JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles

An Information Theoretical approach to Factor Analysis

François Bavaud

IMM - Lettres - Université de Lausanne - CH-1015 Lausanne - Switzerland

Abstract

We show how the introduction of the power divergence family proposed by Cressie and Read (1984) permits to link various aspects of log likelihood model selection and factorial data description. Our approach, illustrated on bigram textual frequencies, generalizes Factorial Correspondance Analysis beyond the independence model, as exemplified by the symmetry model and an "independence-within classes" model, the latter seeming promising for classification purposes. We introduce a "psi square" measure of inertia, alternative to the usual phi square. The concept of "sharp contradiction" as well as a presumably new Rényi-like measure of dependence are discussed in the framework of Information Theory. An "eigenvalues doubling" phenomenon associated to the symmetry model is elucidated.

Keywords: Entropy, Factorial Correspondance Analysis, independence-within-class model, Kullback-Leibler divergence, log likelihood, marginal homogeneity, model selection, power divergence family, Rényi's entropy, symmetry model, textual data analysis, variety.

1. Introduction

Markov chain models, Information Theory and Factorial Correspondance Analysis (FCA) share a remarkable feature, namely to have first emerged as solutions of statistical problems about textual data: Markov (1913) about the quantification of the consonants/vowels sequences in Russian; Shannon (1951) about the entropy of written English; Benzécri (cited in Greenacre (1984) p.9) about the consonants/vowels contingency tables in Chinese modern language manuals.

Hierarchical classification methods aside, French research on textual data mainly relies upon FCA (as e.g. attested in Lebart and Salem (1994)) while Information Theory is the most popular tool in Anglo-Saxon research (as e.g. attested in Manning and Schütze (1999)). With the hope of a better understanding of both approachs, we present a framework originally aimed at linking FCA to Information Theory.

Typical information theoretical expressions, such as the relative entropy (or Kullback-Leibler dissimilarity) do not lend themselves to factorial decomposition. However, the relative entropy is just one member among the power divergence family $\{I_s\}$ proposed by Cressie and Read (1984); on the other hand, FCA is nothing but factor analysis on contingency tables for a particular model (namely the independence model) and a particular "total variance" measure, namely the phi square, also belonging to the power divergence family.

Those circumstances enable to compare information theoretical expressions (obtained for $s = 0$ or $s = -1$) to factorial, data analytical formulations (obtained for $s = 1$); we will also meet another measure (closely related to the Freeman-Tuckey or Escofier (1978) dissimilarity) obtained for $s = -1/2$, we shall call "psi square", which permits another approach to factor analysis, alternative to the traditional practice based upon the phi square.

We are of course well aware that the models we are discussing here (such as bigram independence or symmetry) are little adapted to texts: they are simply aimed at illustrating methodological points on familiar data and familiar models. Also, our choice of units (i.e. letters instead of words) can be criticized from a *modelling* point of view (realistic models for sequences of words are arguably easier to produce by an human subject than realistic models for sequences of letters), but not, in our opinion, from the point of view of *information*, since entropy-like quantities can be converted without loss from a categorization system to another. For instance, the *entropy rate* (see e.g. Cover and Thomas (1991)) satisfies

$$\text{entropy rate per word} = \text{entropy rate per letter} \times \text{average number of letters per word}$$

2. Information theory and model selection

Notations: let n_{jk} be an $(m_1 \times m_2)$ contingency table, with relative frequency $f_{jk} := n_{jk}/n$, row profiles $w_{jk} := n_{jk}/n_{j\bullet}$, column profiles $w^*_{jk} := n_{kj}/n_{\bullet j}$ and marginal profiles $\rho^*_j := n_{j\bullet}/n = f_{j\bullet}$ and $\rho_k := n_{\bullet k}/n = f_{\bullet k}$, where $n := n_{\bullet\bullet}$ is the grand total. By construction, $f_{jk} = \rho^*_j w_{jk} = \rho_k w^*_{kj}$; also, the row and column profiles transform as $w_{jk} = \rho_k w^*_{kj}/\rho^*_j$ and $w^*_{kj} = \rho^*_j w_{jk}/\rho_k$, which simply expresses Bayes' rule on conditional profiles w_{jk} and w^*_{kj}.

