. Numerical investigation in the NB model shows that there is a lower bound $b(\mu_S(\mathbf{x}))$ such that when $\sigma_S(\mathbf{x}) < b(\mu_S(\mathbf{x}))$ there is no solution to (3.7). Such a situation has been observed for up to 2.5% of the samples in simulations with small ($n < 20$) sample sizes.

The second drawback is that we have an identifiability problem again, due to the fact that S-estimates “collapse” when more than half of the data have the same value. Indeed, if more than $n/2$ observations are equal to a certain value x , then

$$\mu_S(\mathbf{x}) = x \quad \text{and} \quad \sigma_S(\mathbf{x}) = 0.$$

In the negative binomial family, a whole subset of models have

$$P(\text{NB}_{m,\alpha} = 0) > 0.5.$$

More precisely, $P(\text{NB}_{m,\alpha} = 0) = (1 + \alpha m)^{-1/\alpha}$ is an increasing function of α which tends to 1 for $\alpha \rightarrow \infty$, for any fixed finite value of m . This implies that there exists a function $\alpha_l(m)$ such that

$$P(\text{NB}_{m,\alpha} = 0) > 0.5 \quad \forall \alpha > \alpha_l(m).$$

Moreover, since

$$\lim_{\alpha \rightarrow 0} P(\text{NB}_{m,\alpha} = 0) = \exp(-m).$$

and since, for fixed α , $P(\text{NB}_{m,\alpha} = 0)$ is a decreasing function of m , we have that

$$P(\text{NB}_{m,\alpha} = 0) > 0.5 \quad \forall (m, \alpha) \in (0, \log(2)) \times \mathbb{R}_+.$$

In count data, situations where we have a high proportion of zero counts are not rare. For instance, in section (8.2) we model chemical mutagenicity data in drosophila, and the proportion of zero counts in that example is nearly 70%. These data could not have been analyzed with the method described in this section.

Thus, again, we decide not to consider this method any further.

3.4 Minimum disparity estimators

Yet another approach is to estimate directly the model parameters by minimizing a disparity measure between the observed distribution and the model. This seems to be the most natural approach in the discrete setting (and also the most popular in the literature, see the Introduction). In this section we present in some details different properties of the minimum disparity estimators (MDEs), such as their breakdown point, their asymptotic bias under contamination, their efficiency. We also establish some new results, valid in the negative binomial setting, showing that in that framework some MDEs resist to extremely high proportions of outliers. In section 3.4.3 we introduce a new disparity measure, called the *linearized negative exponential disparity*, which will be seen to have nice robustness and efficiency properties.

We shall consider a general class of disparities introduced by Lindsay (1994). Define the *Pearson residual function* $\delta(x)$ as

$$\delta(x) = \frac{d(x) - m_\beta(x)}{m_\beta(x)},$$

and define the *disparity measure* between the probability densities $m_\beta(x)$ and $d(x)$ as

$$\rho(d, m_\beta) = \sum m_\beta(x)G(\delta(x)), \quad (3.8)$$

where G is a real-valued thrice-differentiable strictly convex function on $[-1, \infty)$ with $G(0) = 0$. Applying Jensen's inequality to ρ shows that it

is non-negative (for any pair of densities) and an argument by Csiszár (1963) shows that it is zero only when $d(x) = m_\beta(x) \forall x \in X$. Therefore, the estimator of β defined as

$$T(d) = \arg \min_{\beta} \rho(d, m_\beta), \quad (3.9)$$

is Fisher-consistent. We call T a *minimum disparity estimator* (MDE).

3.4.1 Breakdown point

We need a criterion to measure outlyingness in the framework of discrete distributions. The Pearson residual

$$\delta(x) = \frac{d(x)}{m_\beta(x)} - 1 \quad (3.10)$$

offers a natural measure of how surprising the proportion of observations at x is with respect to a given model. More precisely, if the observed frequency $d(x)$ is too large compared to the prediction of the model $m_\beta(x)$, the Pearson residual will be large. Accordingly, in the remainder of this thesis, the term *outlier* will denote an element of the sample space - not a single observation - with a large Pearson residual. Note that this definition is model dependent.

Lindsay (1994) carried out a thorough investigation of MDE's. Notably, he proved an important result about the breakdown properties of certain MDEs, which we expose hereafter, before we give some extensions.

For $\epsilon \in [0, 1)$ and the data $d(x)$, define the ϵ -contaminated data $d_j(x)$ as

$$d_j(x) = (1 - \epsilon)d(x) + \epsilon\chi_{x_j}, \quad (3.11)$$

where $\chi_{x_j}(x)$ is the indicator function for x_j . Let

$$d_\epsilon^*(x) = (1 - \epsilon)d(x),$$

let T be a MDE and ρ be the corresponding disparity.

Assumption 1. $G(-1)$ is finite and $\lim_{\delta \rightarrow \infty} G(\delta)/\delta = 0$.

Assumption 2. $\rho(d_j, m_\beta)$ and $\rho(d_\epsilon^*, m_\beta)$ are continuous in β , with the latter having unique absolute minimum at $T(d_\epsilon^*) = b^*$.

Consider a sequence $\{x_j : j = 1, 2, \dots\}$ of elements of the sample space X .

Definition 3. $\{x_j\}$ constitutes an *outlier sequence* for the model $m_\beta(x)$ and the data $d(x)$ if $m_\beta(x_j) \rightarrow 0$ and $d(x_j) \rightarrow 0$ as $j \rightarrow \infty$.

Note that definition 3 is equivalent to requiring $d(x_j) \rightarrow 0$ and $\delta_j(x_j) \rightarrow \infty$ as $j \rightarrow \infty$, where $\delta_j(x)$ is the Pearson residual corresponding to the ϵ -contaminated data

$$d_j(x) = (1 - \epsilon)d(x) + \epsilon\chi_{x_j}(x)$$

and the model $m_\beta(x)$. Thus the elements of an outlier sequence will get larger and larger Pearson residuals if a finite mass is placed on them, in accordance with our definition of outliers at the beginning of this section.

Remark A. In our sample space $X = \{0, 1, 2, \dots\}$, the requirement that $m_\beta(x) > 0 \forall \beta \in \Theta, \forall x \in X$, implies that $\forall d(x)$ and $\forall m_\beta(x)$, $\{x_j\}$ is an outlier sequence iff $\lim_{j \rightarrow \infty} x_j = \infty$.

Now consider the following asymptotic situation: let m_{β_0} be an element of the family of models m_β , let $\{x_j\}$ be an outlier sequence, and consider the *ϵ -contaminated model*

$$m_{\beta_0j}(x) = (1 - \epsilon)m_{\beta_0}(x) + \epsilon\chi_{x_j}(x).$$

Lindsay's result: under Assumptions 1 and 2 (with m_{β_0j} instead of d_j and $m_{\beta_0\epsilon}^* = (1 - \epsilon)m_{\beta_0}$ instead of d_ϵ^*) and some mild assumptions about the model m_β , it holds that

$$\lim_{j \rightarrow \infty} T(m_{\beta_0j}) = \beta_0$$

as soon as $\epsilon < 0.5$. (See Bhandari, Basu, and Sarkar (2006) for an analogous result for continuous distributions.)

Lindsay's result shows that some MDE's are asymptotically unaffected by extreme outliers up to a proportion as large as 0.5.

Proposition 4. Lindsay's result remains valid if the contamination is a finite sum of outlier sequences, i.e. at the model

$$m_{\beta_0 S_j}(x) = (1 - \epsilon)m_{\beta_0}(x) + \epsilon S_j(x),$$

where $S_j(x) = \sum_{i=1}^n \frac{\epsilon_i}{\epsilon} \chi_{x_j+a_i}(x)$, with n finite, $a_i \in \mathbb{N}$, $\epsilon_i \in [0, \epsilon]$, $\sum_{i=1}^n \epsilon_i = \epsilon$ and $\{x_j\}$ an outlier sequence. (All the sequences $\{x_j + a_i\}$ are outlier sequences as soon as $\{x_j\}$ is, see Remark A.)

Proof. Simply proceed as in Lindsay (1994) (proofs of his Proposition 12 and Lemma 20), by doing the summations separately on the contaminated and uncontaminated parts of the sample space, the only difference being that in Lindsay (1994) the contaminated part consists of one sample space element and here it consists of at most n elements (n if all a_i are different).

■

If we set $m_\beta = \text{NB}_{m,\alpha}$, Lindsay's result can be extended.

Theorem 5. Let $m_\beta = \text{NB}_{m,\alpha}$, let $d(x)$ be the observed data and

$$d_{S_j}(x) = (1 - \epsilon)d(x) + \epsilon S_j(x)$$

with S_j as in Proposition 4. Then under Assumptions 1 and 2, it holds that

$$\lim_{j \rightarrow \infty} T(d_{S_j}) = b^*,$$

with b^* as in Assumption 2, for any $\epsilon \in [0, 1)$.

The proof is given in Appendix B.

Remark B. A complete determination of the breakdown point of MDEs would require to investigate the behavior of the estimates under contamination of the model with any probability density. In this thesis, we will concentrate on the effect of large outliers, and henceforth the term “breakdown point” (bdp) will refer to the minimum value of ϵ for which there can be a breakdown in the presence of a contamination S_j as in proposition 4, when $j \rightarrow \infty$.

Theorem 5 shows that in the framework of estimation in the negative binomial family, some MDEs resist to the presence of extreme outliers regardless of their proportion in the sample. This may seem surprising, as usually the highest possible (and sensible) value for a bdp is 0.5. For higher values the estimator is not fitting the majority of the data anymore, and one may question whether this is a desirable property. Well, in the framework of density estimation it may be. Imagine we know from previous investigation that a certain phenomenon we are interested in has a certain typical shape. In the presence of highly corrupted data, an estimator with a very high bdp is able to recognize that shape and fit it even if it is followed by less than half the observations.

To gain some insight into the mechanism that causes the bdp to be that high, recall from the proof of theorem 5 that this property is linked to the uniform convergence

$$\text{NB}_{m,\alpha}(x_j) \rightarrow 0 \text{ as } x_j \rightarrow \infty$$

over the whole parameter space. In other words, no model in the negative binomial family can nicely accommodate observations going to infinity.

Let us stress that this result is valid for contamination of any sample d with finite $T(d_\epsilon^*)$, not only asymptotically like Lindsay’s general result. Moreover Lemma 21 in Lindsay (1994) shows that if d is a model density m_{β_0} then

$$b^* = \beta_0$$

so the bias due to contamination tends to zero as $j \rightarrow \infty$.

Finally, let us mention that Theorem 5 is also valid if the model is the Poisson density or the geometric density, as these are particular cases of the negative binomial family.

3.4.2 Bias under contamination

It will be useful, for what follows, to formulate the minimization problem (3.9) in terms of estimating equations. Under differentiability of the model, minimizing the disparity (3.8) is equivalent to solving the equations

$$\sum A(\delta(x)) \nabla m_\beta(x) = 0, \quad (3.12)$$

where ∇ denotes differentiation with respect to β ,

$$A(\delta) = \frac{\tilde{A}(\delta) - \tilde{A}(0)}{\tilde{A}'(0)} \quad (3.13)$$

for $\tilde{A}(\delta)$ given by

$$\tilde{A}(\delta) = (1 + \delta)G'(\delta) - G(\delta).$$

It is easy to see from the requirements on G and from (3.13) that A is a strictly increasing twice-differentiable function on $[-1, \infty)$ with $A(0) = 0$ and $A'(0) = 1$. Lindsay (1994) called such a function a *residual adjustment function* (RAF). As indicated by (3.12), for a given model, many of the properties of the estimator are determined by A . An important special case is obtained with $A(\delta) = \delta$, which yields the maximum likelihood estimator (MLE).

The formulation (3.12) provides some insight into the mechanism that gives robustness to certain MDEs. In (3.12), it appears that a disparity for which $A(\delta) < \delta$ for large δ gives a lower weight to the contributions of outliers than the likelihood disparity. This property can have direct bearing on the robustness properties of the corresponding estimators. Let us look at some examples of disparities to illustrate this point.

An important class of disparities, the Cressie-Read family of power-divergence measures (Cressie and Read, 1984; Read and Cressie, 1988), is obtained by using

$$G_\lambda(\delta(x)) = \frac{(1 + \delta(x))^{\lambda+1} - 1}{\lambda(\lambda + 1)}$$

for $G(\delta(x))$ in (3.8). The corresponding RAFs are given by

$$A_\lambda(\delta) = \frac{(1 + \delta)^{\lambda+1} - 1}{\lambda + 1}.$$

Many well known measures are obtained for specific values of λ :

- $\lambda = 1$: Pearson's chi-squared (divided by 2)

$$\frac{1}{2} \sum \frac{(d(x) - m_\beta(x))^2}{m_\beta(x)}$$

- $\lambda = 0$: Likelihood disparity

$$\sum d(x) \left[\log(d(x)) - \log(m_\beta(x)) \right]$$

- $\lambda = -\frac{1}{2}$: Squared Hellinger distance (multiplied by 2)

$$2 \sum \left[\sqrt{d(x)} - \sqrt{m_\beta(x)} \right]^2$$

- $\lambda = -1$: Kullback-Leibler divergence

$$\sum m_\beta(x) \left[\log(m_\beta(x)) - \log(d(x)) \right]$$

- $\lambda = -2$: Neyman's chi-squared (divided by 2)

$$\frac{1}{2} \sum \frac{(d(x) - m_\beta(x))^2}{d(x)}$$

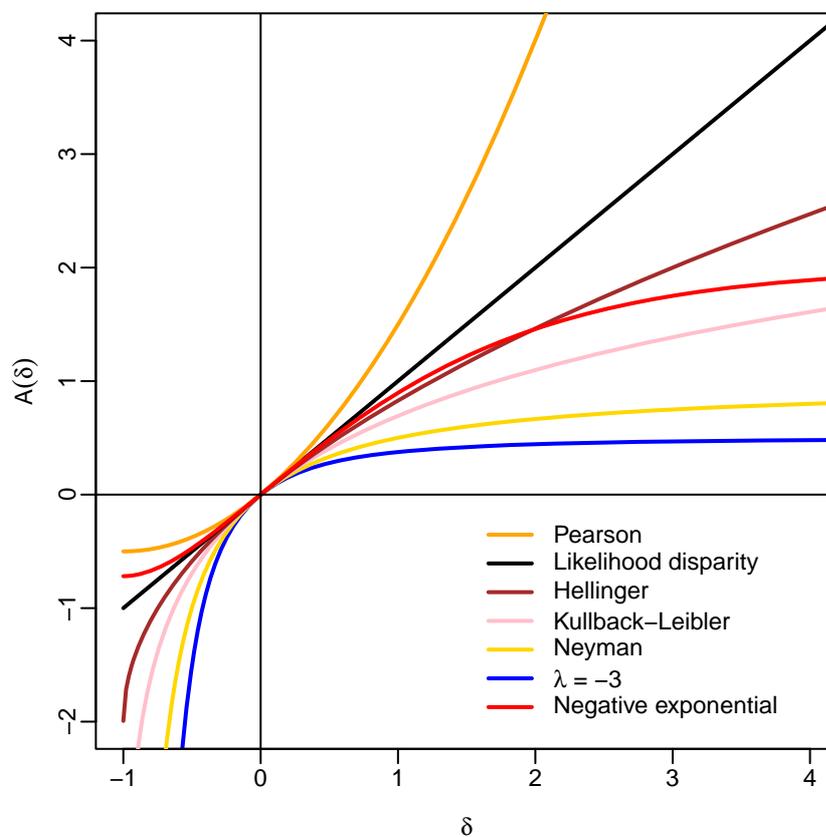


Figure 3.2: The RAF's of various disparities

Another important disparity, the *negative exponential disparity measure* (NE) (Lindsay, 1994), is obtained by using

$$G_{\text{NE}}(\delta) = e^{-\delta} - 1$$

or the RAF

$$A_{\text{NE}}(\delta) = 2 - (2 + \delta)e^{-\delta}.$$

Figure 3.2 shows the RAF's of the disparities mentioned above, as well as the one obtained from the Cressie-Read family with $\lambda = -3$. On figure 3.2, it appears that in the Cressie-Read family, as soon as $\lambda < 0$, the contributions of outliers are given a lower weight in the estimating equations

than with maximum likelihood estimation. Moreover, the lower λ , the more those contributions are downweighted. In relation to the bdp, it is easily checked that the Cressie-Read disparities with $-1 < \lambda < 0$ and the negative exponential disparity satisfy Assumption 1, and thus give rise to high bdp estimators. The Cressie-Read disparities with $\lambda \leq -1$ can also be shown to yield high bdp estimators, even though they do not satisfy Assumption 1, since $G_\lambda(-1) = \infty$ if $\lambda \leq -1$.³

Let us show how the downweighting of outliers influences the bias under point contamination. Figures 3.3 and 3.4 show the asymptotic biases of a selection of MDEs in the contaminated NB model

$$\text{NB}_{m,\alpha j}(x) = (1 - \epsilon)\text{NB}_{m,\alpha}(x) + \epsilon\chi_{x_j}(x),$$

for a wide range of contamination positions x_j , the two models $(m, \alpha) = (5, 2/3)$ and $(m, \alpha) = (7, 1.5)$, and different values of ϵ . In addition to the already defined MDEs, a MDE called linNEG, to be defined below, is also plotted. It is readily seen that the downweighting of outliers has direct bearing on the bias curves. It also appears that, as predicted, even with a proportion of outliers larger than 0.5 the bias tends to zero as the outliers go to infinity, except in the case of the MLE, which diverges. Finally, let us mention that the largest bias generally occurs for a contamination at 0. For better readability of the graphs, these biases are not shown on Figures 3.3 and 3.4. The bias at zero of the MDEs is of comparable size to that of the MLE.

3.4.3 Trade-off between robustness and efficiency

If we consider Figures 3.3 and 3.4, it seems appropriate to choose a very low negative value of λ in the Cressie-Read family, in order to get a very robust

³ $\delta(x) = -1$ occurs when $d(x) = 0$, i.e. if the cell at x is empty. Thus, in the present form, the disparities with $\lambda \leq -1$ are not defined as soon as there is an empty cell. A modified definition is proposed a bit further, for which a proof of high bdp is given.

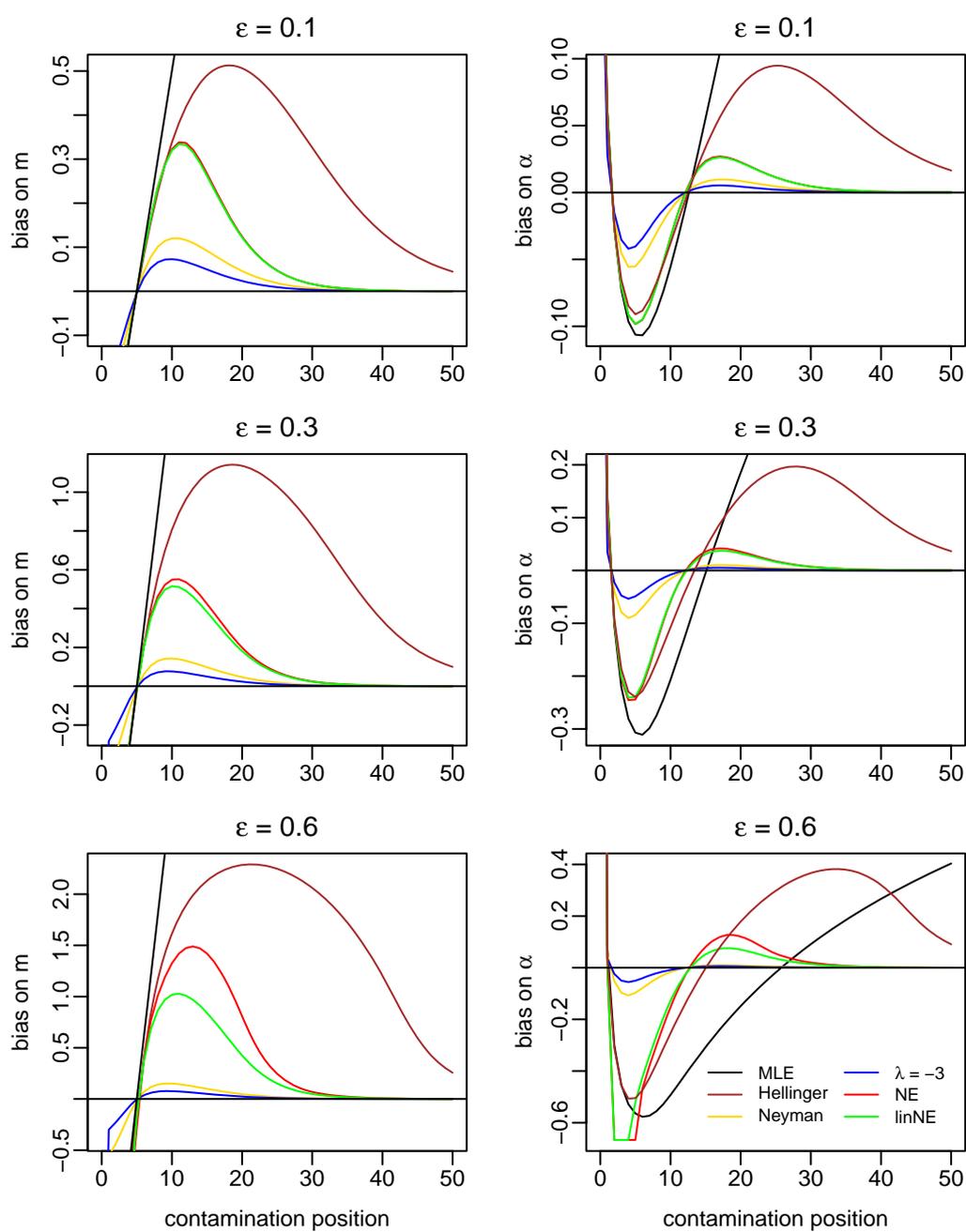


Figure 3.3: Asymptotic bias for different MDEs under point contamination of model $NB_{5, \frac{2}{3}}$.

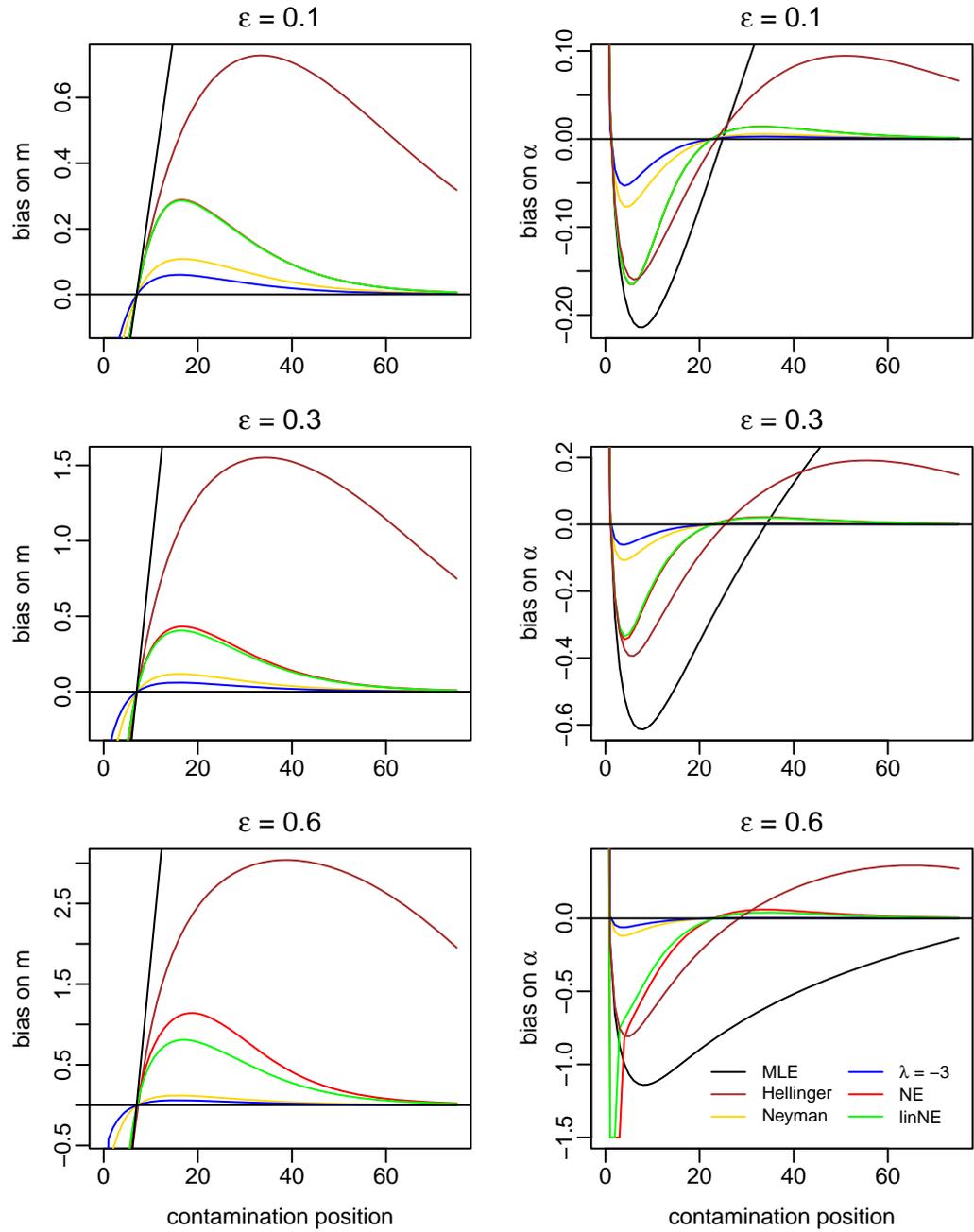


Figure 3.4: Asymptotic bias different MDEs under point contamination of model $NB_{7,1.5}$.

initial estimator.