Entropy. $H(_{\text{column}}) := -\sum_k \rho_k \ln \rho_k \leq \ln m_2$ is the entropy on columns, and $H(_{\text{row}}) := -\sum_j \rho^*_j \ln \rho^*_j \leq \ln m_1$ is the entropy on rows. $H(_{\text{column}}|j) := -\sum_k w_{jk} \ln w_{jk} \leq \ln m_2$ is the conditional entropy on columns given row j and $H(_{\text{column}}|_{\text{row}}) := \sum_j \rho^*_j H(_{\text{column}}|j)$ is the conditional entropy on columns given the rows. Similarly, $H(_{\text{row}}|k) := -\sum_j w^*_{kj} \ln w^*_{kj} \leq \ln m_1$ is the conditional entropy on rows given column k and $H(_{\text{row}}|_{\text{column}}) := \sum_k \rho_k H(_{\text{row}}|j)$ is the conditional entropy on rows given the columns. Also, $H(_{\text{row, column}}) := \sum_{jk} f_{jk} \ln f_{jk}$ is the total entropy. Simple algebra yields the well-known relations:

$$H(_{\text{column}}|_{\text{row}}) = H(_{\text{row, column}}) - H(_{\text{row}}) \qquad H(_{\text{row}}|_{\text{column}}) = H(_{\text{row,column}}) - H(_{\text{column}}) \qquad (1)$$

Kullback-Leibler divergence. The canonical information-theoretical measure of dissimilarity between two theories f and g, supposed here defined by discrete distribution probabilities on modalities i as $f_i \geq 0$ with $\sum_i f_i = 1$ and $g_i \geq 0$ with $\sum_i g_i = 1$ is the Kullback-Leibler dissimilarity $K(f||g) := \sum_i f_i \ln(f_i/g_i)$. The functional $K(f||g)$ is non-negative, asymmetric, with the property $K(f||g) = 0$ iff $f \equiv g$. It can be interpreted as a measure of the information gained (or the surprise generated) when the distribution f replaces the prior distribution g. Its form can be justified from many points of view (see e.g. Cover and Thomas (1991)); for instance, maximum likelihood estimation f^{theo} obtains from the data (specified by the empirical distribution f) as well as from the model (specified by a family of distributions $f(\theta)$ possessing $\dim(\Theta)$ free parameters $\theta \in \Theta$) as

$$f^{\text{theo}} = f(\theta_0) \qquad \text{where} \quad K(f||f(\theta_0)) = \min_{\theta\in\Theta} K(f||f(\theta)) \qquad (2)$$

Also, the p-value associated to the test of H_0 : "data follow model g" asymptotically behaves as $p \sim \exp(-n\, K(g^\star||g))$, where $g^\star$ is the true theoretical distribution and n the sample size. The p-value thus decays exponentially whenever $0 < K(g^\star||g) < \infty$. When $K(g^\star||g) = 0$, the tested theory g turns out to be the true one $g^\star$ and p should not decrease with the sample size, as expected. Oppositely, if $g^\star$ *sharply contradicts* g, namely if there exists an outcome i_0 held for impossible by the tested theory g (i.e. $g_{i_0} = 0$) but actually possible (i.e. $g^\star_{i_0} > 0$), then,

sooner or later, theory g should be eliminated consecutively to the observation of outcome i_0 (deterministic or Poperian refutation). Satisfactorily enough, one gets $K(g^\star||g) = \infty$ in that case, meaning that the p-value asymptotically decays faster than exponentially.

Model selection. Maximum likelihood model selection consists in computing $L^2(H_0) := 2nK(f||f^{\text{theo}})$, and comparing its value to the threshold $\chi^2[\text{df}]_{1-\alpha}$, where $\text{df} = \dim(\text{data}) - \dim(\Theta)$ is the difference between the number of parameters $\dim(\text{data})$ of the saturated model fitting perfectly the data and the number of free parameters $\dim(\Theta)$ available in the model $f(\theta)$. That is, df is the number of constraints expressed in H_0 : "data follow model $f(\theta)$, where $\theta \in \Theta$". Model H_0 survives at level α as long as

$$2nK(f||f^{\text{theo}}) \leq \chi^2[\text{df}]_{1-\alpha} \qquad (\text{or } 2nK(f^{\text{theo}}||f) \stackrel{\sim}{\leq} \text{df} \quad \text{in the simplified version}) \tag{3}$$