However, the shape of the RAF has other important implications than robustness properties. In what follows we report a few important points, most of which are developed in Lindsay (1994).

- All MDEs are asymptotically fully efficient, having the same influence function as the MLE at the model⁴. This shows the limitations of using the influence function as a measure of robustness, as some MDEs are highly robust both in terms of breakdown point and in terms of bias under contamination, as was shown in the course of this section.
- $A_2 = A''(0)$, the second derivative of the RAF at 0, provides a (somewhat questioned, see Basu and Sarkar (1994)) trade-off between the robustness and the efficiency of MDEs in finite samples (see also Basu and Lindsay (1994)). A_2 is linked to the *second-order efficiency* of the estimator in the sense of Rao (1961, 1962) (itself subject to some controversy, see Berkson (1980)). The lower the absolute value of A_2 , the higher the second-order efficiency. If $A_2 = 0$ the estimator has the same second-order efficiency as the MLE, which is optimal under this criterion. In the Cressie-Read family, $A_2 = \lambda$. Therefore the higher resistance to outliers of the low λ members of the Cressie-Read family can be disturbed by their lower efficiency and show poorer performances in terms of mean square error.
- Harris and Basu (1994), Basu, Harris, and Basu (1996) and Basu and Basu (1998) noted that the more robust MDEs in the Cressie-Read family can show poor performances (in terms of efficiency) when the sample size is small. They linked this fact to the shortcomings of those estimators in the treatment of *inliers*, i.e. cells with a lower observed

⁴This is linked to the fact that all RAFs have, by definition, the same first order Taylor expansion $A(\delta) \approx \delta$.

frequency than expected under the model (this was also noted by Lindsay (1994)). Indeed, it is clear from Figure 3.2 that the MDEs in the Cressie-Read family with negative values of λ give higher weight to inliers than does the MLE.

In the NB model, there appears to be one more shortcoming of the lower λ MDEs from the Cressie-Read family, which is the presence of an important bias under the uncontaminated model, when the sample size is small. This is likely to be caused by the treatment of inliers again, as this problem does not affect the minimum NE, the minimum linNE or the maximum likelihood estimators. As will be illustrated with simulations in Chapter 7, this can lead to “reverse effects” of contamination of the model with outliers: when the bias under the model is negative, the presence of a contamination can reduce the bias.

The above considerations seem to point out the negative exponential disparity as a good choice: it shows important downweighting of large outliers, it is second order efficient ($A_2 = 0$) and it also downweights the contributions of inliers, compared to the likelihood disparity. However, our main focus in this thesis is to build a good outlier resistant estimator; the problem of inliers is not our direct concern here.

Accordingly, we define a new disparity, whose RAF is equal to the RAF of the MLE for $-1 < \delta \leq 0$ and to the RAF of the NE for $\delta > 0$. This disparity is designed to have the same desirable properties as the NE - outlier downweighting, second order efficiency - while being similar to the MLE in the treatment of inliers. We call this new disparity the *linearized negative exponential disparity* (linNE). The RAF of the linNE satisfies the corresponding requirements, in particular it is twice differentiable, since the NE is second order efficient. As noted, the linNE is also second order efficient. As can be seen on Figures 3.3 and 3.4, the asymptotic bias under point contamination for linNE is quite similar to the bias for NE (it is slightly lower at the highest contamination rate).

In light of the foregoing, we decide to keep five different initial estimators for further investigation (in chapters 6 and 7):

- Three members of the Cressie-Read family: the minimum Hellinger distance estimator, as it has been pointed out by Lindsay (1994) as presenting a nice balance between robustness and efficiency, and two more robust estimators, the minimum Neyman's chi-squared estimator and the MDE with $\lambda = -3$, as efficiency is not our main target for the initial estimator. A lack in efficiency could be fixed in the second phase of the estimation process.
- The minimum negative exponential disparity estimator
- The minimum linearized negative exponential disparity estimator, obtained from the RAF

$$A_{\text{linNE}}(\delta) = \begin{cases} \delta & \text{if } -1 \leq \delta \leq 0 \\ 2 - (2 + \delta)e^{-\delta} & \text{if } \delta > 0 \end{cases}$$

3.4.4 Cressie-Read disparities with $\lambda \leq -1$

As noted before, the disparities in the Cressie-Read family with $\lambda \leq -1$ are not defined as soon as there is an empty cell with $d(x) = 0$, as then $\delta(x) = -1$ which causes $G_\lambda(x)$ to become infinite and $A_\lambda(x)$ to go to $-\infty$. An immediate remedy for this issue is to exclude empty cells from the definition of the disparities, thus summing only over $\mathbf{X}_F = \{x \in \mathbf{X} : d(x) \neq 0\}$ in (3.8), i.e. minimizing

$$\rho_{\mathbf{X}_F}(d, m_\beta) = \sum_{\mathbf{X}_F} m_\beta(x) G(\delta(x)), \quad (3.14)$$

over β to find the estimate. However, if we do so, then applying Jensen's inequality to the disparity does not provide us with a 0 lower bound but rather with the lower bound

$$G_\lambda \left(\frac{1}{\sum_{\mathbf{X}_F} m_\beta(x)} - 1 \right) \sum_{\mathbf{X}_F} m_\beta(x),$$

which is zero only if $\sum_{\mathbf{X}_F} m_\beta(x) = 1$ and is negative otherwise, since G_λ is decreasing if $\lambda < 0$, and $G_\lambda(0) = 0$. It then becomes unclear what the minimum over β is, but as long as $\sum_{\mathbf{X}_F} m_\beta(x) \neq 1$, i.e. on any finite sample, it cannot correspond to $m_\beta(x) = d(x) \forall x \in \mathbf{X}_F$, as this situation is impossible since $\sum_{\mathbf{X}_F} d(x) = 1$. Moreover, It can be seen from (3.14) that if $m_\beta(x)$ goes to 0 all over \mathbf{X}_F the disparity goes to 0, as $\lim_{\delta \rightarrow \infty} G_\lambda(\delta)/\delta = 0$ if $\lambda < 0$. Depending on $d(x)$, this can cause the minimum of the disparity to correspond to a model for which $\sum_{\mathbf{X}_F} m_\beta(x)$ is small i.e. a model lying away from the observations. To prevent this type of behavior, we can minimize the disparity calculated with the conditioned model

$$\tilde{m}_\beta(x) = \frac{m_\beta(x)}{\sum_{\mathbf{X}_F} m_\beta(x)}.$$

Then the disparity cannot be made small by making the model small all over \mathbf{X}_F , since $\sum_{\mathbf{X}_F} \tilde{m}_\beta(x) = 1$, and applying Jensen's inequality to it yields 0 as lower bound. This lower bound is attained if $\tilde{m}_\beta(x) = d(x) \forall x \in \mathbf{X}_F$, which asymptotically becomes equivalent to $m_\beta(x) = d(x) \forall x \in \mathbf{X}$.

In many cases this is a good cure, yet there can be another problem. It may happen that the conditioned model reaches a limiting distribution on \mathbf{X}_F when certain components of β go to infinity. If that distribution is similar to the observed distribution, the MDE can diverge. To solve this issue, we propose to add a ‘‘departure penalty’’ to the disparity, which will penalize the models that lie apart from the observations, i.e. for which $S_{\mathbf{X}_F}(\beta) = \sum_{\mathbf{X}_F} m_\beta(x)$ is small. We suggest to minimize the following expression:

$$\rho_p(d, m_\beta) = \sum_{\mathbf{X}_F} \tilde{m}_\beta(x) G(\tilde{\delta}(x)) + P(S_{\mathbf{X}_F}(\beta)), \quad (3.15)$$

where $\tilde{\delta}(x) = \frac{d(x) - \tilde{m}_\beta(x)}{\tilde{m}_\beta(x)}$ and $P(S_{\mathbf{X}_F}(\beta)) = \frac{1 - S_{\mathbf{X}_F}(\beta)}{S_{\mathbf{X}_F}(\beta)}$. Figure 3.5 shows the shape of the penalty $P(S_{\mathbf{X}_F}(\beta))$. Let us emphasize that asymptotically there are no empty cells, and so expression (3.15) becomes equivalent to the standard disparity (3.8) and the influence function of the estimator at the model is still the same as the influence function of the MLE.

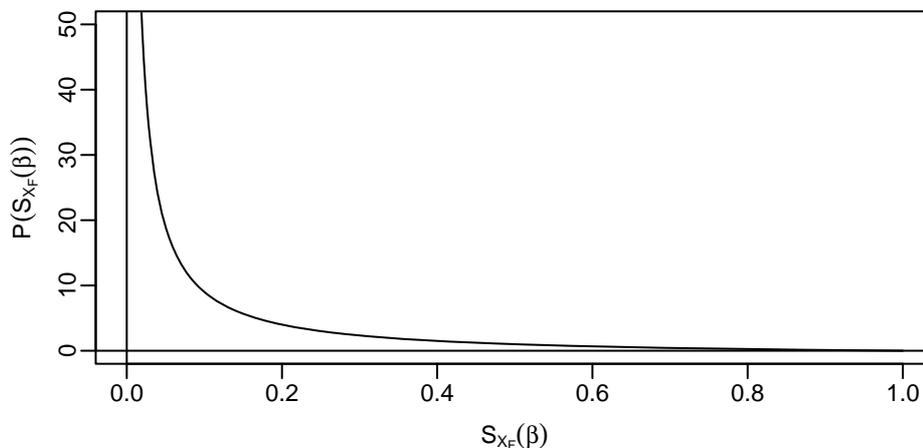


Figure 3.5: The “departure penalty”.

Figure 3.6 shows two examples of situations where the above adjustments are necessary to prevent the estimates from taking extreme values, in the NB_β model, with $\beta = (m, \alpha)$. The two represented samples were obtained by generation of 50 (upper panel) and 40 (lower panel) pseudo-random numbers from the model $\text{NB}_{4,1.5}$. The three following estimates were computed on each of the two samples:

- $\beta_{\mathbf{X}_F} = \arg \min_{\beta} \rho_{\mathbf{X}_F}(d, \text{NB}_\beta)$
- $\tilde{\beta}_{\mathbf{X}_F} = \arg \min_{\beta} \rho_{\mathbf{X}_F}(d, \widetilde{\text{NB}}_\beta)$
- $\beta_p = \arg \min_{\beta} \rho_p(d, \widetilde{\text{NB}}_\beta)$

where $\widetilde{\text{NB}}_\beta(x) = \frac{\text{NB}_\beta(x)}{\sum_{\mathbf{X}_F} \text{NB}_\beta(x)}$ and we have used the Cressie-Read disparity with $\lambda = -2$, corresponding to Neyman’s chi-squared. It may be seen that both $\beta_{\mathbf{X}_F}$ and $\tilde{\beta}_{\mathbf{X}_F}$ are way too large in parameter m , the latter even diverging in the second example (the actual minimum is probably reached for $m = \infty$). On both graphs, we also plotted the distribution corresponding to $\widetilde{\text{NB}}_{\tilde{\beta}}$, which is indeed quite close to the data in both cases. In both examples, the model NB_{β_p} is seen to be in good agreement with the data.

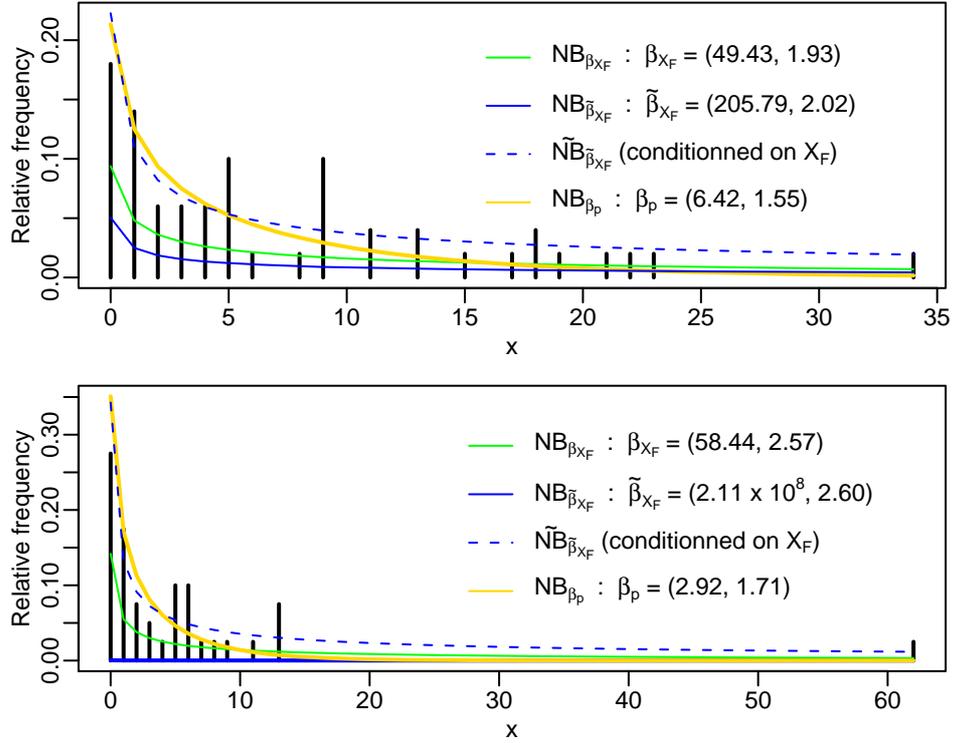


Figure 3.6: Two examples of the necessity for adjustments when $\lambda < -1$. The vertical bars are the relative frequencies of pseudo-random numbers generated from $NB_{4,1.5}$.

In the remainder of this thesis, MDEs from the Cressie-Read family with $\lambda \leq -1$ will be calculated by minimizing (3.15). In the NB model, the MDEs with $\lambda \leq -1$ defined in this way resist to outliers going to infinity independently of their proportion in the sample, like the MDEs with $-1 < \lambda < 0$. A proof is given in Appendix C.

Chapter 4

The final estimator

4.1 Outlier rejection rules

We suppose that $\beta_1(d)$ is a consistent high bdp initial estimate of the model parameters. In what follows we consider two outlier rejection methods, based on the initial estimate. Once outliers have been removed (or downweighted) a corrected maximum likelihood estimator (the final estimator) is computed with the remaining observations. This final estimator is presented in the next section.

The first rejection method calculates an adaptive cut-off and rejects observations larger than the cut-off, so that outlier sequences end up being removed from the sample. This method is based on a proposal by Marazzi and Yohai (2004). At the model, for increasing sample sizes, the cut-off tends to infinity, and so asymptotically no observations are removed. We call the final estimator based on this method the *cut-off weighted maximum likelihood estimator* (WMLc). This approach is presented mainly for comparison purposes with the second method, which yields a final estimator that generally outperforms the WMLc.

The second method calculates adaptive weights for each element of the sample space. The weights are based on standardized differences between the

expected frequencies under the initial model and the observed frequencies; too large differences being downweighted. This way, not only are outlier sequences eventually removed from the sample, but the influence of too large observed frequencies at any place is lowered, causing the final estimator to have very low bias under contamination. At the model, for increasing sample sizes, all the weights tend to 1 and so asymptotically no observations are removed or downweighted. We call the final estimator based on this method the *weighted maximum likelihood estimator* (WML).

4.1.1 Adaptive cut-off

In the context of regression with asymmetric errors, Marazzi and Yohai (2004) propose a method to determine an adaptive cut-off based on the distribution of the negative log-likelihood of the residuals calculated with the initial estimate, and to reject observations with a lower likelihood than the cut-off value. They used the log-likelihood so that the correction of the final estimator is independent from the distribution of the covariates. In the framework of density estimation this is not an issue and one can apply the method directly to the distribution of the data, and reject the observations which are larger than the cut-off¹. Start with a fixed cut-off η , defined as a large quantile of the initial model, and let F_n be the empirical cdf and F_{β_1} be the cdf of the initial model. The adaptive cut-off t_n is determined by comparing the tails of F_n and F_{β_1} . Let $F_{n,t}$ denote F_n truncated at t , i.e.

$$F_{n,t}(x) = \begin{cases} F_n(x)/F_n(t) & \text{if } x \leq t, \\ 1 & \text{otherwise.} \end{cases} \quad (4.1)$$

¹One could nevertheless choose to impose a cut-off on negative log-likelihoods. This would then correspond to a lower and an upper cut-off on the observations, thus protecting against low outliers as well as large ones. This represents only a small modification of the method exposed hereafter, which concentrates on large outliers. The second rejection method also protects against low outliers.

t_n is the largest t for which $F_{n,t}(x) \geq F_{\beta_1}(x)$ for all $x \geq \eta$, i.e.

$$t_n = \sup\{t \mid F_{n,t}(x) \geq F_{\beta_1}(x) \text{ for all } x \geq \eta\}.$$

Note that t_n is always greater or equal to η , and one could consider defining the cut-off independently of η , i.e. using

$$t_n^* = \sup\{t \mid F_{n,t}(x) \geq F_{\beta_1}(x) \text{ for all } x > 0\}$$

instead of t_n . However, Marazzi and Yohai (2004) report simulation results with a cut-off defined analogously to t_n^* , indicating that the value of the cut-off was often too low for “clean” samples. They advised to keep the parameter η in the definition of the cut-off to ensure high efficiency in small samples.

On the other hand, this has the drawback that contaminations at positions lower than η cannot be eliminated. The method proposed in the next section allows to do so, without lessening the small sample efficiency of the final estimator. Figure 4.1 illustrates the determination of the adaptive cut-off.

Once the cut-off has been determined, we define weights

$$\omega_d(x) = I(x \leq t_n)$$

where $I(x \leq t_n)$ is the indicator function for the set $\{x : x \leq t_n\}$, and we reject the observations such that $\omega_d(x) = 0$.

In the context of regression with asymmetric errors, Marazzi and Yohai (2004) proved that this method yields a cut-off which is asymptotically infinite at the model, so that no observations are removed. They also show that the bdp of the final estimator calculated with the remaining observations is not less than the bdp of the initial estimator. Finally they show that the influence function (IF) at the model is equal to the IF of the MLE, which strongly suggests full asymptotic efficiency. We conjecture that these properties also hold in the discrete density estimation setting. Simulations support

this conjecture. We do not give proofs for this estimator (the WMLc), which we consider mainly for comparison with the WML, based on the outlier rejection method presented in the next section.

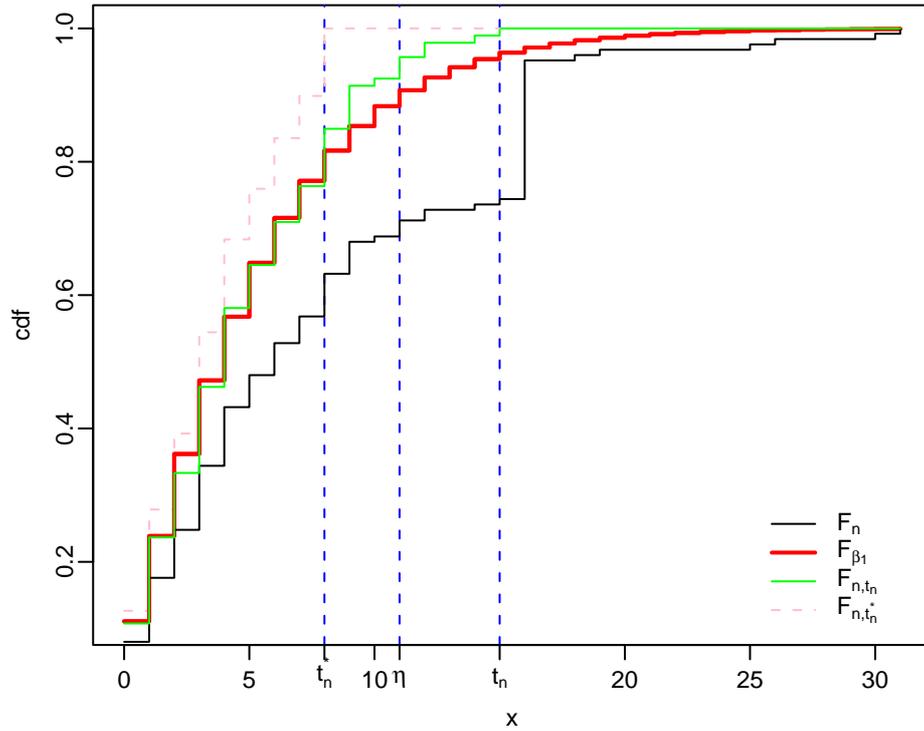


Figure 4.1: Illustration of the adaptive cut-off method, with point contamination at $x = 16$. η corresponds to the 0.9 quantile of m_{β_1} . We see that $t_n = 15$ correctly eliminates the outlier. $t_n^* = 8$ is too low and eliminates many observations which are in good agreement with the model.

4.1.2 Adaptive weights

The outlier rejection method we propose attributes an adaptive downweighting factor to each sample space element. Like the adaptive cut-off, the adaptive weights are determined by comparing the empirical distribution to the

distribution under the initial model. However, each sample space element is considered individually, and attributed a downweighting factor $\omega_d(x)$, which we propose to define in the following way:

$$\omega_d(x) = W(A(x, n, d)), \quad (4.2)$$

where

$$A(x, n, d) = f(n) \frac{d(x) - m_{\beta_1(d)}(x)}{\sqrt{m_{\beta_1(d)}(x)(1 - m_{\beta_1(d)}(x))}}, \quad (4.3)$$

W is a decreasing function with $W(x) = 1$ if $x \leq a$ for some $a \geq 0$ and $W(x) = 0$ if $x > b$ for some $b \geq a$, n is the sample size and f is a bounded positive increasing function.

The idea behind the form of $A(x, n, d)$ is that if $f(n) = \sqrt{n}$ and if $\beta_1(d)$ is consistent, then $A(x, n, d)$ has an approximate $\mathcal{N}(0, 1)$ distribution for large n at the model, which provides a benchmark to set the constants a and b . However, we want to build adaptive weights, i.e. weights which are asymptotically equal to 1 at the model, so that we do not reject or downweight any observation. This is why the function f has to be bounded: if β_1 converges to β in probability, which we suppose to be the case, then, under continuity of the model in β , $A(x, n, d)$ converges to 0 in probability, and hence the weights converge to 1. We therefore propose to define f as follows:

$$f(n) = \begin{cases} \sqrt{n} & \text{if } n \leq n_{\max} \\ \sqrt{n_{\max}} & \text{if } n > n_{\max} \end{cases} \quad (4.4)$$

for some maximum sample size n_{\max} .

The constants a , b and n_{\max} provide a trade-off between robustness and finite sample efficiency of the final estimator. In fact, they regulate the threshold on the standardized difference $\frac{d(x) - m_{\beta_1(d)}(x)}{\sqrt{m_{\beta_1(d)}(x)(1 - m_{\beta_1(d)}(x))}}$, above which the contribution of a sample space element x gets downweighted or removed in the calculation of the final estimator. The lower the threshold, the higher the robustness and the lower the efficiency, and conversely.

The choice of these constants can be simplified by using “hard” rejection weights, i.e. imposing $a = b$, so that sample space elements are either kept or removed completely. The value of $a = b$ has a quantile interpretation, which provides a natural guide for its choice. As to the choice of n_{\max} , it is done by numerical investigation and asymptotic bias analysis in the model of interest. In chapters 6, 7 and 8, we show that the choice $a = b = 3.5$ and $n_{\max} = 200$ provides very satisfactory results in the NB model, in terms of asymptotic robustness, finite sample mean square error and in applications. The case $a \neq b$ has not been investigated, yet we show with an example in section 8.2 that it has some promising properties.

Finally lets us comment the shape of the weight function W . Firstly, it ranges between 0 and 1, in order to get more easily interpretable weights $\omega_d(x)$. Indeed, these weights provide a diagnostic of outlyingness of the proportion of observations at x . Secondly, $W(x) = 1 \forall x < 0$, which implies that only positions with an excessive proportion of observations compared to the initial model can be downweighted. With this definition, we can be protected only against outliers, not against inliers, but as already mentioned in section 3 our main focus in this thesis is resistance against outliers; adding resistance against inliers would be at the price of sacrificing some more finite sample efficiency.

4.2 The final estimator

Suppose the weights $\omega_d(x)$ have been calculated with either of the two methods exposed in the previous section. In this section we define a method to reduce the influence of the outliers in the maximum likelihood equations. Let $s_i(x, \beta) = \frac{\partial}{\partial \beta^i} \log m_\beta(x)$, $i = 1, \dots, p$, be the score functions corresponding to the model m_β , where p is the dimensionality of β . The maximum likelihood equations are

$$\sum d(x) s_i(x, \beta) = 0, \quad i = 1, \dots, p. \quad (4.5)$$

We propose to consider instead the following weighted likelihood equations:

$$\sum \left(\frac{d(x)}{\sum d(x)\omega_d(x)} - \frac{m_\beta(x)}{\sum m_\beta(x)\omega_d(x)} \right) \omega_d(x) s_i(x, \beta) = 0, \quad i = 1, \dots, p. \quad (4.6)$$

If the data are generated by a model m_{β_0} , then all the weights are asymptotically equal to 1 and (4.6) becomes equivalent to (4.5), so that the estimator is asymptotically equal to the MLE and thus Fisher-consistent. If the data follow a contaminated model

$$m_{\beta_{0j}}(x) = (1 - \epsilon)m_{\beta_0}(x) + \epsilon\chi_{x_j}(x),$$

and that $\omega_{m_{\beta_{0j}}}(x_j)$, the asymptotic weight at $x = x_j$, is zero, then the solution to (4.6) is asymptotically β_0 , so the asymptotic bias at the contaminated model is also zero. Of course, the same is true in the presence of a multiple contamination as in Proposition 4 if the weights of all contaminated positions are zero asymptotically.

Another feature which makes equations (4.6) appealing is their asymptotic correspondence with the maximum likelihood equations of a conditioned model. Let us write $\omega_\infty(x)$ for the asymptotic weights. Suppose again we are in a situation where the weights of all contaminated positions - if any - are asymptotically zero. Then the equations (4.6) are asymptotically equivalent to the ML equations of the uncontaminated conditioned model

$$\check{m}_\beta(x) = \frac{\omega_\infty(x)m_\beta(x)}{\sum \omega_\infty(x)m_\beta(x)},$$

where the weights $\omega_\infty(x)$ are fixed. Indeed, the β -dependent terms of the corresponding log-likelihood are

$$\log m_\beta(x) - \log \left(\sum \omega_\infty(x)m_\beta(x) \right),$$

and so the asymptotic value of the log-likelihood for an i.i.d. sample generated with \check{m}_{β_0} , for some parameter value β_0 , is

$$\frac{\sum \omega_\infty(x)m_{\beta_0}(x) \log m_{\beta_0}(x)}{\sum \omega_\infty(x)m_{\beta_0}(x)} - \log \left(\sum \omega_\infty(x)m_{\beta_0}(x) \right). \quad (4.7)$$

It is easy to check that differentiating (4.7) with respect to β yields equations (4.6) with ω_∞ instead of ω_d and m_{β_0} instead of d . This asymptotic correspondence will be useful in the next chapter for the demonstration of a statement about the asymptotic bdp of the final estimator.

Now, although the correspondence is valid only asymptotically (and if all contaminated positions have 0 asymptotic weights), the differentiation of

$$\frac{\sum \omega_d(x)d(x) \log m_\beta(x)}{\sum \omega_d(x)d(x)} - \log \left(\sum \omega_d(x)m_\beta(x) \right) \quad (4.8)$$

with respect to β yields equations (4.6), regardless of the weights and of the contamination. This provides a criterion to choose the right solution in case (4.6) has multiple roots. Thus, we define our final estimate as the value of β that maximizes (4.8), or equivalently that minimizes the *negative weighted log-likelihood*²

$$\text{wl}(d, m_\beta) = \log \left(\sum \omega_d(x)m_\beta(x) \right) - \frac{\sum \omega_d(x)d(x) \log(m_\beta(x))}{\sum \omega_d(x)d(x)}. \quad (4.9)$$

When the weights are obtained with the “adaptive weights” method described in section 4.1.2, we call the estimator the *weighted maximum likelihood estimator* (WML). When they are defined via the “cut-off” method, we call the estimator the WMLc.

In the following chapters, we concentrate on the WML. In chapter 5, we show that the WML has a breakdown point at least as high as the bdp of the

²It is interesting to note that if we use “hard truncation” weights, i.e. weights that can be either 0 or 1, so that we are just removing some observations and keeping the others, then minimizing (4.9) actually corresponds to fitting the model

$$\check{m}_\beta(x) = \frac{\omega_d(x)m_\beta(x)}{\sum \omega_d(x)m_\beta(x)}$$

to the remaining observations by maximum likelihood (with the weights $\omega_d(x)$ fixed). This property will be used in the next chapter to prove a statement about the finite sample bdp of the final estimator.

initial estimator. In section 6.1, we show that its influence function (IF) at the model is the same as the IF of the MLE, and give some arguments which suggest that it is asymptotically normal. In section 6.2, we show with several examples that it has a low asymptotic bias under point contamination. We also show that the bias is generally much lower than the bias on the MDEs at the same contamination rate (actually, it is exactly zero as soon as the rate exceeds a certain model dependent threshold). In chapter 7, we present simulation results which show that the WML has a much lower root mean square error under contamination than all the considered initial estimators, and that the same is true at the “clean” model for 4 initial estimators out of 5. In section 6.2 and in chapter 7, we also present the results for the WMLc, for comparison.

When the weights are obtained with the “adaptive cut-off” method of section 4.1.1, it appears (see chapter 7) that the final estimator (the WMLc) has quite weak performances when a contamination is present close to τ , the quantile of m_β of the same level as η for m_{β_1} (see section 4.1.1). In that case, the performances can be enhanced by solving analogous estimating equations to (4.6), but where we replace β by β_1 in the second term, thus solving

$$\sum \frac{d(x)}{\sum d(x)\omega_d(x)}\omega_d(x)s_i(x, \beta) = \sum \frac{m_{\beta_1}(x)}{\sum m_{\beta_1}(x)\omega_d(x)}\omega_d(x)s_i(x, \beta_1), \quad (4.10)$$

$$i = 1, \dots, p.$$

for β . This seems to mitigate the effect of outliers at positions lower than τ , by reducing the flexibility of the estimator with a fixed right hand side in the equations. We call the estimator obtained by solving (4.10) the *truncated maximum likelihood estimator* (TML). Nevertheless, the TML still shows rather poor performances under this type of contaminations, and this is also visible on the asymptotic bias curves in section 6.2.2.

In the absence of contamination, the TML is Fisher-consistent, since the initial estimator is. In that situation, the TML generally improves the per-

formances of the initial estimators (it fails to do so for one of them, like the WML). But again, is outperformed by the WML.

The problem with the TML (and the WMLc) is that it acts on the whole tail of the distribution at once. If an outlier is present close to η , but at a larger position, then all observations which are larger than the outlier are removed, even if their proportion is not too large relative to the model. This results in a greater loss in efficiency than would be necessary to get rid of the outlier. Conversely, if an outlier is present at a lower position than η , it is not removed, as the adaptive cut-off can never be lower than η . Hence the bad performances of the TML under this kind of contaminations.

The discrete setting offers quite naturally the possibility to act on each sample space element separately, thus getting a more flexible outlier rejection procedure which allows to reduce the influence of departures from the model at any position, without rejecting abusively observations which are in agreement with the model.

Remark C. If the distribution is continuous and we want to define adaptive weights, we have to group the data into categories, and there could be several ways to do it. A proposal is sketched in section 10.1.

Chapter 5

Breakdown point

In this chapter, we establish that the breakdown point of the WML is at least as high as the the bdp of the initial estimator, under some conditions on the model and the weight function. As in chapter 3, the term “breakdown point” refers to the quantity defined in Remark B. All proofs are given for the case of contamination with one single outlier sequence. The extension of the results to contaminations as in Proposition 4 is then straightforward.

Section 5.1 addresses the asymptotic bdb, and section 5.2 concerns the finite sample bdp. In both cases, a stronger result is established for the NB model.

As presented in section 4.2, the WML is obtained by minimization of the quantity

$$\text{wl}(d, m_\beta) = \log \left(\sum \omega_d(x) m_\beta(x) \right) - \frac{\sum \omega_d(x) d(x) \log(m_\beta(x))}{\sum \omega_d(x) d(x)}, \quad (5.1)$$

where the weights $\omega_d(x)$ are calculated from the initial estimate $\beta_1(d)$ as

$$\omega_d(x) = W \left(f(n) \frac{d(x) - m_{\beta_1(d)}(x)}{\sqrt{m_{\beta_1(d)}(x)(1 - m_{\beta_1(d)}(x))}} \right),$$

where W is a decreasing function with $W(x) = 1$ if $x \leq a$ for some $a \geq 0$

and $W(x) = 0$ if $x > b$ for some $b \geq a$, n is the sample size and f is given by

$$f(n) = \begin{cases} \sqrt{n} & \text{if } n \leq n_{\max} \\ \sqrt{n_{\max}} & \text{if } n > n_{\max} \end{cases}$$

for some maximum sample size n_{\max} .

5.1 Asymptotic breakdown point

Let $m_{\beta_0}(x)$ be a member of the considered family of probability densities, and let

$$m_{\beta_{0j}}(x) = (1 - \epsilon)m_{\beta_0}(x) + \epsilon\chi_{x_j}(x)$$

be the corresponding contaminated model with $\{x_j\}$ an outlier sequence and χ_{x_j} the indicator function for x_j . Let $\omega_j(x)$ be the weights defined from $m_{\beta_{0j}}$.

Let ϵ_1 be the asymptotic bdp of the initial estimator and impose $\epsilon < \epsilon_1$.

Assumption 6. $\sup_{\beta \in \mathbf{B}} m_{\beta}(x) \rightarrow 0$ as $x \rightarrow \infty$ for any compact set of parameter values \mathbf{B} .

This corresponds to an intuitive assumption on the model structure that, as x gets large, it becomes less and less likely to have arisen from a model distribution with β close to any finite value ν (see Lindsay (1994)). Now consider

$$\omega_j(x_j) = W \left(\frac{\sqrt{n_{\max}}(1 - \epsilon)m_{\beta_0}(x_j) + \epsilon - m_{\beta_1(m_{\beta_{0j}})}(x_j)}{\sqrt{m_{\beta_1(m_{\beta_{0j}})}(x_j)(1 - m_{\beta_1(m_{\beta_{0j}})}(x_j))}} \right). \quad (5.2)$$

Since $\epsilon < \epsilon_1$, the sequence $\{|\beta_1(m_{\beta_{0j}})|\}$, $j = 1, 2, \dots$, is bounded and thus Assumption 6 and the fact that $W(x) = 0$ for $x > b$ imply that

$$\exists x_0 \text{ such that } \omega_j(x_j) = 0 \forall x_j > x_0. \quad (5.3)$$

Theorem 7. Let ϵ_{WML} be the asymptotic bdp of the WML. Under Assumption 6, $\epsilon_{\text{WML}} \geq \epsilon_1$. Moreover, if $\epsilon < \epsilon_1$, then $\lim_{x_j \rightarrow \infty} \text{WML}(m_{\beta_{0j}}) = \beta_0$.

Proof. As soon as $x_j > x_0$, the solution of the minimization problem is $\beta = \beta_0$. Indeed, if $\omega_j(x_j) = 0$, the asymptotic negative weighted log-likelihood becomes

$$\text{wl}(m_{\beta_0j}, m_\beta) = \log \left(\sum \omega_j(x) m_\beta(x) \right) - \frac{\sum \omega_j(x) m_{\beta_0}(x) \log(m_\beta(x))}{\sum \omega_j(x) m_{\beta_0}(x)},$$

which is exactly equal to the asymptotic negative log-likelihood for the family of distributions

$$\check{m}_\beta(x) = \frac{\omega_j(x) m_\beta(x)}{\sum \omega_j(x) m_\beta(x)}$$

at the uncontaminated model $\check{m}_{\beta_0}(x) = \frac{\omega_j(x) m_{\beta_0}(x)}{\sum \omega_j(x) m_{\beta_0}(x)}$, where the weights ω_j are considered fixed (see section 4.2). Since the log-likelihood function of the model $\check{m}_\beta(x)$,

$$\check{l}(x; \beta) = \log(m_\beta(x)) - \log \left(\sum \omega_j(x) m_\beta(x) \right),$$

satisfies

$$E_{\check{\beta}}(\nabla_\beta l(x; \check{\beta})) = 0$$

for any parameter value $\check{\beta}$, the corresponding maximum likelihood estimator is Fisher-consistent and so the minimum of $\text{wl}(m_{\beta_0j}, m_\beta)$ is attained at $\beta = \beta_0$. ■

In the case of the NB model, a stronger statement can be proved.

Theorem 8. In the NB model, the asymptotic bdp of the WML is equal to 1 as soon as the asymptotic bdp of the initial estimator is non-zero.

Proof. Consider the argument of the weight function in (5.2)

$$A(x_j, n, m_{\beta_0j}) = \sqrt{n_{\max}} \frac{(1 - \epsilon) m_{\beta_0}(x_j) + \epsilon - m_{\beta_1(m_{\beta_0j})}(x_j)}{\sqrt{m_{\beta_1(m_{\beta_0j})}(x_j) (1 - m_{\beta_1(m_{\beta_0j})}(x_j))}}. \quad (5.4)$$

$\omega_j(x_j)$ will be 0 if $A(x_j, n, m_{\beta_0 j}) > b$, so equation (5.4) implies that if the contamination rate ϵ is larger than the upper limit

$$B(x_j, \epsilon, \beta_0) = \frac{1}{1 - m_{\beta_0}(x_j)} \left[\frac{b}{\sqrt{n_{\max}}} \sqrt{m_{\beta_1(m_{\beta_0 j})}(x_j)(1 - m_{\beta_1(m_{\beta_0 j})}(x_j))} + m_{\beta_1(m_{\beta_0 j})}(x_j) - m_{\beta_0}(x_j) \right], \quad (5.5)$$

the contribution of x_j will be suppressed. (Note that B depends on ϵ through $\beta_1(m_{\beta_0 j})$.)

Since in the NB model we have that

$$\sup_{\beta \in \Theta} m_{\beta}(x_j) \rightarrow 0 \text{ as } x_j \rightarrow \infty \quad (5.6)$$

uniformly in β over the whole parameter space Θ (see the proof of Theorem 5 in Appendix B), we have

$$B(x_j, \epsilon, \beta) \rightarrow 0 \text{ as } x_j \rightarrow \infty \quad (5.7)$$

uniformly in β , regardless of ϵ .

From Theorem 7, we have that $\epsilon_{\text{WML}} \geq \epsilon_1$, so if $\epsilon < \epsilon_1$ no breakdown occurs for $x_j \rightarrow \infty$. If $\epsilon \geq \epsilon_1$, (5.7) implies that

$$\exists \tilde{x} \text{ such that } \forall x_j > \tilde{x}, \forall \beta_0, B(x_j, \epsilon_1, \beta_0) < \epsilon_1.$$

As soon as $x_j > \tilde{x}$, we have $\omega_j(x_j) = 0$, and then, from the proof of Theorem 7, the WML is equal to β_0 , so that again no breakdown occurs for $x_j \rightarrow \infty$.

■

5.2 Finite sample breakdown point

Let $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ be an observed sample and let $d(x)$ be the proportion of observations equal to x . Define $d_j(x)$ as

$$d_j(x) = (1 - \epsilon)d(x) + \epsilon\chi_{x_j},$$

where $\{x_j\}$ is an outlier sequence. Let $x_M = \max(\mathbf{x}) < \infty$.

Let ϵ_1^* be the finite sample bdp of the initial estimator and impose $\epsilon < \epsilon_1^*$.

Analogous statements as in the asymptotic case can be proved in the finite sample case, but two additional assumptions are needed.

Assumption 9. The weight function $W(x)$ is a “step function”

$$W(x) = I(x \leq b) \quad (5.8)$$

where $I(x \leq b)$ is the indicator function for the set $\{x : x \leq b\}$. This corresponds to setting $a = b$ in the definition of the weight function, and the obtained weights are referred to as “hard” weights.

Assumption 10. Let $\omega(x)$ be a function from the sample space into the two-element set $\{0, 1\}$. Let $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ be a finite sample such that $\max(\mathbf{y})$ is finite. Then the maximum likelihood estimate of β in the model

$$\check{m}_\beta(x) = \frac{\omega(x)m_\beta(x)}{\sum \omega(x)m_\beta(x)},$$

calculated with the observations in \mathbf{y} for which $\omega(y_i) \neq 0$ is finite.

This assumption is reasonable since the sample \mathbf{y} does not contain large outliers.

Theorem 11. Let ϵ_{WML}^* be the finite sample bdp of the WML. Under Assumptions 6, 9 and 10, $\epsilon_{\text{WML}}^* \geq \epsilon_1^*$.

Proof. Assumption 6 and similar arguments as in the asymptotic case imply that

$$\exists x_0^* \text{ such that } \omega_{d_j}(x_j) = 0 \forall x_j > x_0^*. \quad (5.9)$$

As soon as $x_j > x_0^*$, we have

$$\text{wl}(d_j, m_\beta) = \log \left(\sum \omega_{d_j}(x)m_\beta(x) \right) - \frac{\sum \omega_{d_j}(x)d(x) \log(m_\beta(x))}{\sum \omega_{d_j}(x)d(x)},$$

which in the case of hard weights is exactly equal to the negative log-likelihood of the model

$$\check{m}_\beta(x) = \frac{\omega_{d_j}(x)m_\beta(x)}{\sum \omega_{d_j}(x)m_\beta(x)}$$

at the sample consisting of the observations in \mathbf{x} for which $\omega_{d_j}(x) \neq 0$, with the weights $\omega_{d_j}(x)$ considered fixed (see footnote 2 in section 4.2)). Since $\max(\mathbf{x}) = x_M$ is finite, the theorem is proved. ■

Again, a stronger statement can be proved in the case of the NB model.

Theorem 12. Under Assumptions 9 and 10, in the NB model, the finite sample bdp of the WML is equal to 1 as soon as the finite sample bdp of the initial estimator is non zero.

Proof. The proof is completely analogous to the proof of Theorem 8. Note that Assumption 6 is no longer needed as the property (5.6) of the NB model is stronger. ■

Chapter 6

Asymptotic behavior

6.1 Influence function

We denote by $IF(x_0, Z, g)$ the influence function at x_0 of a functional $Z(g)$ when the data are distributed according to the probability distribution g , and we use the abbreviation I_Z^g for $IF(x_0, Z, g)$. Consider as usual a model m_β . The WML estimate of parameter β is noted $\hat{\beta}$ and its value at distribution g is noted $\hat{\beta}_g$, and the latter convention is applied to all other functionals of the probability distribution.

Theorem 13. The influence function of the WML is given by

$$IF(x_0, \hat{\beta}, g) = M^{-1}\mathbf{c}(x_0), \quad (6.1)$$

where

$$M = \frac{1}{S_m^2} \left[S_m \sum \mathbf{s}(x, \hat{\beta}_g) \mathbf{s}(x, \hat{\beta}_g)^t \omega_g(x) m_{\beta_g}(x) - \sum T_g(x, \hat{\beta}_g) \omega_g(x) m_{\beta_g}(x) \right] \\ - \sum \left(\frac{g(x)}{S_g} - \frac{m_{\hat{\beta}_g}(x)}{S_m} \right) \omega_g(x) H(x, \hat{\beta}_g),$$

where

$$S_m = \sum \omega_g(x) m_{\hat{\beta}_g}(x),$$

$$S_g = \sum \omega_g(x) g(x),$$

$\mathbf{s}(x, \cdot)$ is the score function for model $m_\beta(x)$,

$H(x, \cdot)$ is the Hessian matrix of $\log m_\beta(x)$,

$$T_g(x, \beta) = \mathbf{u}_g(\beta) \mathbf{s}(x, \beta)^t, \text{ with } \mathbf{u}_g(\beta) = \sum \nabla m_\beta(x) \omega_g(x),$$

and

$$\begin{aligned} \mathbf{c}(x_0) &= \frac{1}{S_g^2} \left(S_g \omega_g(x_0) \mathbf{s}(x_0, \hat{\beta}_g) - \omega_g(x_0) \sum g(x) \omega_g(x) \mathbf{s}(x, \hat{\beta}_g) \right) \\ &\quad - \frac{1}{S_g^2} \left(\sum d(x) \omega_g(x) \mathbf{s}(x, \hat{\beta}_g) \right) \sum d(x) I_\omega^g(x) \\ &\quad + \frac{1}{S_m^2} \left(\sum m_{\hat{\beta}_g}(x) \omega_g(x) \mathbf{s}(x, \hat{\beta}_g) \right) \sum m_{\hat{\beta}_g}(x) I_\omega^g(x) \\ &\quad + \sum \left(\frac{g(x)}{S_g} - \frac{m_{\hat{\beta}_g}(x)}{S_m} \right) I_\omega^g(x) \mathbf{s}(x, \hat{\beta}_g). \end{aligned}$$

The influence function of the weight function $\omega(\cdot)$ is given by

$$\begin{aligned} IF(x_0, \omega(\cdot), g) &= \frac{W' \left(f(n) \frac{g(x) - m_{\beta_1(g)}(x)}{\sqrt{m_{\beta_1(g)}(x)(1 - m_{\beta_1(g)}(x))}} \right) f(n)}{\sqrt{m_{\beta_1(g)}(x)(1 - m_{\beta_1(g)}(x))}} \\ &\quad \left[\nabla m_{\beta_1(g)}(x) \cdot I_{\beta_1}^g \left(\frac{1}{2} (2m_{\beta_1}(x) - 1)(g(x) - m_{\beta_1}(x)) - 1 \right) \right. \\ &\quad \left. + \chi_{x_0}(x) - g(x) \right], \end{aligned}$$

If we are using a MDE as initial estimator β_1 , its influence function $I_{\beta_1}^g$ can be found in Lindsay (1994).

Proof. The proof is straightforward (but lengthy) differentiation of the estimating equations (4.6). ■

If the distribution g is a model point m_β , then the consistency of the initial estimator and of the WML imply that $\beta_g = \beta_1 = \beta$, $w_g(x) \equiv 1$, $m_{\beta_g}(x) \equiv g(x)$, and then (6.1) becomes equal to

$$i(\beta)^{-1}\mathbf{s}(x_0, \beta),$$

where $i(\beta) = E(\mathbf{s}(X, \beta)\mathbf{s}(X, \beta)^t)$ is the Fisher information matrix, with X distributed according to m_β . Thus, the WML has the same influence function as the MLE at the model, which strongly suggests full asymptotic efficiency. A proof of asymptotic normality is still lacking, but simulations support this conjecture. Moreover, a theorem by Rao (1961) (his Lemma 3 p.539) states that in the multinomial model with a *finite* number of cells, any estimator which has the same influence function as the MLE is asymptotically normal (and therefore fully efficient). Finally, if we use “hard rejection weights”, then footnote 2 p.46 shows that the WML is closely related to a maximum likelihood estimator, which strengthens our confidence that it is asymptotically normal.

If we suppose that the WML is asymptotically normal, its asymptotic covariance matrix at the model m_β is given by (Hampel et al., 1986)

$$\sum IF(x, \beta, m_\beta)IF(x, \beta, m_\beta)^t m_\beta(x).$$

In practice, the covariance matrix of the WML can be estimated either by

$$\frac{1}{n} \sum IF(x, \hat{\beta}_d, m_{\hat{\beta}_d})IF(x, \hat{\beta}_d, m_{\hat{\beta}_d})^t m_{\hat{\beta}_d}(x) \quad (6.2)$$

or by

$$\frac{1}{n} \sum IF(x, \hat{\beta}_d, d)IF(x, \hat{\beta}_d, d)^t d(x), \quad (6.3)$$

where the observed frequencies are given by $d(x)$ and n is the sample size. We shall see with two examples in chapter 8 that formula (6.2) works well for the WML and the more efficient MDEs (NE, linNE, Hellinger) already

for moderate sample sizes, but that it underestimates the variance of the less efficient MDEs (Neyman, “ $\lambda = 3$ ”).¹

Finally, note that if we use “hard rejection weights”, $IF(x_0, \omega(\cdot), g) \equiv 0 \forall g$ and $\mathbf{c}(x_0)$ simplifies to

$$\mathbf{c}(x_0) = \frac{1}{S_g^2} \left(S_g \omega_g(x_0) \mathbf{s}(x_0, \hat{\beta}_g) - \omega_g(x_0) \sum g(x) \omega_g(x) \mathbf{s}(x, \hat{\beta}_g) \right).$$

6.2 Asymptotic bias under contamination

In this chapter we investigate the asymptotic bias of the WML under point contamination, in the NB model. We consider different contamination positions, excluding contaminations at 0 for the following reasons: In the NB model, there are typically two ways an estimator of the parameters can be caused to breakdown: observations going to infinity and observations accumulating at zero. While both are well handled by the MDEs, the TML and the WML in terms of breakdown point, which is 1 in both cases, the latter is still an issue in terms of bias. As noted in section 3.4.2, for a given contamination level ϵ , the largest bias of the MDEs is often observed for contamination at zero, where the MDEs do not do better than the MLE. Naturally, this shortcoming is transferred to the WML if it is based on a MDE. Thus, in cases where a strong contamination at 0, and only there, is expected, the methods proposed in this thesis are not preferable to the MLE. If one wishes to find a good model for the other data, which is the robust philosophy we are following, one should probably fit a zero-truncated model to the non zero observations, and then our methods are again advisable. In what follows we only consider contaminations at positions different from zero.

Remark D. An asymptotic bias investigation in the NB model, of the type

¹As noted in section 3.4.3, Lindsay (1994) showed that the MDEs also have the same IF as the MLE at the model, and so their covariance matrix can also be estimated with formula (6.2), with the corresponding estimate of β instead of $\hat{\beta}_d$.

presented in this chapter, together with a numerical investigation as in chapter 7, led to the choice $b = 3.5$ and $n_{\max} = 200$ for the tuning constants of the weight function and its argument (see (4.2) and (4.4)). The examples presented in this chapter and in chapters 7 and 8 use this choice.

6.2.1 Maximum asymptotic bias

In this section we consider the maximum bias that can be induced on the WML estimates of the parameters of the negative binomial model by a contamination at position x_j . We show that this bias is quite low, due to the fact that the maximum contamination rate that will not be suppressed by the WML itself is low.

Consider the contaminated model

$$m_{\beta_{0j}}(x) = (1 - \epsilon)m_{\beta_0}(x) + \epsilon\chi_{x_j}(x),$$

where χ_{x_j} is the indicator function for x_j .

In what follows we take m_{β_0} in the NB model and consider $\text{MB}_{\text{WML}}(x_j, \beta_0)$, the maximum asymptotic bias that can be caused to the WML by a contamination at position x_j .

As noted in the proof of Theorem 8, $\omega_{m_{\beta_{0j}}}(x_j)$, the weight at the contamination position, is zero as soon as

$$\epsilon > B(x_j, \epsilon, \beta_0), \quad (6.4)$$

where the upper limit $B(x_j, \epsilon, \beta_0)$ is given by

$$B(x_j, \epsilon, \beta_0) = \frac{1}{1 - m_{\beta_0}(x_j)} \left[\frac{b}{\sqrt{n_{\max}}} \sqrt{m_{\beta_1(m_{\beta_{0j}})}(x_j)(1 - m_{\beta_1(m_{\beta_{0j}})}(x_j))} + m_{\beta_1(m_{\beta_{0j}})}(x_j) - m_{\beta_0}(x_j) \right], \quad (6.5)$$

which depends on ϵ through $m_{\beta_{0j}}$. Thus, a contamination at x_j will contribute to the estimating equation only if

$$\epsilon \leq B(x_j, \epsilon, \beta_0). \quad (6.6)$$

In the NB model, with a MDE as initial estimator, it seems that $\epsilon^*(x_j, \beta_0)$, the largest value of ϵ for which (6.6) holds, is quite low for all positions x_j (except 0) for the choice $b = 3.5$ and $n_{\max} = 200$. A low $\epsilon^*(x_j, \beta_0)$ generally causes $\text{MB}_{\text{WML}}(x_j, \beta_0)$, the maximum asymptotic bias under contamination at position x_j , to be quite low as well.

Figure 6.1 shows the maximum bias on each of the parameters as a function of the contamination position, in different NB models. In these examples, the initial estimator is the minimum negative exponential estimator. The results were very similar with each of the other MDEs we chose to consider in section 3.4.3. In Figure 6.1 it can be seen that in our examples the maximum relative bias under point contamination at $x_j \neq 0$ almost never exceeds 10% and is generally much lower, which indicates that our choice of b and n_{\max} is reasonable in terms of asymptotic robustness.

For comparison purposes, consider the bias $b_{\text{MDE}}(x_j, \epsilon, \beta_0)$ caused to a MDE by a contamination at x_j with rate ϵ , and consider the limit $\epsilon \rightarrow 1$ in the NB model. In that case all considered MDE estimates $(m_{\text{MDE}}, \alpha_{\text{MDE}})$ tend to $(x_j, 0)$, and the same is true for the corresponding TMLs, regardless of β_0 .² The corresponding biases are plotted in Figure 6.1 only for parameter m , and appear as almost vertical lines. The equivalent for α would be a horizontal line at $-\alpha$ (outside the plotting region). Note that the bias of the WML in the limit $\epsilon \rightarrow 1$ is zero in all these examples, as the maximum ϵ that will not be cut by the WML is much lower than 1. This maximal contamination rate is plotted as a function of contamination position in the third column of Figure 6.1. The largest contamination proportion to ever go

²For the MDEs, recall that in the NB model the maximum over $\beta = (m, \alpha)$ of $m_\beta(x_j)$ is attained at the limit $\beta \rightarrow (x_j, 0)$ for all $x_j \neq 0$ (see the proof of Theorem 5 in Appendix B). It is easy to check that if $d(x_j) \rightarrow 1$ the minimum disparity, for all disparities considered, is attained by maximizing $m_\beta(x_j)$. For the TML, supposing the contamination at x_j is not rejected (which is reasonable since it corresponds to the mode of the initial model), we arrive to the same conclusions, again with the arguments exposed in the proof of Theorem 5.

through in our examples is 0.13.

6.2.2 Asymptotic bias for fixed ϵ

In Figures 6.2, 6.3 and 6.4, we give “standard” asymptotic bias curves for two different NB models, $(m, \alpha) = (5, 2/3)$ and $(m, \alpha) = (7, 1.5)$, and different contamination rates ϵ , for point contamination from $x_j = 1$ to $x_j = 75$. As the plots are all quite similar in shape, we chose two very different contamination rates, $\epsilon = 0.1$ and $\epsilon = 0.5$, and 3 different initial estimators, the minimum Hellinger distance (Figure 6.2), the minimum NE (Figure 6.3) and the minimum Cressie-Read disparity with $\lambda = -3$ (Figure 6.4). These initial estimators span the range of resistance to contamination among the 5 we chose to investigate (see Figures 3.3 and 3.4). Again, we used $a = b = 3.5$ and $n_{\max} = 200$. All graphs share the following common characteristics:

- The TML and the WMLc have the same asymptotic bias as the MLE up to the point where the contamination position gets larger than the adaptive cut-off, thus having large biases for low contamination positions.
- When the cut-off is reached, the bias of the WMLc drops directly to zero, as expected (an estimator that minimizes the negative weighted log-likelihood (4.9) has zero asymptotic bias if all contaminated positions get a zero weight). This is not the case for the TML, which is still influenced by the bias of the initial estimate.
- As we move on from a more biased initial estimate (Hellinger) to a more robust one (“ $\lambda = -3$ ”), the point where the “drop down” occurs gets lower, and this is the main difference, as far as asymptotic bias is concerned.
- The bias of the WML is exactly zero as soon as the contamination gets cut. In our examples the only situation where this is not the case

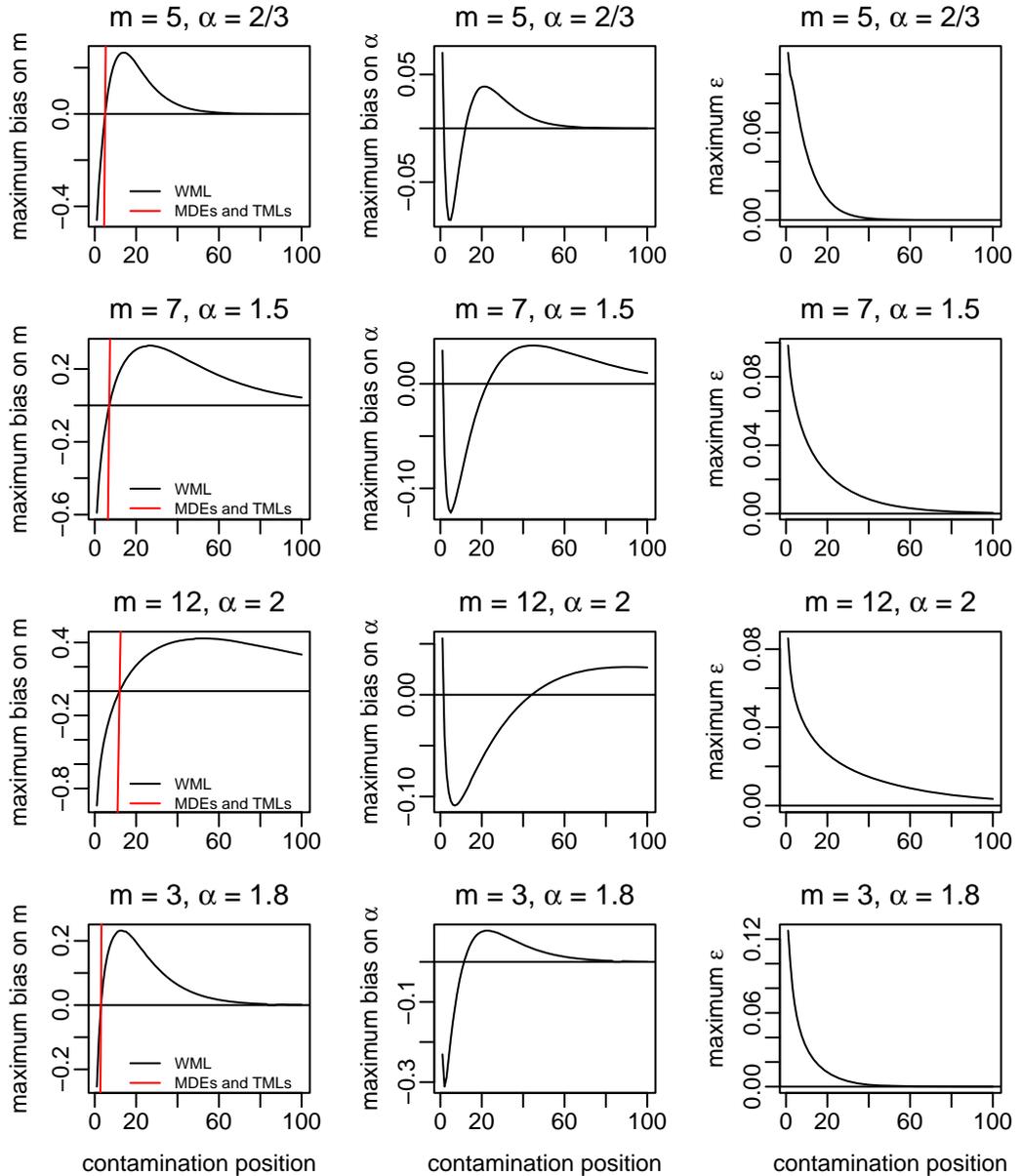


Figure 6.1: The maximum bias that can be caused to the WML estimates of m and α in the NB model, as a function of contamination position, the 0 position being excluded (first two columns). The third column shows the corresponding rates of contamination, which are the maximum rates for which the contamination at the corresponding position is not removed in the calculation of the WML. All examples with weight function parameters $a = b = 3.5$, $n_{\max} = 200$. In red, the analogous curves for the MDEs and the TMLs. In the case of α these curves lie outside the plotting region.

is for contamination at $x_j = 1$ or $x_j = 2$ with $\epsilon = 0.1$ at the model $(m, \alpha) = (5, 2/3)$. In all other cases, the contamination rate is above the upper limit (6.5), consistently with the third column of Figure 6.1.

- In our examples, a change of initial estimator makes almost no difference for the WML. (The only difference is that with the more robust “ $\lambda = -3$ ” initial estimator, the contamination at $x_j = 2$ gets cut.)
- Finally, note that although the graphs for $\epsilon = 0.1$ and $\epsilon = 0.5$ look very similar, there is a big difference in scale.

Coming back to our discussion of section 3.4.3 about the best choice for the initial estimator, we see that the differences in robustness among the MDEs have rather small influence on the robustness of the final estimators, particularly for the WML. Thus, the decisive argument for this choice will be the performances of the estimators in finite samples, to be presented in chapter 7, to which we postpone this question.

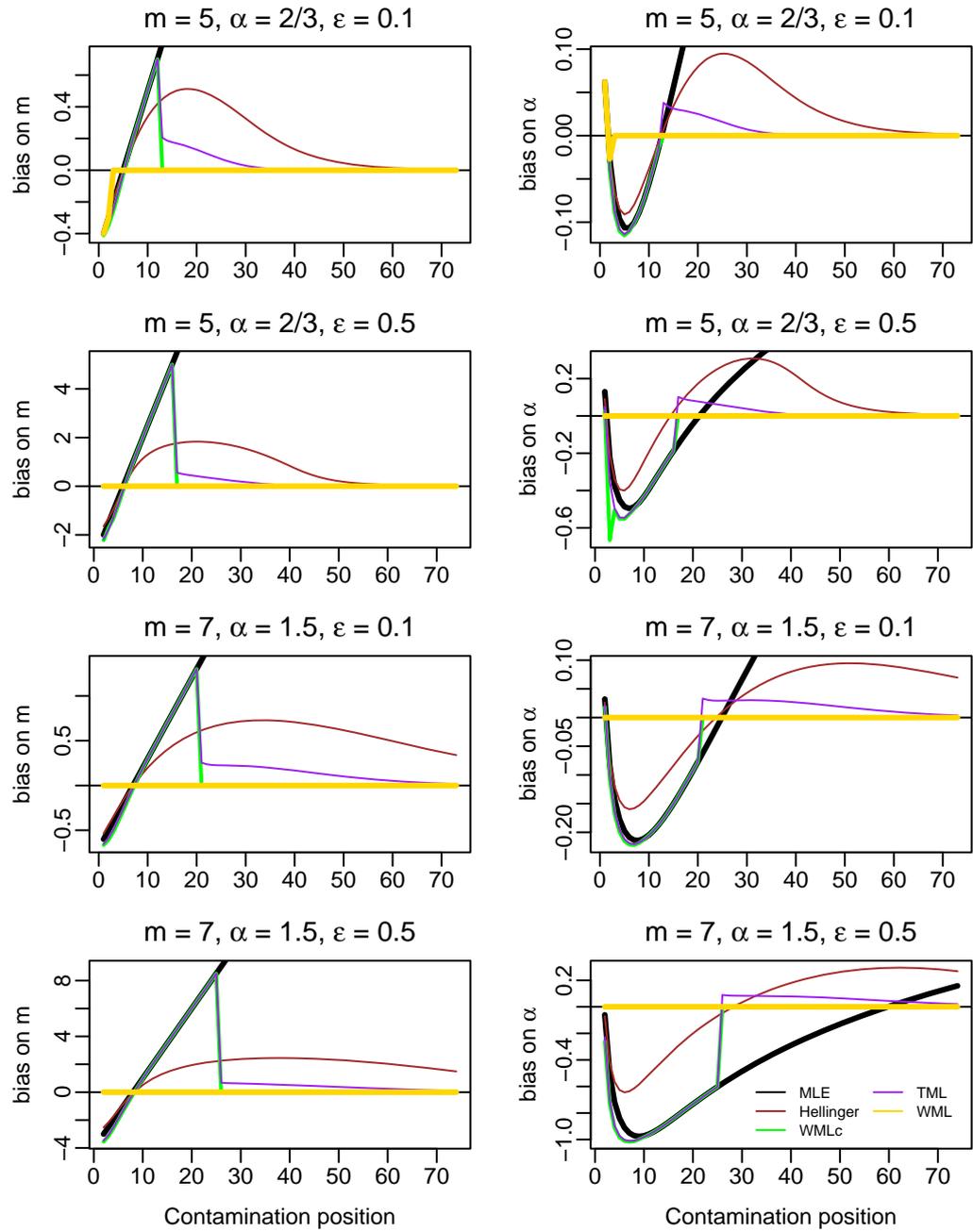


Figure 6.2: Asymptotic bias plots. The initial estimate is minimum Hellinger distance.

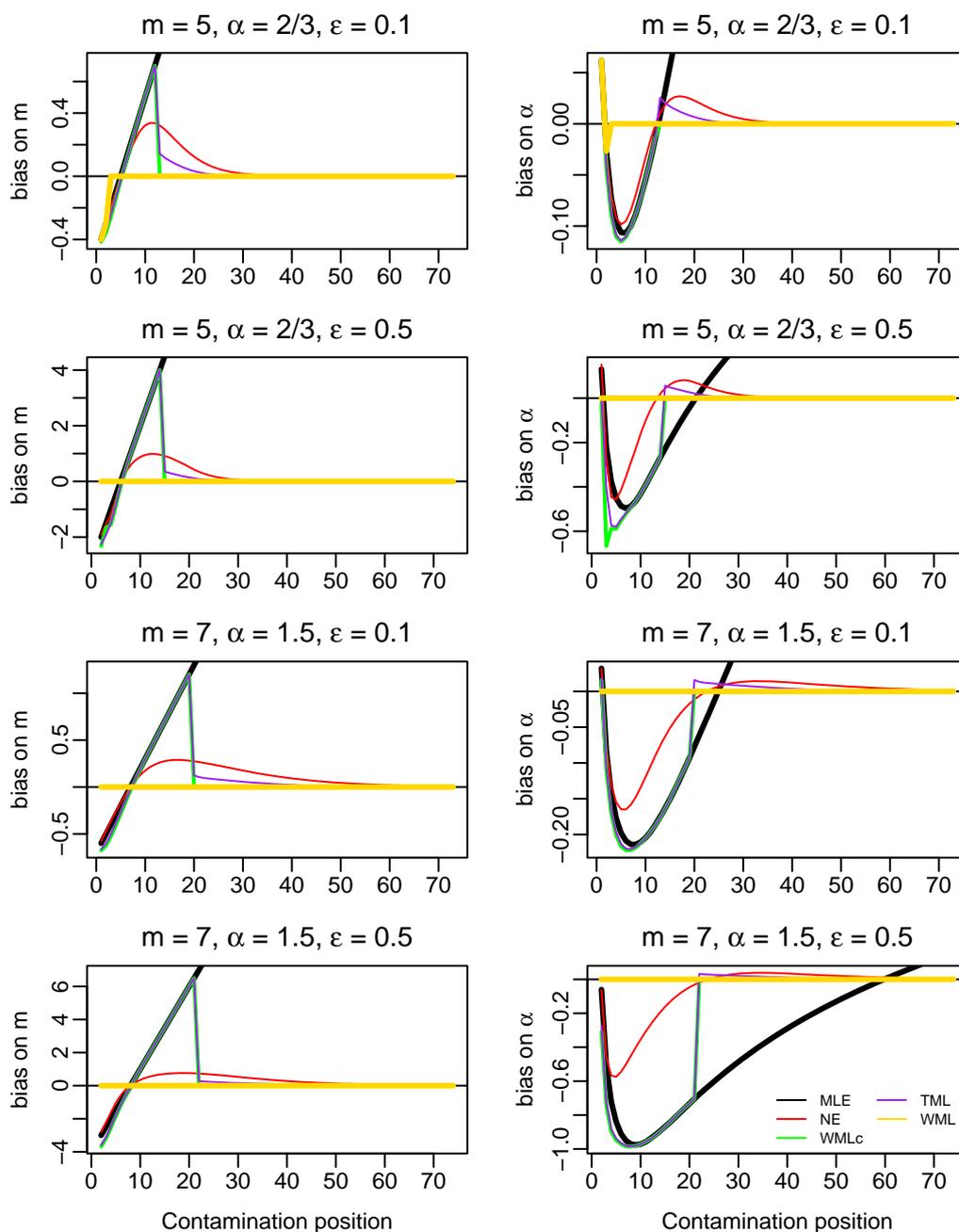


Figure 6.3: Asymptotic bias plots. The initial estimate is minimum NE.

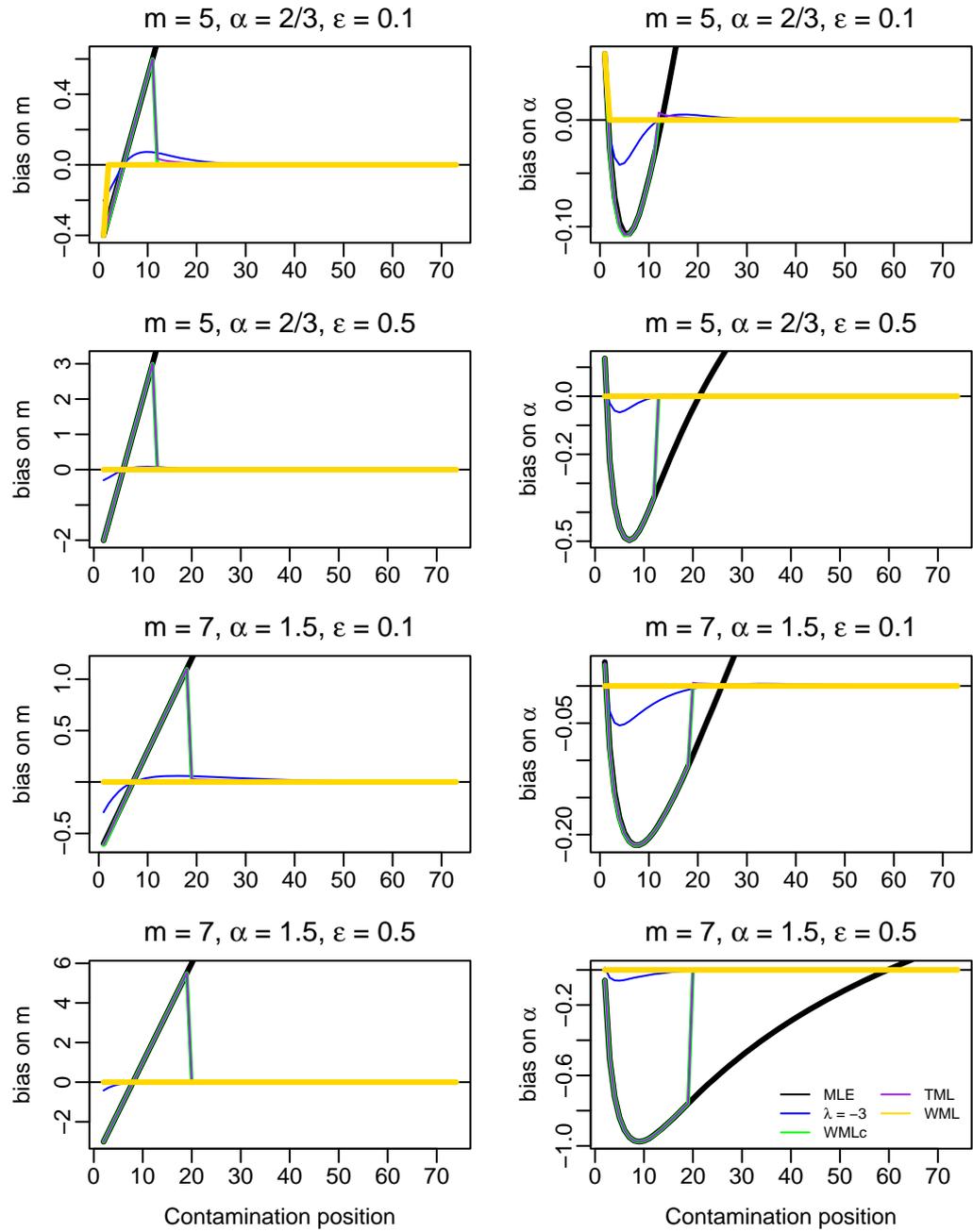


Figure 6.4: Asymptotic bias plots. The initial estimate is minimum Cressie-Read disparity with $\lambda = -3$.

Chapter 7

Empirical results

We did some simulations in order to explore the finite sample behavior of the different estimators. We are interested in the following aspects:

- Efficiency in terms of mean square error (MSE efficiency), relative to the MLE, at the model. To explore this point for different sample sizes, we have run simulations at increasing sample sizes from 100 to 2000, by steps of 100. The results are presented on Figures D.1, D.2 and D.3 in Appendix D. The MSE efficiency of an estimator $\hat{\beta}$ of a parameter β is defined as

$$\frac{\text{MSE}(\beta_{\text{ML}})}{\text{MSE}(\hat{\beta})},$$

where β_{ML} is the maximum likelihood estimator for β .

- “Standard” efficiency at the model for increasing sample sizes. Efficiency is defined as

$$\frac{\text{Var}(\beta_{\text{ML}})}{\text{Var}(\hat{\beta})}.$$

The corresponding results (based on the same simulations as for the previous point) are presented on Figures D.4 and D.5 in Appendix D.

- Bias at the model for increasing samples sizes. The corresponding results (again based on the same simulations as for the first point) are presented on Figures D.6 and D.7 in Appendix D.
- Behavior in contaminated samples. To explore this point we have run simulations at point contaminated models for contamination position ranging from 1 to 50, by steps of 1.¹ Different aspects of the corresponding results are presented on Figures D.8, D.9, D.10, D.11 and D.12 in Appendix D.

All simulations have been performed on the two NB models $(m, \alpha) = (5, 2/3)$ and $(m, \alpha) = (7, 1.5)$, with 500 replications, using the pseudo-random number generator provided by the R software. In all cases, we have calculated the five initial estimators we consider and the corresponding TMLs, WMLc's and WMLs. As the number of simulations is quite high and it would be very time consuming to check all results for possible numerical problems, all the quantities mentioned in this section (MSEs, biases, ...) are in fact trimmed versions, where the most extreme values have been removed².

We now comment these results, using the following conventions:

- The MDEs are referred to as NE, linNE, Hellinger, Neyman and “ $\lambda = -3$ ”.
- These denominations are extended to the corresponding final estimators, when no confusion is possible.

¹contaminations at zero have not been considered, for already exposed reasons (see the comment at the beginning of section 6.2).

²The values exceeding the whiskers of a box-plot have been suppressed before calculation of each of the quantities of interest. The lower whisker is defined as the first quartile minus 1.5 times the interquartile range (IC), and the upper whisker as the third quartile plus 1.5 times the IC.

7.1 Simulations at the model

Although all considered estimators are asymptotically equivalent at the model, they show very different performances in finite samples. Figures D.1 and D.2 show the MSE efficiencies of the estimators. On the two upper panels of both figures, we see that except in the case of the NE, the WMLs show better performances than the MDEs on which they are based. In addition, their MSE efficiencies are very close to each other, compared to the efficiencies of the MDEs. This is in fact a general pattern that will also appear in contaminated sample situations: the choice of the initial estimator does not seem to have a strong influence on the performances of the WML. However, the ordering of the estimators is preserved: for both the MDEs and the WMLs the best estimator is NE, followed by linNE, Hellinger, Neyman and “ $\lambda = -3$ ”. This also is a general pattern which is observed on most of our figures.

This calls for some comments on the MDEs. It appears that the most robust MDEs are the ones that show the weakest performances in finite samples of the sizes considered. This is a consequence of the robustness-efficiency trade-off which we mentioned in section 3.4.3. The Cressie-Read disparities with the lowest values of λ , Neyman and “ $\lambda = -3$ ”, have higher finite sample variance and bias, as predicted by their lower second order efficiency and their shortcomings in the treatment of inliers. However, the poor performances of Hellinger are to be explained solely by the inlier problem, which seems to cause bias. In particular, on Figures D.4 and D.5, it is seen to be the most efficient of all our estimators (even more than the MLE). This contradicts the prediction made by second order efficiency, as the MLE, the linNE and the NE are second order efficient and have higher variances. This type of behavior was also noted by Basu and Sarkar (1994). Moreover, the linNE, designed to be closer to the MLE while keeping the outlier downweighting properties of the NE, has weaker performances than the NE in terms of MSE, although its efficiency is slightly higher (see Figures D.4 and D.5). Thus the

NE appears to present an excellent balance for what concerns second order efficiency and the treatment of inliers; indeed, it is often the best estimator in our examples. We shall see that it also does very well in the presence of contamination, but then it is outperformed by the WML.

One more remark on the MDEs: some of them (“ $\lambda = -3$ ”, Neyman, Hellinger) do not seem to show a convergence of the MSE efficiency to 1. This is probably due to their important finite sample bias (see Figures D.6 and D.7). This phenomenon slows up the convergence. On Figures D.4 and D.5 we see that in terms of “standard” efficiency, the convergence is more visible.

As a matter of fact, finite sample bias seems to be the main drawback of MDE - and of WML - estimation. The bias is particularly large for Hellinger at the smallest sample sizes represented in Figures D.6 and D.7, but in fact all MDEs have large biases as the sample size gets small, and this effect is stronger for more over-dispersed models (i.e. models with a larger α). The WML makes the situation better³, yet it can still have important biases; like for the MDEs, this shortcoming gets stronger as α increases. For example, a simulation with 500 replications for sample size $n = 50$ at the models $(m, \alpha) = (5, 2/3)$ and $(m, \alpha) = (7, 1.5)$ yielded biases on the WML (based on NE) as high as 5% on m and 10% on α for the former model and 13% on m and 17% on α for the latter. (Let us mention however that the biases on the Hellinger MDE were between two and three times larger).

This phenomenon has also been noted, for MDEs, by Basu, Basu, and Chaudhuri (1997), in the context of estimation of the Poisson model. These authors explore the Cressie-Read family in the range $\lambda \in (-1, 0)$, i.e. between The Kullback-Leibler divergence and the MLE, and propose to minimize “penalized” versions of the disparities, where the impact of the empty cells

³sometimes much better, e.g. Hellinger, upper left panel of Figure D.6, at a sample size of 100, where the MDE’s bias is about 12% of the parameter value ($m = 5$), and the WML’s is about 4.5%.

is reduced. They show in their Table 3 p.23 that the penalized disparities have lower - and more stable over λ - values of the mean square error in a simulation at the Poisson model with mean 5 and a sample size of 20. In particular, Hellinger has an MSE of 0.3250 and its penalized version has an MSE of 0.2781; a MDE with $\lambda = -0.8$ (corresponding to their $\alpha = 0.2$) has an MSE of 0.5272 while its penalized version has 0.2949. In a simulation in the same conditions we obtain 0.2952 for Hellinger and 0.2723 for the corresponding WML, and for $\lambda = -0.8$ we get 0.4852 for the MDE and 0.2966 for the WML, so we see that the same kind of MSE reduction and stabilization over λ is offered by the WML in this situation⁴. A second remark is that the same simulation for the NE yields an MSE of 0.2811, which is very similar to the MSE of the penalized Hellinger, and yet the NE offers stronger large outlier downweighting than Hellinger (see Figures 3.2, 3.3 and 3.4). This points out, once more, the NE as a very performing estimator. In passing, the WML based on the NE in the previous simulation had an MSE of 0.2835, i.e. almost the same as the NE.

Coming back to the problem of the small sample bias, we see that, at least in the Poisson case, the penalized estimators of Basu et al. (1997) do not perform better than the WML or the NE; we do not know whether they would offer an improvement in a markedly over-dispersed situation, where our estimators can suffer important bias. Anyway an alternative method should be developed for estimation in that type of situation. Note that in the NB model, a value of α around 1 can already imply much over-dispersion, depending on the value of m (recall the variance equals $m + m^2\alpha$). Let us point out, however, that in the case of low over-dispersion, the WML performs very well in quite small samples, as we shall see in the next chapter with two examples with real data.

Finally, some comments about the TML and the WMLc. The influence of

⁴We estimated a NB model on Poisson data, as the programs have for the time being been developed only for NB. We see that this does not cause a large efficiency loss.

the initial estimator is much stronger on those estimators than on the WML, as can be seen on the lower panels of figures D.1 and D.2. In that case, for visibility reasons, we show the curves only for the NE and the “ $\lambda = -3$ ”. Again, the above ordering is respected, but the differences are much larger than for the WML, and no shrinkage is visible in the considered range of sample sizes. The TML is seen to be superior to the WMLc, and this will be the case in all our examples.

On Figure D.3, we have plotted the NE, which is the best estimator in terms of MSE efficiency, as well as the corresponding WML and TML. The WML shows comparable, yet slightly weaker, performances than the NE. However, it is superior to the TML, and this is the case in all our examples.

7.2 Simulations at contaminated models

All simulations have been done with a contamination rate of 10%, apart from the ones presented on Figure D.9 where the rate is 20%.

The main result is shown on Figure D.8, which presents the root mean square errors (RMSE) of the MDEs and the corresponding WMLs, as a function of contamination position. In almost all cases, the RMSE of the WML is globally lower than the RMSE of the MDE. Thus the WML is seen to improve the initial estimate both at the model and in the presence of contamination.

The only MDE which is sometimes better than the WML is the NE, but even then, its RMSE is larger than the RMSE of the WML by a factor as large as 1.5 for certain contamination positions. Moreover, when the contamination rate is 20% the WML is globally better than the NE over the whole range of tested contamination positions, improving it by factors sometimes close to 2 (see Figure D.9).

Figures D.10, D.11 and D.12 show different aspects of the simulations with 10% contamination. On Figure D.10, we plotted all MDEs and the

most extreme WMLs: the NE and the “ $\lambda = -3$ ”. All the WML curves are inside the envelope formed by the NE and the “ $\lambda = -3$ ” (filled in lightgrey on the plots). Again the differences between the MDEs are much larger than the differences between the WMLs, showing the limited influence of the choice of the MDE on the performances of the WML.

Thus, our choice of the tuning constants $b = 3.5$ and $n_{\max} = 200$ is appropriate. As noted in section 4.1.2, b and n_{\max} control the threshold (on the standardized difference between the observed and the predicted frequencies) above which a sample space element gets a zero weight. Here this threshold is large enough so that in the absence of contamination not too many observations are removed (even if the initial estimator is biased), thus allowing high efficiency, and yet small enough for contaminations rates of 10% to be successfully suppressed.

One point that might seem surprising when considering the MDE curves on Figure D.10 is that they sometimes show very different patterns, some having local minima where some others have local maxima. This is again imputable to the strong bias which affects some MDEs in finite clean samples (hereafter: the “base bias”). Figure D.11 shows the respective contributions of bias and standard deviation to the RMSE curves of the MDEs, for the m parameter at the contaminated model $(m, \alpha) = (5, 2/3)$. The two lower graphs have the same scale, to allow visual comparison of the curves therein. We see that the patterns of the RMSE curves are mainly dictated by the bias patterns. These bias patterns are similar, but differently positioned with respect to 0, which implies quite different patterns for the RMSEs. (For example Hellinger’s “base bias” is such that its bias under contamination ends up being zero at the point where the outlier has the largest influence.)

Finally, Figure D.12 shows the respective performances of the linNEG and the corresponding TML, WMLc, and WML. The situation is very similar for the other MDEs. We see that the WMLc is again weaker than the TML, which in turn is weaker than the WML. The WMLc and the TML curves

show the same shortcoming which was already noted on the asymptotic bias curves (Figures 6.2, 6.3, 6.4), and anticipated in chapter 4.2: for low contamination positions, they follow the MLE curve (generally doing worse) up to the point where the contamination starts being eliminated by the adaptive cut-off procedure. The reason why they do worse than the MLE before the drop down is the following: the fact that the contamination is not eliminated does not mean that no observations are removed. Depending on the sample, the adaptive cut-off can have a value not far above the contamination position, resulting in an important loss of information and thus an increase in variability.

In the case of the linNE, the WMLc and the TML show worse performances than the MDE on which they are based. The same is true for the NE. For Hellinger, Neyman and “ $\lambda = -3$ ”, they do a little better for large contamination positions. But they never do better than the WML.

Coming back once again to our discussion about the choice of the initial estimator, we see that the most robust members of the Cressie-Read family are ruled out by their finite sample shortcomings, even in the presence of contamination. The WML based on these MDEs offers a large improvement of their performances, yet the WMLs based on MDEs which are more efficient and less biased in clean samples are better. Therefore we advise to start with either the NE or the linNE.

Chapter 8

Illustration with real data

We present two examples of application of the WML.

8.1 Chemical mutagenicity data

In the sex-linked recessive lethal test in *Drosophila* (fruit flies), male flies are exposed to different doses of a chemical to be screened. Each male is then mated with unexposed females. One observes the number of daughter flies carrying a recessive lethal mutation on the X chromosome. Details of the experimental protocol can be found in Woodruff, Mason, Valencia, and Zimmering (1984). These data were previously analyzed by Simpson (1987). A striking feature of these data is the occasional occurrence of exceptionally large counts. Woodruff et al. (1984) referred to these exceptional counts as “clusters”. They conjectured that, unlike the majority of the recessive lethals, which result from mutations during meiosis, a cluster results “from a single spontaneous premeiotic event” (p. 195). Consequently, they advocated the exclusion of observations identified as clusters.

Table 8.1 reports the observed frequencies of daughters with lethal mutation in one such experiment, and we note the presence of a very large outlier (having value 91).

Table 8.1: Observed distribution of the number of daughters carrying a recessive lethal mutation on the X chromosome.

Number of daughters	0	1	2	91
Frequency	23	7	3	1

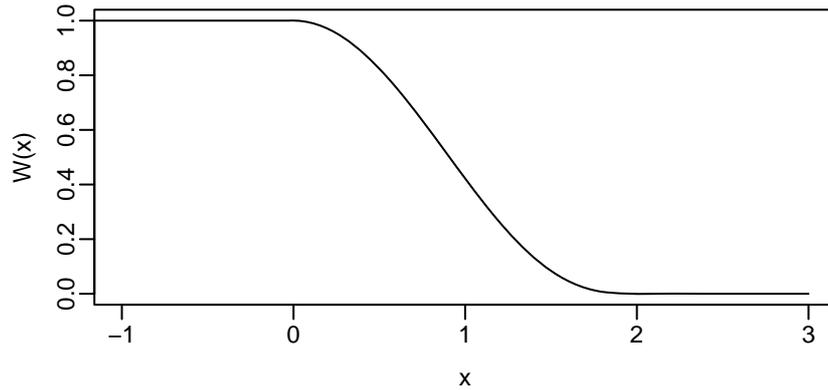


Figure 8.1: The weight function of the WML in the drosophila example.

We considered a NB fit to these data. We computed the 5 MDEs (NE, linNE, Hellinger, Neyman, “ $\lambda = -3$ ”) and the corresponding WMLs, as well as the MLE and the MLE after removal of the outlier (MLE*). In this example, we used smooth weights in the computation of the WML, for reasons to be clarified further. In (4.2), we used

$$W(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ 1 - \rho_k(x) & \text{if } x > 0, \end{cases} \quad (8.1)$$

where

$$\rho_k(x) = \begin{cases} 1 - [1 - (x/k)^2]^3 & \text{if } |x| \leq k \\ 1 & \text{if } |x| > k \end{cases} \quad (8.2)$$

is Tukey’s biweight function. We used the value $k = 2$, and $n_{\max} = 200$ in (4.4). See the shape of $W(x)$ on Figure 8.1.

Table 8.2: Estimates of the NB parameters for the drosophila data.

Estimator		m	sd(m)	sd(m)	α	sd(α)	sd(α)
		(as. approx.)	(bootstrap)	(bootstrap)	(as. approx.)	(bootstrap)	(bootstrap)
NE	MDE	0.40	0.12	0.11	0.38	0.81	0.93
	WML	0.37	0.11	0.11	0.18	0.77	0.42
linNE	MDE	0.39	0.11	0.11	0.24	0.76	0.62
	WML	0.36	0.11	0.10	0.15	0.77	0.39
Hellinger	MDE	0.36	0.10	0.10	0.00	0.15	0.68
	WML	0.33	0.10	0.11	0.01	0.75	0.21
Neyman	MDE	0.39	0.12	0.11	0.47	0.87	1.27
	WML	0.37	0.11	0.11	0.16	0.77	0.31
“ $\lambda = -3$ ”	MDE	0.46	0.14	0.13	0.93	0.99	1.72
	WML	0.38	0.11	0.11	0.22	0.76	0.43
	MLE*	0.39	0.11	0.11	0.25	0.76	0.84
	MLE	3.06	2.43	2.62	9.97	4.08	6.42

The results are presented in Table 8.2. Except for the MLE, all estimates are rather similar in value, but a greater similarity is noted amongst the WMLs than amongst the MDEs. The variation amongst the different estimates is more important for parameter α than for m , yet this has little influence on the predicted frequencies, which are all very much alike (see Figure 8.2).

The standard errors of the estimates were estimated with the asymptotic formula (6.2) and also via a bootstrap procedure, using the empirical distribution function¹. As can be seen in Figure 8.2, for parameter m , the asymptotic approximations coincide very well with the bootstrap values (except for the MLE, as one could expect), and all estimates in Figure 8.2 have

¹We generated 1000 pseudo-random samples of size 34, according to the empirical distribution function. (This took about 5 minutes per MDE-WML couple.)

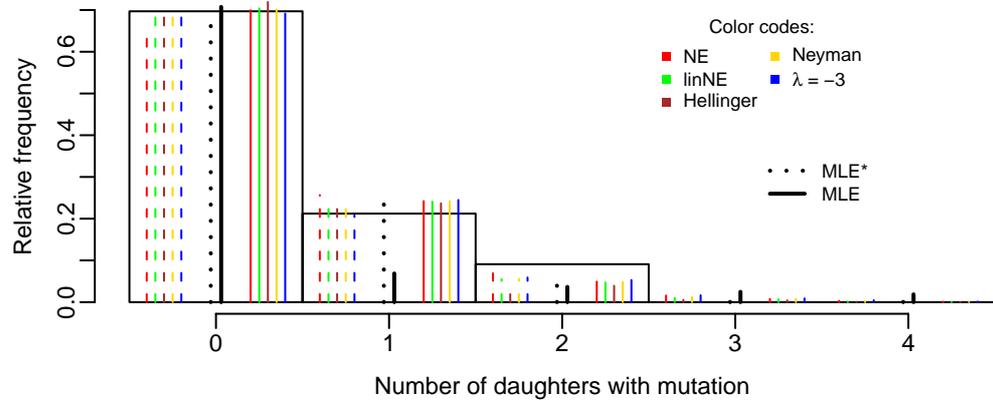


Figure 8.2: Observed and predicted frequencies for drosophila data. The boxes show the observed frequencies (calculated without the outlier in order to correctly assess the fits to the “good” data). Coloured lines show the MDEs (dashed) and the WMLs (solid). The two central black lines show the MLE* (dotted) and the MLE (solid).

about the same variability (which also corresponds quite well to the jackknife estimate of standard error (with value 0.10) given by Simpson (1987), who fitted a Poisson model to these data with minimum Hellinger distance). For parameter α , the asymptotic approximation is rather close to the bootstrap result for the more efficient MDEs (NE, linNE) and for the MLE*, and underestimates the sd of the less efficient MDEs (Neyman, “ $\lambda = -3$ ”). For the Hellinger MDE, the asymptotic approximation also underestimates the sd relative to the bootstrap, but the value of the estimate is extremely close to the lower limit of the parameter set, so that one can expect the asymptotic approximation not to work well. For the WMLs, the sd as estimated by bootstrap is systematically lower than the asymptotic approximation. Here, the WMLs work better than the MDEs and also than the MLE*.

Actually, this feature is linked to the use of smooth weights. Attempts with “hard” weights were not satisfactory in this case, a fact for which we give the following interpretation: as can be seen in Figure 8.2, the frequency

of zeros is very high in the considered situation. For small sample sizes, samples with even higher proportions of zeros will often arise, and this is the source of much instability in the estimates of the α parameter, which will have very large values in such samples. To face this situation, we need to downweight the influence of position 0 in the estimating equations, without removing it completely, as then we would be discarding the great majority of the data, also resulting in a poor efficiency.

Thus, for situations with a high proportion of zeros and a rather small sample size, the WML can be preferable to the MLE even in uncontaminated situations.

Finally, it is visible in Table 8.2 and in Figure 8.2 that the MLE is badly corrupted by the presence of the outlier.

8.2 Lengths of hospital stays

In the second example we consider lengths of hospital stays (LOS). In modern hospital management, stays are often classified into “diagnosis related groups” (DRGs), and LOS is used as a cost indicator. The mean LOS of several hundred DRGs are then used for budgeting purposes and to compare the economic efficiency of different hospitals. LOS distributions often contain outliers whose value and frequency fluctuate from year to year. Thus, robustness is an important issue if one wishes to obtain stable summaries. Since many DRGs must routinely be inspected each year, automatic outlier detection is important in this field.

Table 8.3 shows an example of 32 stays in a Swiss hospital in 1988, classified into DRG “disorders of the nervous system”. A simple visual inspection of this particular DRG identifies three outliers: 115, 198, 374. We considered a NB fit to these data, which were previously analyzed by Marazzi and Ruffieux (1999), who considered a Weibull fit, and Marazzi and Yohai (2010), who also considered a NB fit.

Table 8.3: Lengths of stay of 32 hospital patients.

LOS	1	2	3	4	5	6	7	8	9	16	115	198	374
frequency	2	6	5	5	4	2	2	1	1	1	1	1	1

Negative binomial fits of LOS data are quite frequent in the literature, see for example UCLA (2010), the examples in Hilbe (2007), or Marazzi and Yohai (2010). See Bithell (1969) for a mathematical justification of the NB model for LOS modeling.

Like Marazzi and Yohai (2010), we modeled the distribution of LOS-1 with a NB distribution. Like in the previous example, we computed the five MDEs and the corresponding WMLs, as well as the MLE and the MLE*. Here, the MLE* is the maximum likelihood estimate of the data without the four largest observations (16, 115, 198, 374). Indeed, the maximum predicted frequency for LOS=16 amongst the 10 robust fits we have computed is 0.00015, which points it out as a clear outlier in a sample of size 32.

Here we have used “hard” weights in the computation of the WML, with $b = 3.5$ and $n_{\max} = 200$. (An attempt with smooth weights as given by (8.1) with $k = 3.5$ yielded an estimate with a slightly larger variability than with hard weights.)

Again, the standard deviations of the estimates were estimated with the asymptotic formula (6.2) and by bootstrap². Like for the drosophila data, all robust estimates were very close in value; actually, all WMLs were exactly equal (all of them remove just the four outliers). In Table 8.4, we present the results for the NE, the corresponding WML, the MLE and the MLE*. The corresponding fits are shown in Figure 8.3, where all distributions are conditioned on the interval $[0, 14]$, in order to correctly assess the fits to the “good” data. The MDE, the WML and the MLE* fit the data well, but the

²We generated 1000 pseudo-random samples of size 32, according to the empirical distribution function. (This took about 10 minutes per MDE-WML couple.)

Table 8.4: Estimates of the NB parameters for the LOS data.

Estimator	m	$sd(m)$	$sd(m)$	α	$sd(\alpha)$	$sd(\alpha)$
		(as. approx.)	(bootstrap)		(as. approx.)	(bootstrap)
NE MDE	3.04	0.40	0.42	0.16	0.14	0.12
WML	3.00	0.39	0.43	0.15	0.13	0.11
MLE*	3.00	0.39	0.38	0.14	0.13	0.11
MLE	24.47	19.38	12.54	3.08	0.87	0.80

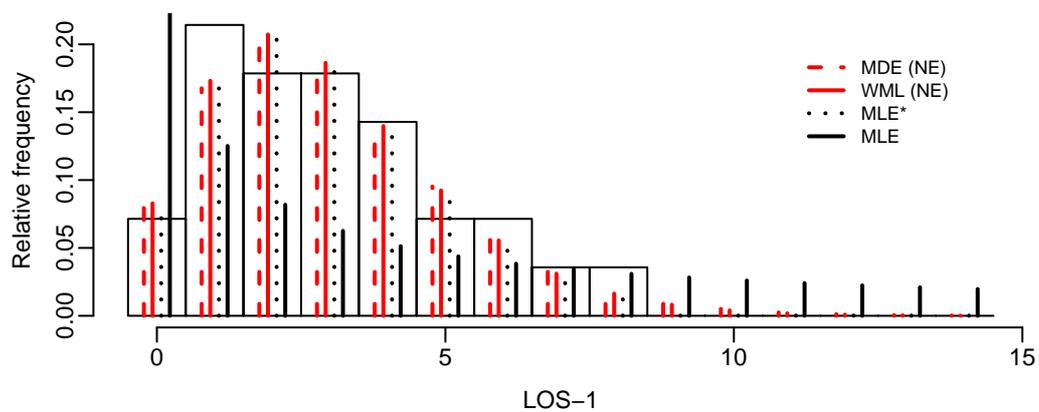


Figure 8.3: Observed and predicted frequencies for LOS data. The boxes show the observed frequencies.

MLE is badly corrupted by the outliers.

In Table 8.4 we see that, except for the MLE, the standard errors of the different estimates are quite similar, and in good agreement with the value given by the asymptotic formula. Moreover, they are smaller than the standard errors obtained by Marazzi and Ruffieux (1999) for estimation of the mean of the same data, with truncated means (a procedure similar to our TML, but with a fixed cut-off) and the Weibull model: their standard deviations range from 0.57 to 2.65, depending on the initial estimator (in Table 8.4, the largest standard error on m (excluding the MLE) is 0.43).

If we compare our estimated values to the ones obtained by Marazzi and Yohai (2010), we see that their most robust estimator (an M-estimator with 80% asymptotic relative efficiency) yields the values 3.58 and 0.44 for m and α , which are markedly larger than the 3.00 and 0.15 yielded by the WML (the latter are much closer to the MLE*). A visual comparison of the fits in Figure 8.3 and in Figure 1 in Marazzi and Yohai (2010) shows that the WML fit is better.

Finally, in this example, the standard deviations of the MDE and the WML are almost equal. Note however that with less efficient initial estimators (Neyman and “ $\lambda = -3$ ”), the WML provided a substantial improvement (and the asymptotic formula underestimated the standard deviation of the MDEs, like in the drosophila example).

Chapter 9

Computation

Programs to compute the estimates have been developed using the R programming language. At the time of writing, these programs have been developed for the specific case of estimation of the negative binomial parameters. The programs are available from the website of the Statistical Unit of the Institute for Social and Preventive Medicine, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland:

http://www.iumsp.ch/Unites/us/msp_us.htm.

These programs use the built-in optimizing function `optim` to minimize the disparities (eq. (3.8)) and the weighted likelihood (eq. (4.9)). `optim` needs to be given a starting point $(m_{\text{start}}, \alpha_{\text{start}})$ for the minimization. It is important that this starting point be robust to avoid convergence of the algorithm to a wrong local minimum in the presence of outliers. We use the estimates of the NB parameters obtained via the S-estimates, presented in section 3.3. An ad-hoc programme has been developed to solve system (3.7) which relates the S-estimates of location and scale to the NB parameters.¹

¹In section 3.3, we pointed out that the S-estimators of location and scale “collapsed” when more than half the data have the same value, and that the system (3.7) did not always have a solution. When one of these situations arises, more specific rules are applied to find the starting point in a robust way.

A help file to use the software can also be found at the same url.

Chapter 10

Conclusion and perspectives

We have developed a two-step estimation procedure for discrete distributions which combines interesting robustness and efficiency properties. Like the minimum disparity estimators, it offers a high breakdown point together with full asymptotic efficiency. However while the finite sample performances of different MDEs can be very different, the performances of the corresponding WMLs are much closer to each other. In a large variety of situations, these performances - in terms of mean square error - are better than those of the MDE used as initial estimator. This effect is particularly important in the presence of contamination in the data: the influence of a contamination on the WMLs is generally much weaker than its influence on the MDEs.

The stability of the WML with respect to change of the initial estimator attenuates the importance of the choice of that estimator, however it appears in simulations that the best performances of the WML are obtained by starting with the minimum negative exponential estimator, which presents a particularly good balance in terms of robustness, efficiency and small sample bias.

The idea of using an initial estimator as a tool for outlier detection and rejection was already present in Marazzi and Ruffieux (1999); Gervini and Yohai (2002); Marazzi and Yohai (2004). However, these authors considered

rejecting - or downweighting - the whole tails of a distribution starting from an observation-dependent cut-off, determined from the initial estimator. The procedure proposed in this thesis, particularly natural in the discrete setting, allows to downweight more specifically the observations which are in contradiction with the initial model. This feature has two positive consequences:

- Outliers can be downweighted regardless of their position in the sample
- Efficiency losses are reduced by the possibility of eliminating outliers more specifically, without removing too many “good” observations.

This procedure offers some flexibility with two constants, b and n_{\max} , which regulate the robustness-efficiency trade-off. The constant b is interpreted as a quantile of a standard normal distribution, which facilitates its choice. The choice of n_{\max} relies on numerical investigation and exploration of the asymptotic bias under contamination in the model of interest. This investigation has been carried out by the author in the negative binomial model, and the values $b = 3.5$ and $n_{\max} = 200$ have been shown to provide highly performing estimators of the NB parameters, in terms of finite sample mean square error, both in the presence and in the absence of contamination, and in terms of asymptotic bias under contamination.

However, in chapter 7, we mentioned the fact that a drawback of the proposed procedure is a possibly important small sample bias in markedly over-dispersed models. This is not just an effect of the bias of the initial estimator; the WML does have some intrinsic small sample bias in such models. This drawback should be explored more precisely, in order to better determine the appropriate domain of application of the WML.

When the data are too much over-dispersed, another procedure should be used. Possible alternatives are M-estimation procedures as in Marazzi and Yohai (2010) or Cadigan and Chen (2001). However, to the author’s knowledge, no precise statement has been made about the breakdown point or the contamination bias of these methods.

Coming back to the WML, future work could also include computing its second order efficiency, which might be an interesting tool to better understand its finite sample efficiency properties. However, recall that second order efficiency is not always an adequate measure of finite sample efficiency: in our simulations the most efficient MDE is Hellinger, which is not second order efficient, while NE is and has a lower finite sample efficiency. This was also noted by Basu and Sarkar (1994) in an empirical study at the normal model.

In this thesis we considered principally “hard” weights, i.e. the observations are either removed or kept, but not downweighted by a factor in $(0,1)$. This was done so mainly for simplicity of exploration, so that the issue of the shape of the weight function is avoided (we are just left with two constants to set). Another advantage is that with hard weights, the WML has an interpretation as the maximum likelihood estimator of a conditioned model, calculated on the remaining observations¹. This correspondence is useful in the proof that the finite sample bdp of the WML is not lower than the finite sample bdp of the initial estimator.

However, we conjecture that this statement holds for a more general class of weight functions. Simulations support this conjecture, and moreover the statement holds for the asymptotic bdp (see section 5.1). We have seen in the drosophila example (section 8.2) that continuous weights can be of great interest in situations where we want to reduce the influence of a certain position without removing it completely. Another situation where continuous weights would probably be of interest is in the presence of spread contamination. In fact, in our numerical study, we considered only point contamination

¹Note that this does not make the WML a real MLE. It would be so if the remaining observations really followed the conditioned model based on the initial estimator, which is not true. Positions which get a zero weight were not “doomed” to be discarded: the probability, before we draw the data, that a remaining observation has that value is not zero. Thus the conditioned model does not correspond to the distribution of the remaining data, even if the sample is generated by the corresponding global (unconditioned) model.

at varying positions. While this kind of contaminations is not unrealistic in the discrete setting, and provides a good picture of the contamination bias pattern of an estimator, it would be interesting to study the behavior of the WML under more spread contaminations. The WML with hard weights could suffer an important efficiency loss in such a situation, if it eliminates numerous positions from the calculation of the final estimator. Working with smooth weight could then improve its performances.

10.1 Perspectives

Possible future developments include the extension of the WML to the continuous distribution setting. The discrete setting offers a natural framework for the WML, as it automatically provides categories - the sample space elements - to define the weights; in the continuous setting, no such natural categorization exists. In what follows, we sketch a possible method to apply the WML to continuous distributions.

To start with, we suppose that we have computed a robust and consistent initial estimate β_1 of the parameters of the continuous model g_β in consideration. This initial estimator could be a MDE, methods to apply this kind of estimators in the continuous framework exist (see e.g. Simpson (1987), Basu and Lindsay (1994)). These methods generally imply comparing the model with a nonparametric density estimate based on the sample.

Let $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ be an observed sample and g_{β_1} be the initial model. Here is a possible procedure:

- Define a fixed width $h(\sigma_1)$, where σ_1 is a dispersion measure of f_{β_1} and h is some increasing function.
- For each observation x_i , consider the interval $[x_i - h(\sigma_1)/2, x_i + h(\sigma_1)/2]$ and calculate p_i , the probability associated to that interval under the initial model. Define d_i as the proportion of observations in the interval.

- Similarly as in section 4.1.2, define a weight for each observation as

$$\omega_i = W \left(f(n) \frac{d_i - p_i}{\sqrt{p_i(1 - p_i)}} \right), \quad (10.1)$$

with $f(n)$ as in (4.4). Let us consider hard weights for simplicity:

$$W(x) = I(x < b)$$

for some positive value b .

- Let $x_{(i)}$ denote the i^{th} smallest observation and $\omega_{(i)}$ its associated weight. Define the function $\omega(x)$ to be equal to $\omega_{(i)}$ when $x = x_{(i)}$ and, between two consecutive observations $x_{(i)}$ and $x_{(i+1)}$, to be equal to 1 if $\omega_{(i)} = \omega_{(i+1)} = 1$, to 0 if $\omega_{(i)} = \omega_{(i+1)} = 0$ and to be 0 on the first half of the interval $[x_{(i)}, x_{(i+1)}]$ and 1 on the second half if $\omega_{(i)} = 0$ and $\omega_{(i+1)} = 1$, and conversely if $\omega_{(i)} = 1$ and $\omega_{(i+1)} = 0$. Define $\omega(x)$ in the tails in an analogous manner.
- Analogously to (4.9), define the WML estimate of β as the value that minimizes

$$\log \left(\int \omega(x) g_\beta(x) dx \right) - \frac{1}{\sum_{i=1}^n \omega_i} \sum_{i=1}^n \omega_i \log g_\beta(x_i).$$

Note that no density estimation is necessary for this procedure.

The above procedure could represent an alternative to the adaptive cut-off method applied to regression with asymmetric errors in Marazzi and Yohai (2004).

Acknowledgements. I thank my thesis advisor Alfio Marazzi for his helpful comments and advice, and Valentin Rousson for useful and constructive discussions.

Appendix A

Identifiability in location-scale families

Let $F_{\mu,\sigma}$ be a location-scale family of distributions and Ω a set of distributions containing $F_{\mu,\sigma}$. Let $m : \Omega \rightarrow \mathbb{R}$ and $s : \Omega \rightarrow \mathbb{R}_+$ be respectively a location and a scale measure on Ω .

Notation: the distribution of a random variable Z is denoted F_Z .

Let $X_1 \sim F_{\mu_1,\sigma_1}$ and $X_2 \sim F_{\mu_2,\sigma_2}$. If $m(F_{\mu_1,\sigma_1}) = m(F_{\mu_2,\sigma_2})$ and $s(F_{\mu_1,\sigma_1}) = s(F_{\mu_2,\sigma_2})$, then

$$\mu_1 = \mu_2 \quad \text{and} \quad \sigma_1 = \sigma_2.$$

Proof. Let $m(F_{\mu_1,\sigma_1}) = m(F_{\mu_2,\sigma_2}) = m_X$ and $s(F_{\mu_1,\sigma_1}) = s(F_{\mu_2,\sigma_2}) = s_X$ and consider the variable $A = \frac{X_1 - \mu_1}{\sigma_1} \sigma_2 + \mu_2$. Since $F_{\mu,\sigma}$ is a location-scale family, we have that $F_A = F_{\mu_2,\sigma_2}$ and so

$$m(F_A) = m_X \quad \text{and} \quad s(F_A) = s_X. \tag{A.1}$$

But since m and s are location and a scale measures, we have

$$m(F_A) = \frac{m_X - \mu_1}{\sigma_1} \sigma_2 + \mu_2 \quad \text{and} \quad s(F_A) = \frac{s_X}{\sigma_1} \sigma_2. \tag{A.2}$$

Combining (A.1) and (A.2) yields the conclusion. ■

Appendix B

Proof of Theorem 5

We shall refer to the proofs given in Lindsay (1994) (his Proposition 12 and Lemma 20). The central point of Lindsay's proof is the convergence

$$\lim_{j \rightarrow \infty} \rho(d_j, m_\beta) = \rho(d_\epsilon^*, m_\beta), \quad (\text{B.1})$$

which he shows to hold under Assumption 1 for any $\epsilon \in [0, 1)$. Then, Lindsay assumes that the convergence (B.1) is uniform in β inside any compact set B of parameter values containing b^* . The uniformity of the convergence, together with Assumption 2, implies that any sequence $\{b_j\}$ of values of β that minimize $\rho(d_j, m_\beta)$ over β in B converges to b^* . Finally, Lindsay builds a lower bound on $\rho(d_j, m_\beta)$ for $\beta \notin B$ and determines the values of ϵ for which b^* is eventually the global minimum when $j \rightarrow \infty$.

In what follows, we prove that if $m_\beta = \text{NB}_{m, \alpha}$ we do not need the lower bound, because the convergence (B.1) is uniform in β in the whole parameter space Θ .

From the proof of (B.1) given by Lindsay, it appears that the convergence will be uniform inside parameter set B if

$$\sup_{\beta \in B} m_\beta(x_j) \rightarrow 0 \text{ as } j \rightarrow \infty, \quad (\text{B.2})$$

and so we have to prove that

$$\sup_{(m,\alpha) \in \mathbb{R}_+^2} \text{NB}_{m,\alpha}(x_j) \rightarrow 0 \text{ as } x_j \rightarrow \infty \quad (\text{B.3})$$

(note that from Remark A, we have that x_j is an outlier sequence iff $\lim_{j \rightarrow \infty} x_j = \infty$). The bdp result can then be extended to the multiple outlier sequence contamination of Proposition 4 in an elementary fashion.

Let \hat{m} and $\hat{\alpha}$ be the maximum likelihood estimates for the sample composed of the single observation x_j . By definition,

$$\text{NB}_{m,\alpha}(x_j) \leq \text{NB}_{\hat{m},\hat{\alpha}}(x_j) \quad \forall (m, \alpha) \in \mathbb{R}_+^2, \quad \forall x_j \in \{1, 2, \dots\}.$$

In the NB model, the maximum likelihood estimate for parameter m is the sample mean, and so $\hat{m} = x_j$. Now we observe that:

- A sample consisting of one single non zero observation has sample mean superior to sample variance.
- When the sample mean is superior to the sample variance, the value of α which maximizes the likelihood in \mathbb{R}_+ is 0, i.e. we get a Poisson distribution (Anscombe, 1950).

Thus, $\text{NB}_{\hat{m},\hat{\alpha}} = P_{x_j}$, where P_{x_j} is the Poisson distribution with mean x_j , and so

$$\text{NB}_{m,\alpha}(x_j) \leq P_{x_j}(x_j) = e^{-x_j} \frac{x_j^{x_j}}{x_j!} \quad \forall (m, \alpha) \in \mathbb{R}_+^2, \quad \forall x_j \in \{1, 2, \dots\}.$$

Now from Stirling's formula (Abramowitz and Stegun, 1964)

$$\lim_{n \rightarrow \infty} \frac{n!}{\sqrt{2\pi n} (n/e)^n} = 1,$$

we get that

$$\lim_{x_j \rightarrow \infty} P_{x_j}(x_j) = 0$$

and (B.3) follows. ■

Appendix C

Breakdown point of MDEs from the Cressie-Read family with $\lambda \leq -1$

In what follows we prove that the finite sample bdp of the MDEs from the Cressie-Read disparity with $\lambda \leq -1$ is 1 in the NB model. We give the proof for the case of contamination of the sample with one single outlier sequence. The result is then easily extended to the multiple outlier sequence contamination of Proposition 4 (see also Remark B).

Let $d(x)$ be the observed frequencies. Like in section 3.4.4, define $\mathbf{X}_{\mathbf{F}} = \{x \in \mathbf{X} : d(x) \neq 0\}$, $\tilde{m}_{\beta}(x) = \frac{m_{\beta}(x)}{\sum_{\mathbf{X}_{\mathbf{F}}} m_{\beta}(x)}$, $\tilde{\delta}(x) = \frac{d(x) - \tilde{m}_{\beta}(x)}{\tilde{m}_{\beta}(x)}$, $S_{\mathbf{X}_{\mathbf{F}}}(\beta) = \sum_{\mathbf{X}_{\mathbf{F}}} m_{\beta}(x)$, $P(S_{\mathbf{X}_{\mathbf{F}}}(\beta)) = \frac{1 - S_{\mathbf{X}_{\mathbf{F}}}(\beta)}{S_{\mathbf{X}_{\mathbf{F}}}(\beta)}$ and consider the estimator defined as the minimum over β of the penalized disparity

$$\rho_p(d, m_{\beta}) = \sum_{\mathbf{X}_{\mathbf{F}}} \tilde{m}_{\beta}(x) G(\tilde{\delta}(x)) + P(S_{\mathbf{X}_{\mathbf{F}}}(\beta)).$$

Next consider the ϵ -contaminated data

$$d_j(x) = (1 - \epsilon)d(x) + \epsilon\chi_{x_j}$$

and define $\mathbf{X}_{\mathbf{F}_j} = \{x \in \mathbf{X} : d_j(x) \neq 0\}$ and $d_\epsilon^*(x) = (1 - \epsilon)d(x)$. In an analogous approach to what is done in Lindsay (1994) and in Appendix B, we assume that $\rho_p(d_\epsilon^*, m_\beta)$ has unique absolute minimum at some point b^* of the parameter space.

In what follows we show that $\forall \epsilon \in [0, 1)$, $\rho_p(d_j, m_\beta) \rightarrow \rho_p(d_\epsilon^*, m_\beta)$ as $x_j \rightarrow \infty$ uniformly in β for β inside some parameter subset \mathbf{B} such that

- $b^* \in \mathbf{B}$
- $\exists x_0$ such that $\forall x_j > x_0, \forall \beta \notin \mathbf{B}$, it holds that $\rho_p(d_\epsilon^*, m_{b^*}) < \rho_p(d_j, m_\beta)$.

The uniformity of convergence and the continuity of $\rho_p(d, m_\beta)$ in β then imply that the absolute minimum of $\rho_p(d_j, m_\beta)$ is eventually b^* and thus the estimator does not breakdown.

Define \mathbf{B} as $\mathbf{B} = \{\beta : \sum_{\mathbf{X}_F} m_\beta(x) \geq \gamma\}$ for some yet to be chosen $\gamma < 1$ and note that since $d(x)$ corresponds to a finite sample, X_F and $\{x_j\}$ are disjoint if x_j is large enough. Choose such an x_j and write

$$\rho_p(d_j, m_\beta) = A_j + B_j + C_j$$

where

$$\begin{aligned} A_j &= \sum_{\mathbf{X}_F} \tilde{m}_\beta^j(x) G \left(\frac{d_\epsilon^*(x)}{\tilde{m}_\beta^j(x)} - 1 \right), \\ B_j &= \tilde{m}_\beta^j(x_j) G \left(\frac{d_\epsilon^*(x_j) + \epsilon}{\tilde{m}_\beta^j(x_j)} - 1 \right), \\ C_j &= P(S_{\mathbf{X}_{\mathbf{F}_j}}(\beta)) \\ &= \frac{1}{\sum_{\mathbf{X}_F} m_\beta(x) + m_\beta(x_j)} - 1, \end{aligned}$$

where

$$\tilde{m}_\beta^j(x) = \frac{m_\beta(x)}{\sum_{\mathbf{X}_F} m_\beta(x) + m_\beta(x_j)}.$$

The first goal is to prove that

$$\lim_{x_j \rightarrow \infty} A_j = \sum_{\mathbf{X}_F} \tilde{m}_\beta(x) G \left(\frac{d_\epsilon^*(x)}{\tilde{m}_\beta(x)} - 1 \right), \quad (\text{C.1})$$

$$\lim_{x_j \rightarrow \infty} B_j = 0, \quad (\text{C.2})$$

$$\lim_{x_j \rightarrow \infty} C_j = P(S_{\mathbf{X}_F}(\beta)) = \frac{1}{\sum_{\mathbf{X}_F} m_\beta(x)} - 1, \quad (\text{C.3})$$

uniformly in β for $\beta \in \mathbf{B}$. Since A_j is a continuous function of $\tilde{m}_\beta^j(x)$, to prove (C.1) it suffices to show that

$$\lim_{x_j \rightarrow \infty} \tilde{m}_\beta^j(x) = \tilde{m}_\beta(x) \quad \forall x \in \mathbf{X}_F \quad (\text{C.4})$$

uniformly in β for $\beta \in \mathbf{B}$. Since G_λ with $\lambda < 0$ satisfies $\lim_{\delta \rightarrow \infty} G(\delta)/\delta = 0$, to prove (C.2) it suffices to show that

$$\lim_{x_j \rightarrow \infty} \tilde{m}_\beta^j(x_j) = 0, \quad (\text{C.5})$$

uniformly in β for $\beta \in \mathbf{B}$. Both proofs, as well as the proof of (C.3), follow the same lines and only the proof of (C.4) is given.

Proof of (C.4)

We need to show that $\forall \zeta > 0, \exists x_1$ such that $\forall x_j > x_1, \forall \beta \in \mathbf{B}$,

$$D_j = |\tilde{m}_\beta^j(x) - \tilde{m}_\beta(x)| < \zeta.$$

We have

$$\begin{aligned} D_j &= \left| \frac{m_\beta(x)}{\sum_{\mathbf{X}_F} m_\beta(x) + m_\beta(x_j)} - \frac{m_\beta(x)}{\sum_{\mathbf{X}_F} m_\beta(x)} \right| \\ &= \left| \frac{m_\beta(x)m_\beta(x_j)}{(\sum_{\mathbf{X}_F} m_\beta(x) + m_\beta(x_j)) \sum_{\mathbf{X}_F} m_\beta(x)} \right| \\ &\leq \frac{m_\beta(x_j)}{(\sum_{\mathbf{X}_F} m_\beta(x))^2} \\ &\leq \frac{m_\beta(x_j)}{\gamma^2} \end{aligned}$$

Now since in the NB model we have that

$$\lim_{x_j \rightarrow \infty} m_\beta(x_j) = 0$$

uniformly in β over the whole parameter space Θ (see the proof of Theorem 5 in Appendix B), it holds that $\forall \zeta > 0, \forall \gamma^2, \exists x_1$ such that $\forall x_j > x_1, \forall \beta \in \mathbf{B}$,

$$m_\beta(x_j) < \zeta \gamma^2$$

and so

$$D_j < \zeta.$$

■

Thus we have proved that $\forall \epsilon \in [0, 1), \rho_p(d_j, m_\beta) \rightarrow \rho_p(d_\epsilon^*, m_\beta)$ as $x_j \rightarrow \infty$ uniformly in β for $\beta \in \mathbf{B}$. Now we just need to show that we can always choose γ in such a way that $b^* \in \mathbf{B}$ and that $\exists x_0$ such that $\forall x_j > x_0, \forall \beta \notin \mathbf{B}$, we have $\rho_p(d_\epsilon^*, m_{b^*}) < \rho_p(d_j, m_\beta)$.

Chose $\kappa \in (0, 1)$ and take γ as the solution of

$$\begin{aligned} P(\gamma + \kappa) &= \rho_p(d_\epsilon^*, m_{b^*}) \\ &= \sum_{\mathbf{X}_F} \tilde{m}_{b^*}(x) G(\tilde{\delta}^*(x)) + P(S_{\mathbf{X}_F}(b^*)), \end{aligned}$$

where $\tilde{\delta}^*(x) = \frac{d_\epsilon^*(x) - \tilde{m}_{b^*}(x)}{\tilde{m}_{b^*}(x)}$. From Jensen's inequality, we get that $\sum_{\mathbf{X}_F} \tilde{m}_{b^*}(x) G(\tilde{\delta}^*(x)) \geq G(-\epsilon)$ which is a positive lower bound¹. Since P is a decreasing function, we then have $S_{\mathbf{X}_F}(b^*) > \gamma + \kappa > \gamma$ and thus $b^* \in \mathbf{B}$.

¹It is easily checked that $G_\lambda(\delta)$ is a decreasing function for $\lambda < 0$. In addition, recall that $G_\lambda(0) = 0$.

Next consider, for $\beta \notin \mathbf{B}$,

$$\begin{aligned}
\rho_p(d_j, m_\beta) &= \sum_{\mathbf{X}_{\mathbf{F}_j}} \tilde{m}_\beta^j(x) G(\tilde{\delta}^j(x)) + P(S_{\mathbf{X}_{\mathbf{F}_j}}(\beta)) \\
&= \sum_{\mathbf{X}_{\mathbf{F}_j}} \tilde{m}_\beta^j(x) G(\tilde{\delta}^j(x)) + P(S_{\mathbf{X}_{\mathbf{F}}}(\beta) + m_\beta(x_j)) \\
&\geq P(S_{\mathbf{X}_{\mathbf{F}}}(\beta) + m_\beta(x_j)) \\
&\geq P(\gamma + m_\beta(x_j)).
\end{aligned} \tag{C.6}$$

For (C.6), note that from Jensen's inequality we have $\sum_{\mathbf{X}_{\mathbf{F}_j}} \tilde{m}_\beta^j(x) G(\tilde{\delta}^j(x)) \geq G(0) = 0$. Now using again the uniform convergence to 0 of $m_\beta(x_j)$ as $x_j \rightarrow \infty$ in the NB model, we get that $\exists x_0$ such that $\forall x_j > x_0, \forall \beta \in \overline{\mathbf{B}}$,

$$m_\beta(x_j) < \kappa$$

and so

$$\rho_p(d_j, m_\beta) > P(\gamma + \kappa) = \rho_p(d_\epsilon^*, m_{b^*}).$$

■

Appendix D

Simulation results

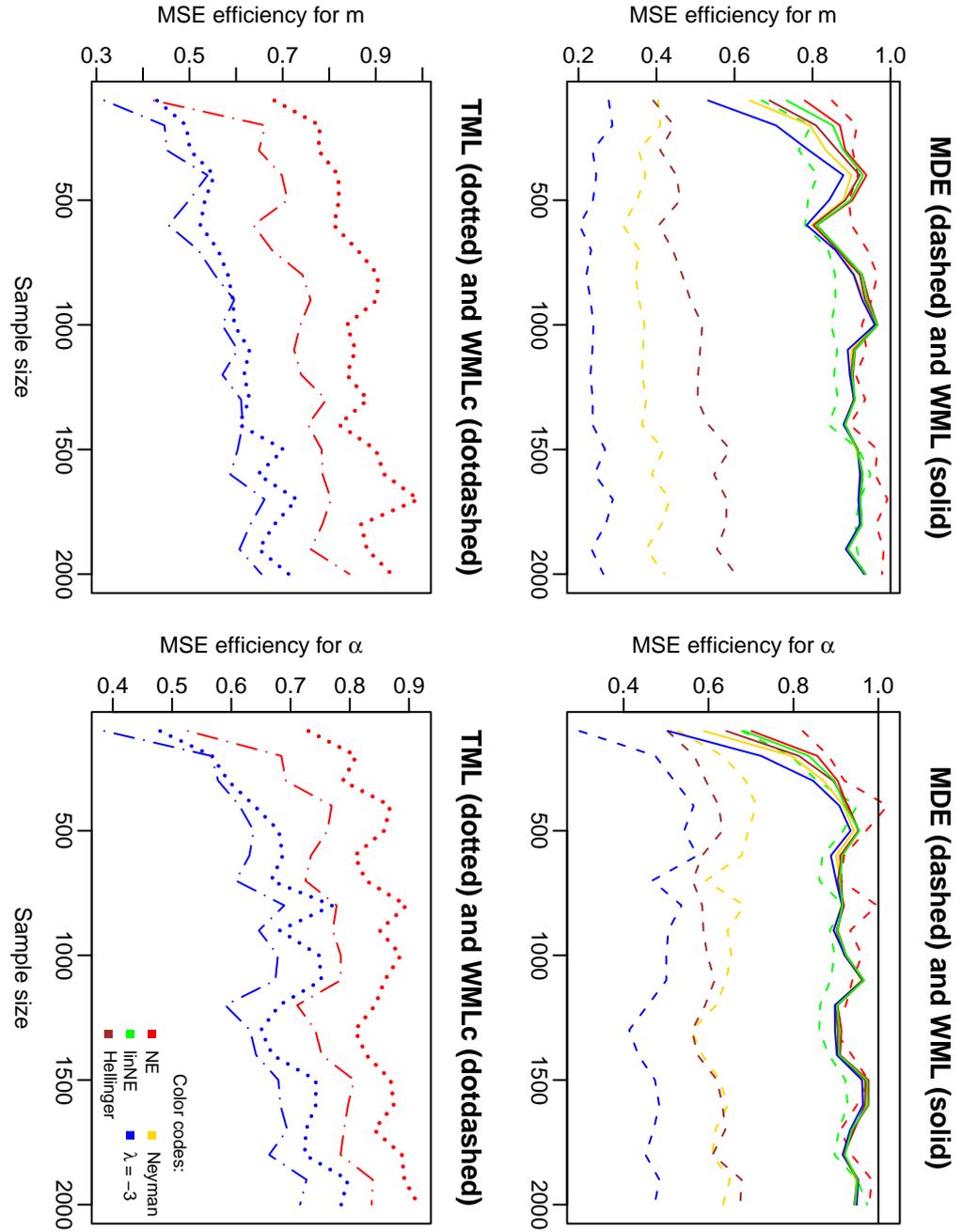


Figure D.1: MSE efficiencies. Simulations at the model $(m, \alpha) = (5, 2/3)$, for increasing sample sizes. A simulation with 500 replications was run for each size between 100 and 2000 by steps of 100.

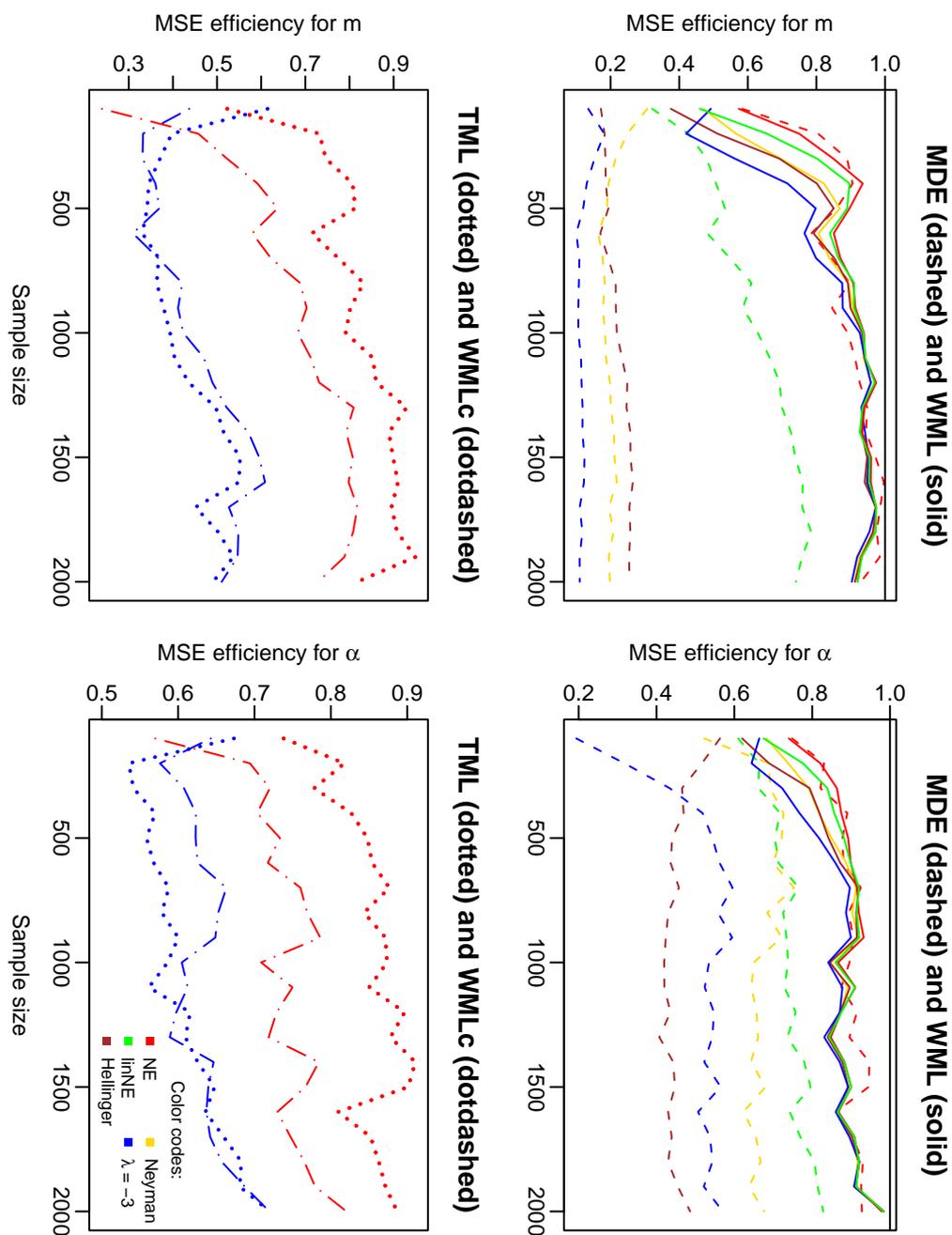


Figure D.2: MSE efficiencies. Simulations at the model $(m, \alpha) = (7, 1.5)$, for increasing sample sizes. A simulation with 500 replications was run for each size between 100 and 2000 by steps of 100.

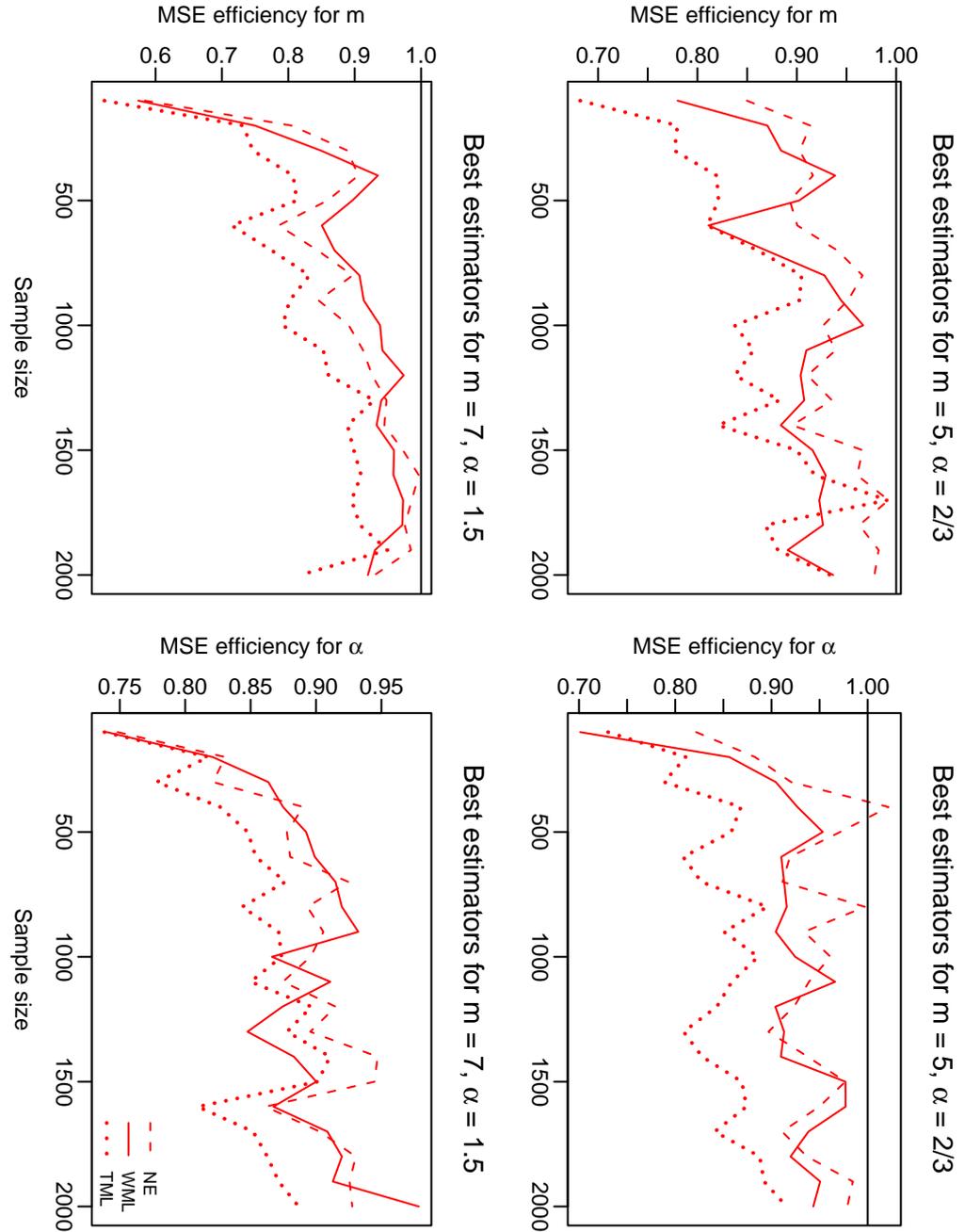


Figure D.3: MSE efficiencies for the best estimators: NE and the corresponding WML and TML. Simulations at the models $(m, \alpha) = (5, 2/3)$ and $(m, \alpha) = (7, 1.5)$, for increasing sample sizes. A simulation with 500 replications was run for each size between 100 and 2000 by steps of 100.

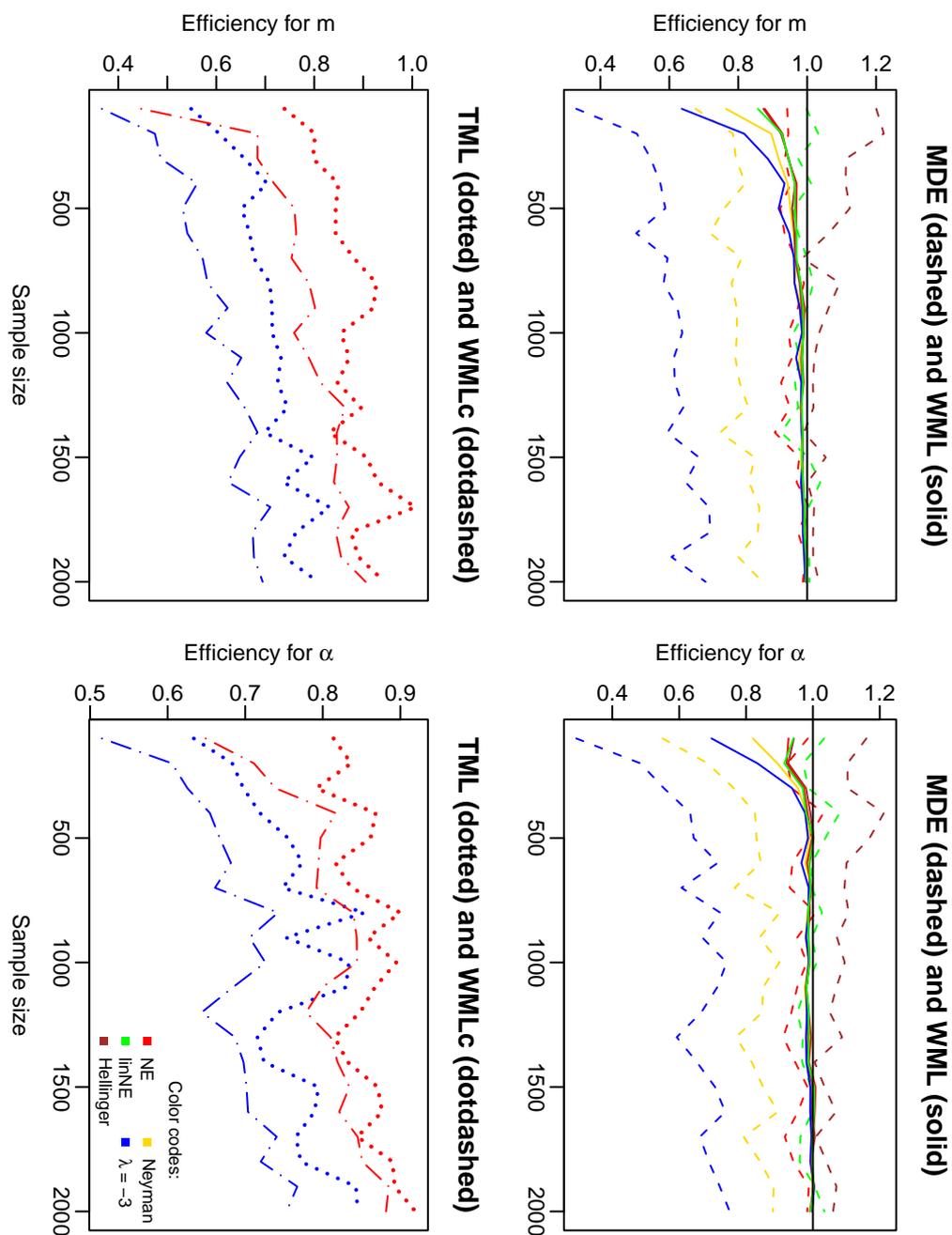


Figure D.4: “Standard” efficiencies. Simulations at the model $(m, \alpha) = (5, 2/3)$, for increasing sample sizes. A simulation with 500 replications was run for each size between 100 and 2000 by steps of 100.

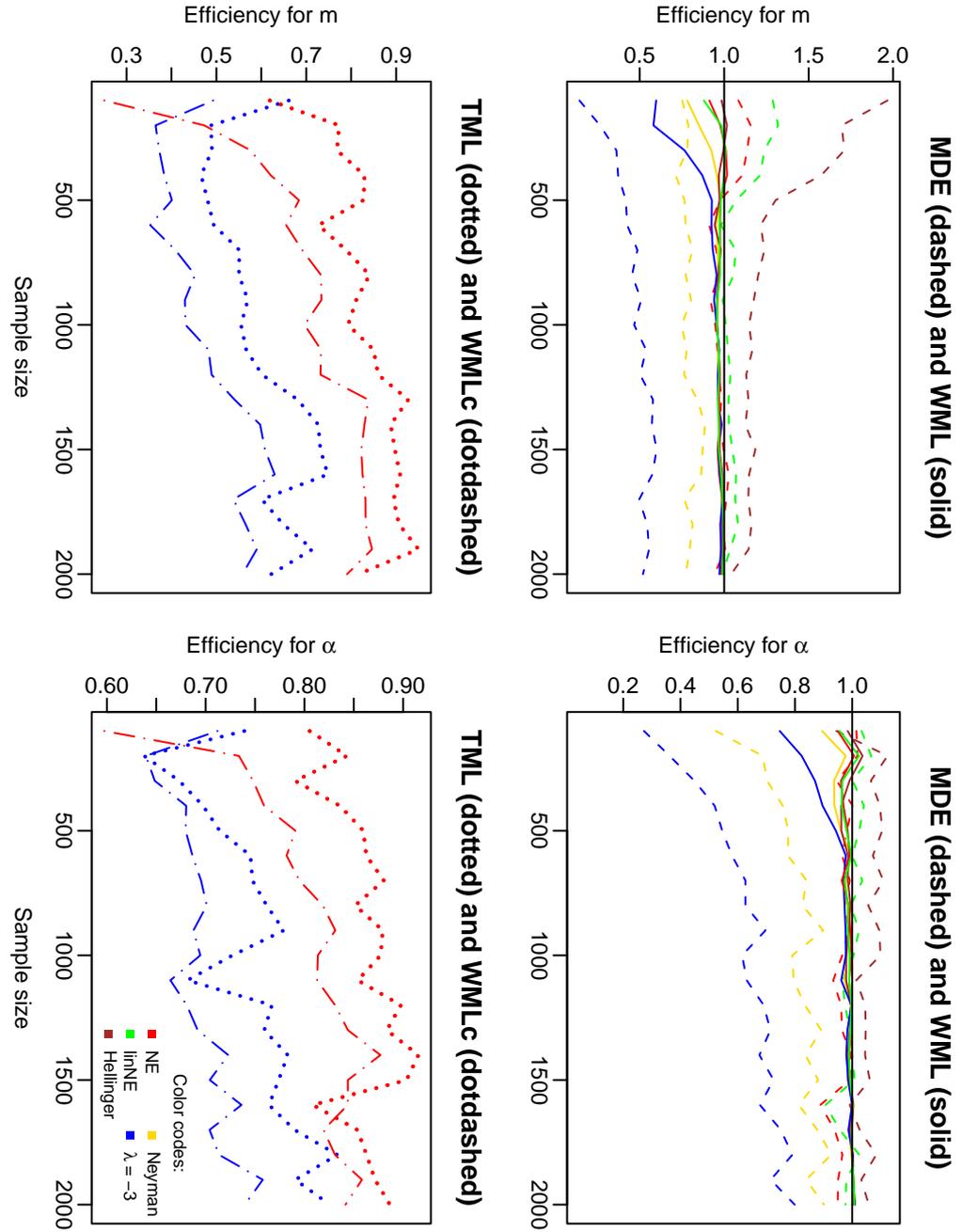


Figure D.5: “Standard” efficiencies. Simulations at the model $(m, \alpha) = (7, 1.5)$, for increasing sample sizes. A simulation with 500 replications was run for each size between 100 and 2000 by steps of 100.

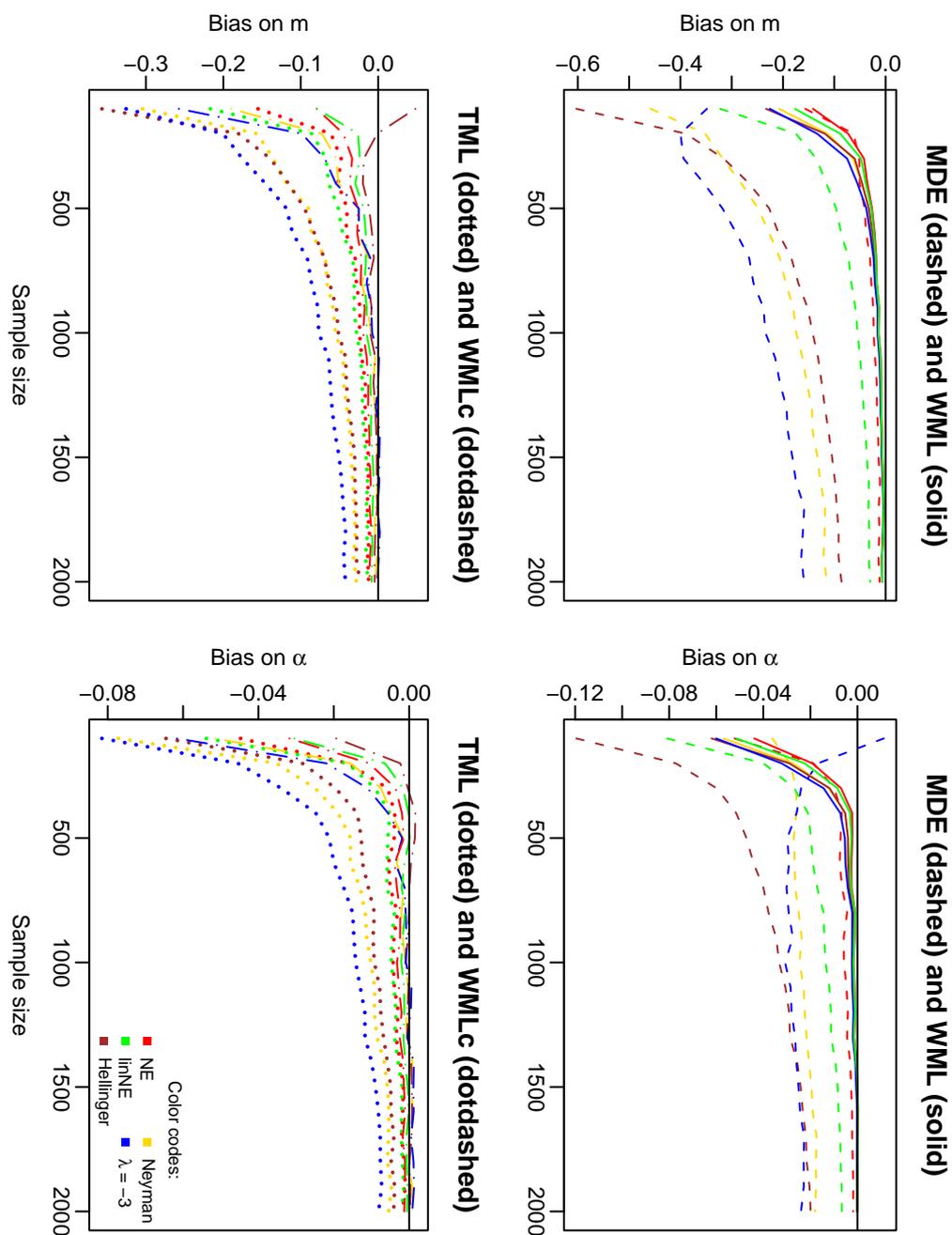


Figure D.6: Bias. Simulations at the model $(m, \alpha) = (5, 2/3)$, for increasing sample sizes. A simulation with 500 replications was run for each size between 100 and 2000 by steps of 100.

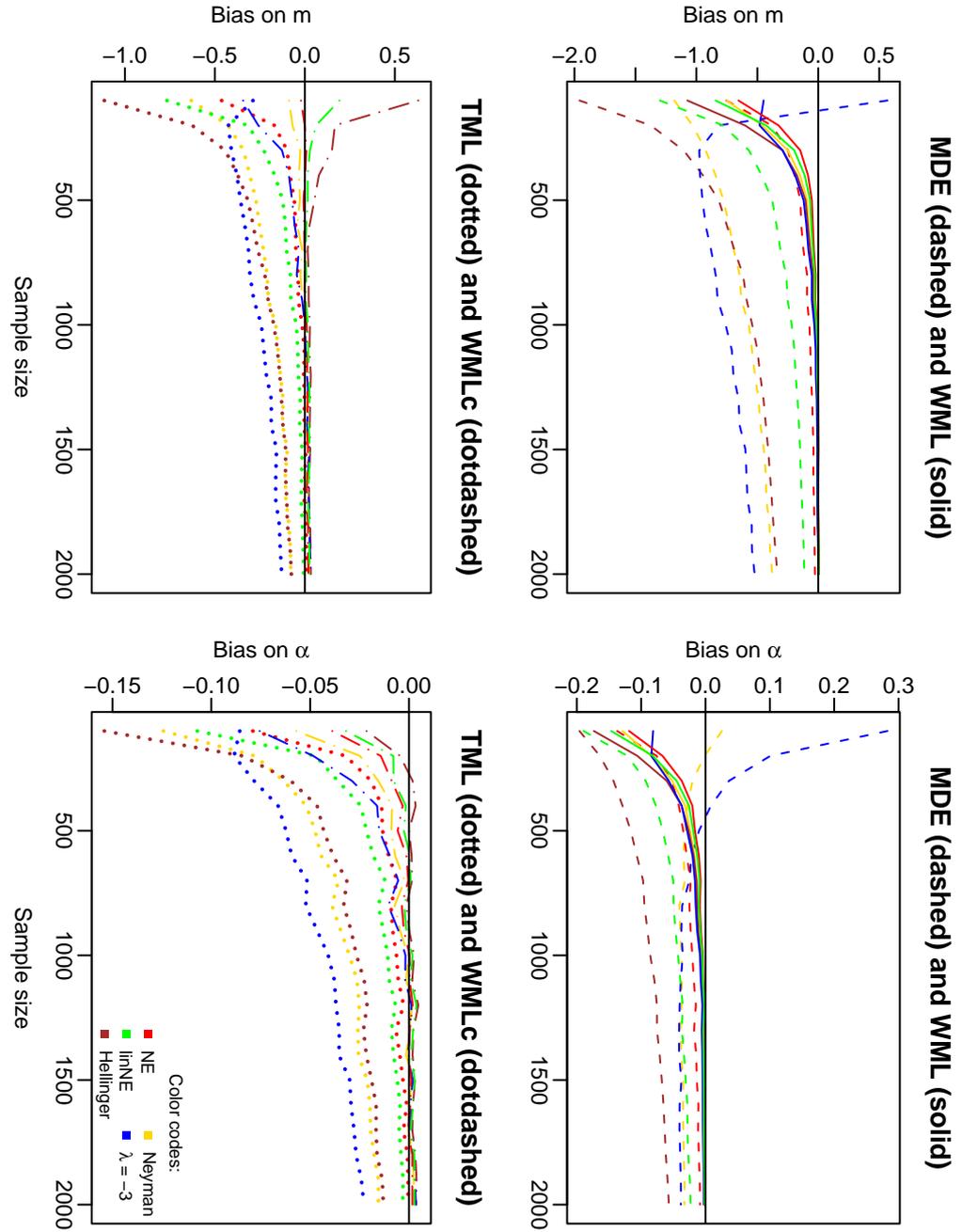


Figure D.7: Bias. Simulations at the model $(m, \alpha) = (7, 1.5)$, for increasing sample sizes. A simulation with 500 replications was run for each size between 100 and 2000 by steps of 100.

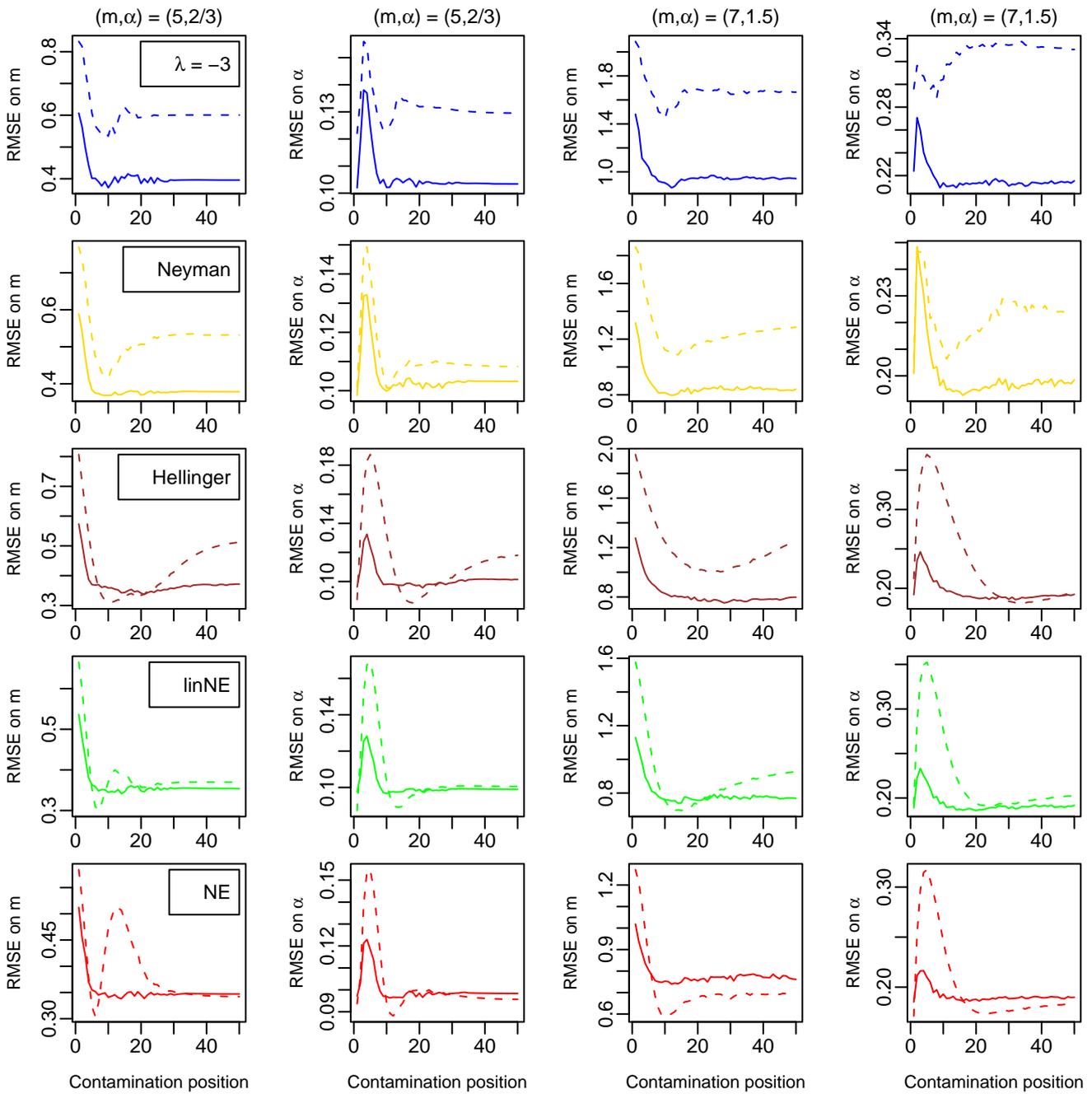


Figure D.8: RMSE of MDEs (dashed) and the corresponding WMLs (solid). Simulations at the point contaminated models $(m, \alpha) = (5, 2/3)$ and $(m, \alpha) = (7, 1.5)$. Contamination rate $\epsilon = 0.1$. A simulation with 500 replications was run for each contamination position between 1 and 50 by steps of 1.

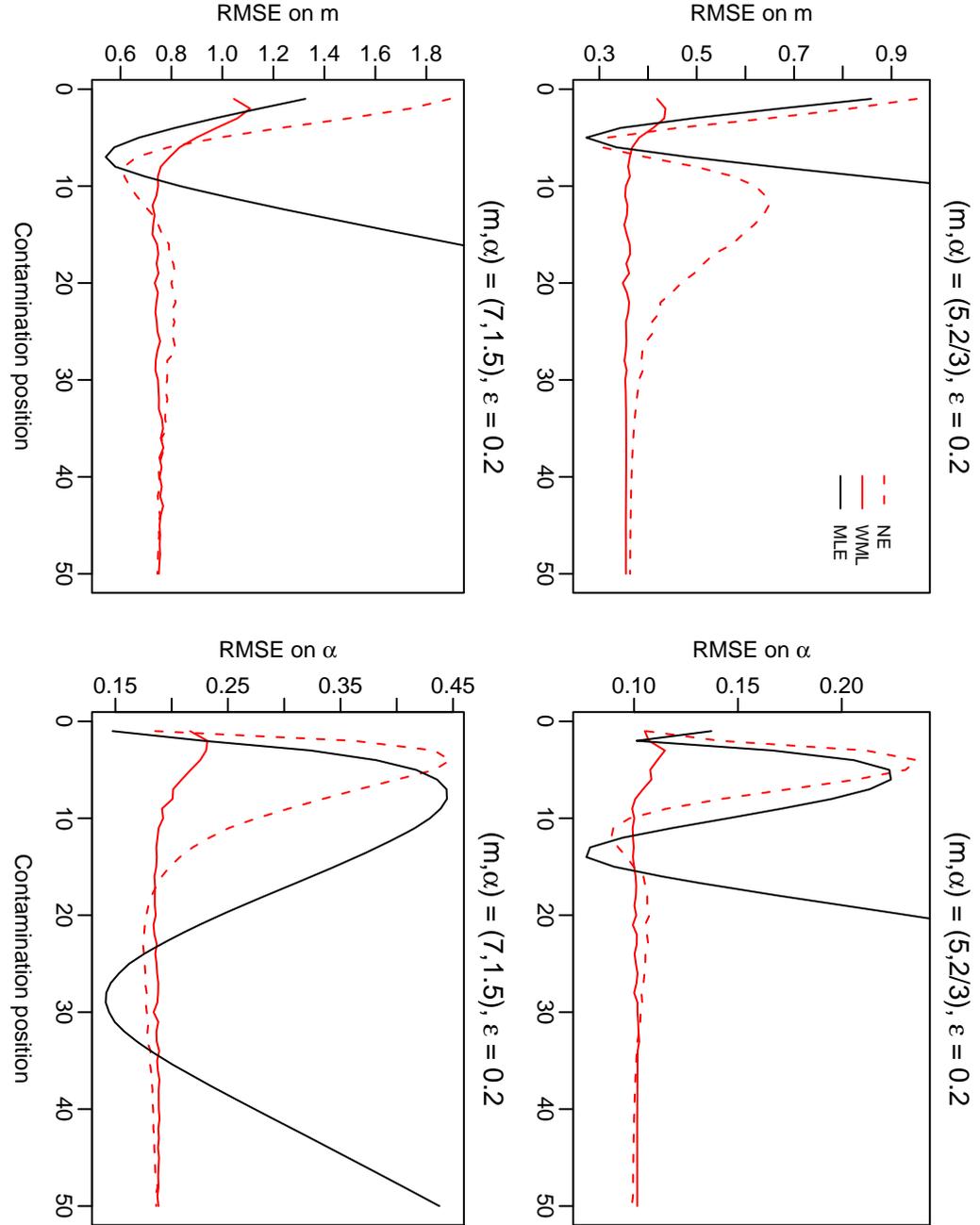


Figure D.9: RMSE of NE and the corresponding WML. Simulations at the point contaminated models $(m, \alpha) = (5, 2/3)$ and $(m, \alpha) = (7, 1.5)$. Contamination rate $\epsilon = 0.2$. A simulation with 500 replications was run for each contamination position between 1 and 50 by steps of 1.

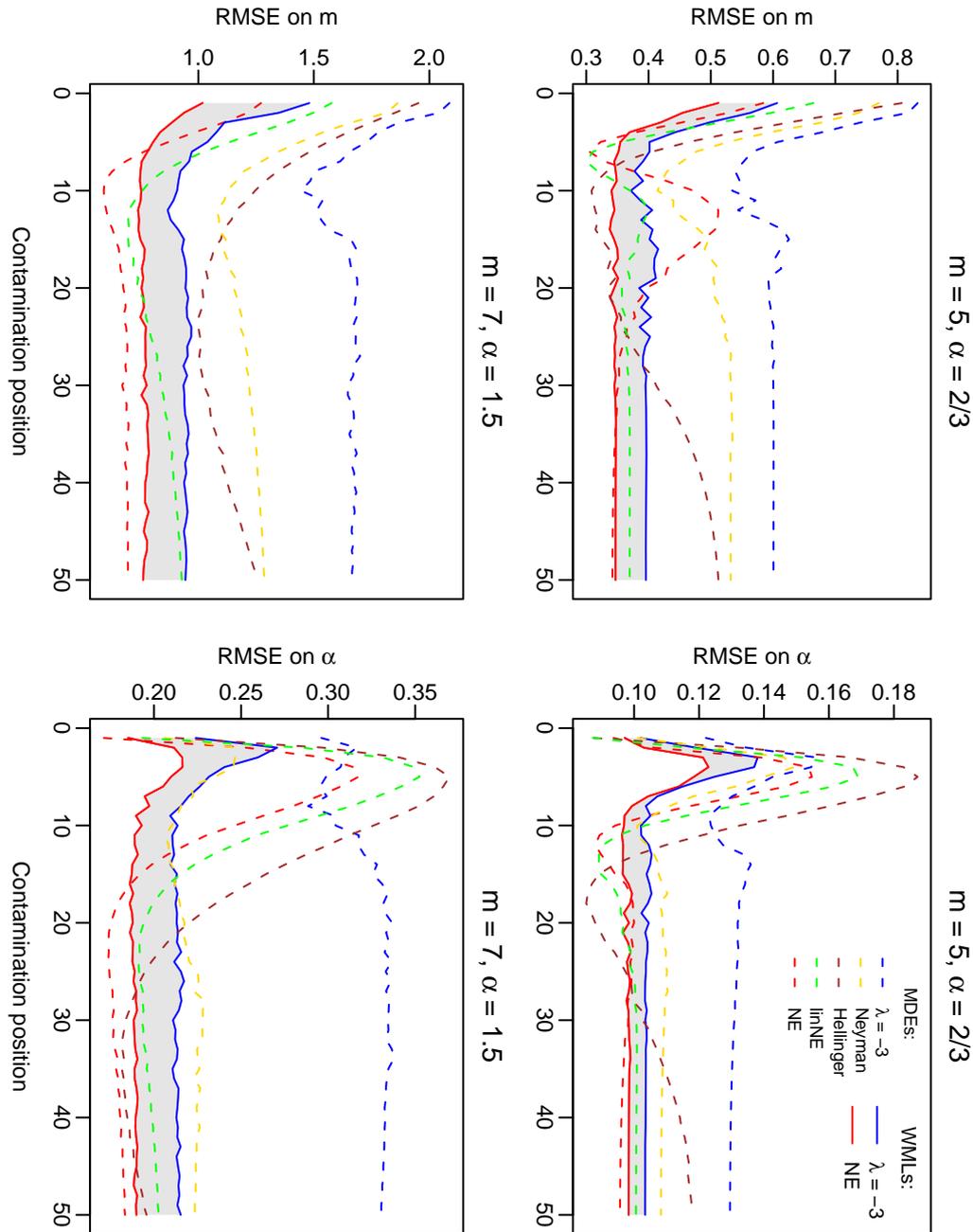


Figure D.10: RMSE of MDEs and WMLs. Simulations at the point contaminated models $(m, \alpha) = (5, 2/3)$ and $(m, \alpha) = (7, 1.5)$. Contamination rate $\epsilon = 0.1$. A simulation with 500 replications was run for each contamination position between 1 and 50 by steps of 1. The grey zone contains the curves of all WMLs.

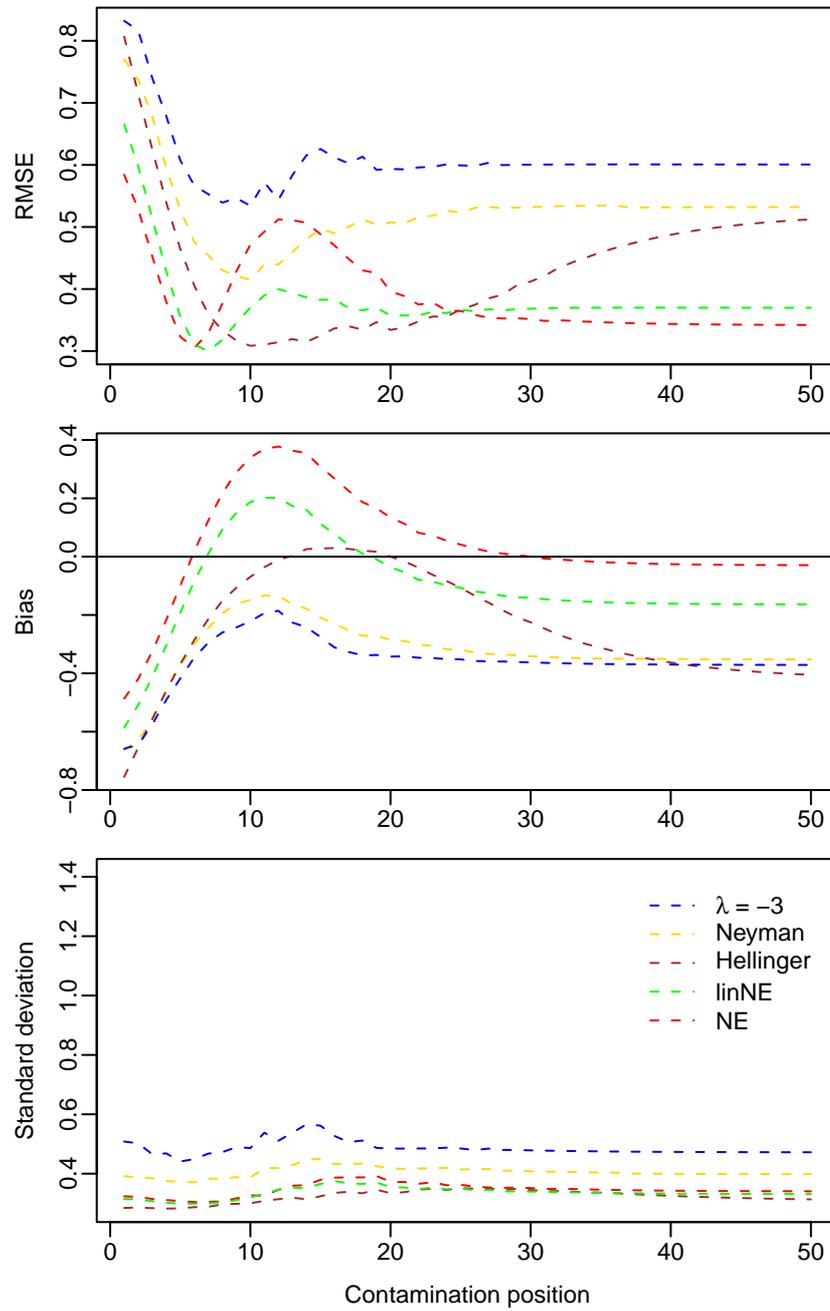


Figure D.11: RMSE, bias and standard deviation of the MDEs for the m parameter. Simulations at the point contaminated model $(m, \alpha) = (5, 2/3)$. Contamination rate $\epsilon = 0.1$. A simulation with 500 replications was run for each contamination position between 1 and 50 by steps of 1.

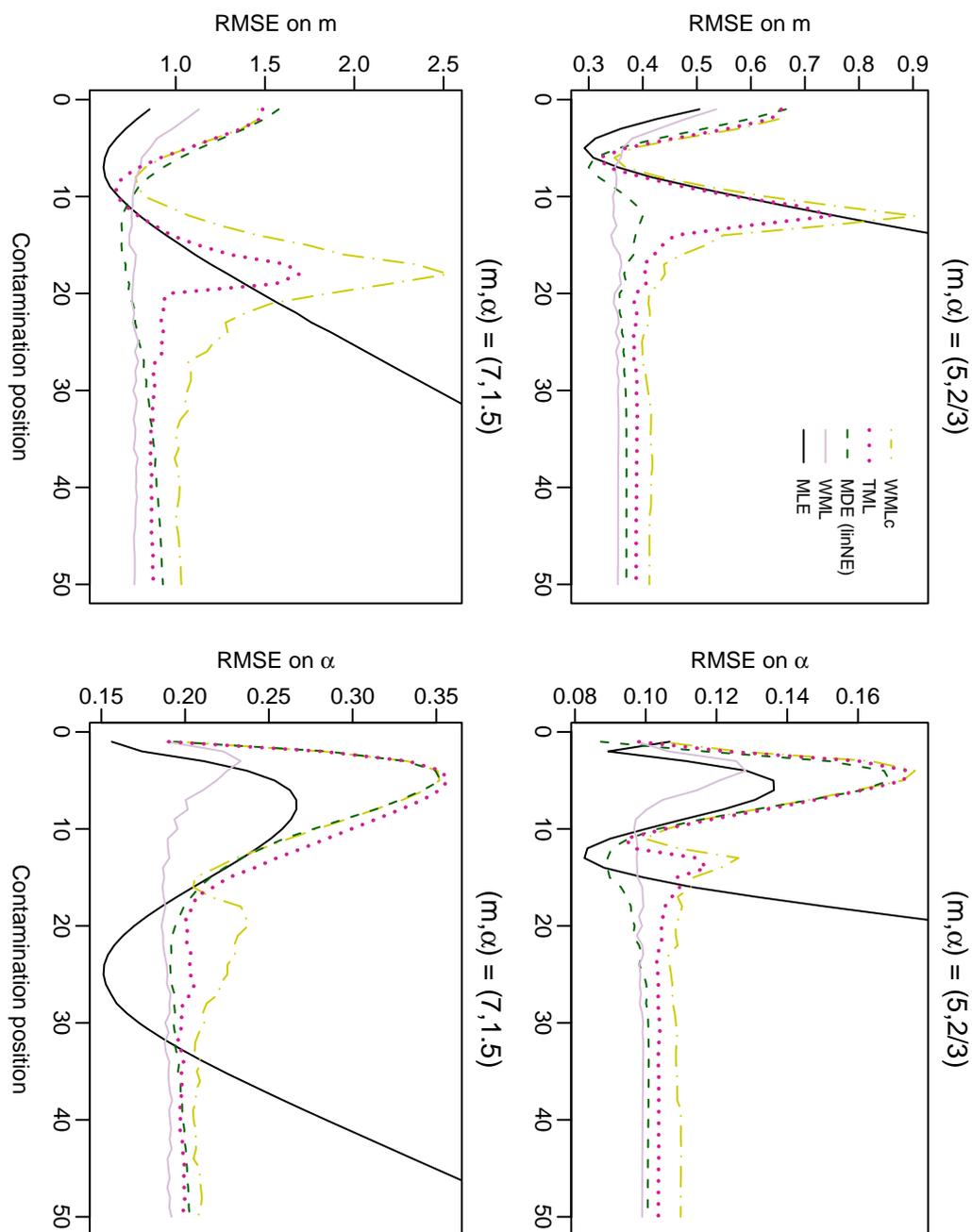


Figure D.12: RMSE of linNE and the corresponding WMLc, TML and WML. Simulations at the point contaminated models $(m, \alpha) = (5, 2/3)$ and $(m, \alpha) = (7, 1.5)$. Contamination rate $\epsilon = 0.1$. A simulation with 500 replications was run for each contamination position between 1 and 50 by steps of 1.

Bibliography

- M. Abramowitz and I.A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover publications, 1964.
- F.J. Anscombe. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 37(3-4):358, 1950.
- A. Basu and S. Basu. Penalized minimum disparity methods for multinomial models. *Statistica Sinica*, 8:841–860, 1998.
- A. Basu and B.G. Lindsay. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46(4):683–705, 1994.
- A. Basu and S. Sarkar. The trade-off between robustness and efficiency and the effect of model smoothing in minimum disparity inference. *Journal of Statistical Computation and Simulation*, 50(3):173–185, 1994.
- A. Basu, I.R. Harris, and S. Basu. Tests of hypotheses in discrete models based on the penalized Hellinger distance. *Statistics & probability letters*, 27(4):367–373, 1996.
- A. Basu, S. Basu, and G. Chaudhuri. Robust minimum divergence procedures for count data models. *Sankhyā: The Indian Journal of Statistics, Series B*, 59(1):11–27, 1997.

- R. Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463, 1977.
- J. Berkson. Minimum chi-square, not maximum likelihood! *The Annals of Statistics*, 8(3):457–487, 1980.
- S.K. Bhandari, A. Basu, and S. Sarkar. Robust inference in parametric models using the family of generalized negative exponential disparities. *Australian & New Zealand Journal of Statistics*, 48(1):95–114, 2006.
- J.F. Bithell. A class of discrete-time models for the study of hospital admission systems. *Operations Research*, 17(1):48–69, 1969.
- NG Cadigan and J. Chen. Properties of robust m-estimators for poisson and negative binomial data. *Journal of Statistical Computation and Simulation*, 70(3):273–288, 2001.
- AR Clapham. Over-dispersion in grassland communities and the use of statistical methods in plant ecology. *The Journal of Ecology*, 24(1):232–251, 1936.
- S.J. Clark and J.N. Perry. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics*, 45(1):309–316, 1989.
- N. Cressie and T.R.C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):440–464, 1984.
- I. Csiszár. Eine Informations theoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.*, 3:85–107, 1963.
- D. Gervini and V.J. Yohai. A class of robust and fully efficient regression estimators. *The Annals of Statistics*, 30(2):583–616, 2002.

- F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics: the approach based on influence functions*, volume 1. Wiley New York, 1986.
- I.R. Harris and A. Basu. Hellinger distance as a penalized log likelihood. *Communications in Statistics-Simulation and Computation*, 23(4):1097–1113, 1994.
- J. Hilbe. *Negative binomial regression*, volume 24. Cambridge University Press New York, 2007.
- J.F. Lawless. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15(3):209–225, 1987.
- B.G. Lindsay. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The annals of statistics*, 22(2):1081–1114, 1994.
- A. Marazzi and G. Barbati. Robust parametric means of asymmetric distributions: estimation and testing. *Estadística*, 54(162-163):47–72, 2003.
- A. Marazzi and C. Ruffieux. The truncated mean of an asymmetric distribution. *Computational Statistics & Data Analysis*, 32(1):79–100, 1999.
- A. Marazzi and V.J. Yohai. Adaptively truncated maximum likelihood regression with asymmetric errors. *Journal of statistical planning and inference*, 122(1-2):271–291, 2004.
- A. Marazzi and V.J. Yohai. Optimal robust estimates using the Hellinger distance. *Advances in Data Analysis and Classification*, pages 1–11, 2010. ISSN 1862-5347.
- R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust statistics*. Wiley New York, 2006.

- W.W. Piegorsch. Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, 46(3):863–867, 1990.
- C.R. Rao. Asymptotic efficiency and limiting information. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, volume 1, pages 531–545. Univ. California Press, Berkeley, 1961.
- C.R. Rao. Efficient estimates and optimum inference procedures in large samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(1):46–72, 1962.
- T.R.C. Read and N.A.C. Cressie. *Goodness-of-fit statistics for discrete multivariate data*. Springer, 1988.
- G.J.S. Ross and D.A. Preece. The negative binomial distribution. *The Statistician*, 34(3):323–335, 1985.
- D.G. Simpson. Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, pages 802–807, 1987.
- R.N. Tamura and D.D. Boos. Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81(393):223–229, 1986.
- UCLA (2010). Annotated stata output. University of California, Los Angeles: Academic Technology Services, Statistical Consulting Group. From http://www.ats.ucla.edu/stat/stata/output/Stata_ztnb.htm (accessed November 14, 2010).
- RC Woodruff, JM Mason, R. Valencia, and S. Zimmering. Chemical mutagenesis testing in *Drosophila*: I. Comparison of positive and negative control data for sex-linked recessive lethal mutations and reciprocal translocations in three laboratories. *Environmental mutagenesis*, 6(2):189–202, 1984.