Example 1: independence model. For the independence model $H_0 = H_{\text{IND}}$, the expected frequencies (2) are $f_{jk}^{\text{theo}} = \rho_j^* \rho_k$, and the corresponding Kullback-Leibler dissimilarity is thus

$$K(f||f^{\text{theo}}) = \sum_{jk} f_{jk} \ln \frac{f_{jk}}{\rho_j^* \rho_k} = H(_{\text{row}}) + H(_{\text{column}}) - H(_{\text{row, column}}) \tag{4}$$

As an illustration, consider the contingency table n_{jk} counting the number of bigrams appearing in the first $n = 15'442$ characters of the French text "La pensée remonte les fleuves" by C.F.Ramuz (1937). Suppressing separators with the exception of the blank character "_", accents and case, we are left with $m_1 = m_2 = 26$ categories (namely "_" together with 25 letters, "k" having no occurrences in the text).

Rows and columns formally coincide. Thus both w_{jk} and w_{kj}^* can be regarded as Markov transition matrices, describing the first-order generation of symbols given the previous one (resp. the next one). The text begins and ends with a blank, and thus satisfies marginal homogeneity, namely $n_{j\bullet} = n_{\bullet j}$. Consequently, $\rho_j^* = \rho_j$, the latter also constituting the stationary distribution of w_{jk} or w_{kj}^* (Bavaud 1998). While $H(_{\text{column}}|j)$ and $H(_{\text{row}}|j)$ do not coincide in general (for instance, $H(_{\text{column}}|\text{"q"}) = 0$ since "q" is always followed by "u", but $H(_{\text{rows}}|\text{"q"}) = 0.69 > 0$ since "q" can follow different symbols), their averages $H(_{\text{column}}|_{\text{row}})$ and $H(_{\text{row}}|_{\text{column}})$ do, with value 2.14. As $H(_{\text{row}}) = 2.70$, one gets $K(f||f^{\text{theo}}) = 2.70 - 2.14 = 0.56$.

The corresponding log likelihood is $L^2(H_{\text{IND}}) = 2nK(f||f^{\text{theo}}) = 17'371.2$ ($\text{df} = 625$): as we well know, successive symbols in a text are highly dependent.

Equation (4) can be generalized by introducing **Rényi's entropy** H_α of parameter $\alpha \in (0, 1)$:

$$H_\alpha(f) := \tfrac{1}{1-\alpha} \ln \textstyle\sum_i f_i^\alpha \tag{5}$$

The interested reader will find helpful to use the freeware *Entropizer 1.1* of A.Xanthos (2000), computing transition tables as well as Rényi's and Shannon's entropies of different orders. From inequality $\sum_{jk} f_{jk}^\alpha \leq (\sum_j f_{j\bullet}^\alpha)(\sum_k f_{\bullet k}^\alpha)$, the quantity

$$R_\alpha(f) := H_\alpha(\rho^*) + H_\alpha(\rho) - H_\alpha(f) = H_\alpha(_{\text{row}}) + H_\alpha(_{\text{column}}) - H_\alpha(_{\text{row, column}}) \tag{6}$$

is non-negative, with value zero iff $f_{jk} = \rho_j^* \, \rho_k$. Thus $R_\alpha(f)$ constitutes a suitable measure of dependence. The limit $\lim_{\alpha \to 1} H_\alpha(f) = H(f)$ yields Shannon entropy again. The limit $\lim_{\alpha \to 0} H_\alpha(f) = \ln V(f)$ makes appear the *variety* $V(f)$ of the system, i.e. the number of distinct categories i such that $f_i > 0$. In this case, (6) simply says that $\ln V(f) \leq \ln V(\rho^*) + \ln V(\rho)$

or equivalently $V(_{\text{row, column}}) \leq V(_{\text{row}})\, V(_{\text{column}})$: the number of distinct cross-modalities observed in the contingency table n_{jk} cannot exceed the number of observed rows times the number of observed columns. Note that $H_\alpha(f)$ and $R_\alpha(f)$ somewhat interpolate between "qualitative measures" for $\alpha = 0$ (taking only into account the presence/absence of a category) and "quantitative measures" for $\alpha > 0$ (taking into account the relative frequency of a category).

Example 2: independence-within-classes model. Suppose vowels on one hand and consonants on the other hand are equivalent to the extent to be entirely substitutable by each other. More generally, consider the set of $m_1 = m_2 =: m$ categories to be partitioned into $M_1 = M_2 =: M < m$ classes, and suppose the counts n_{jk}^{theo} to be independent *conditionally* to the belonging of symbols j and k in classes $J(j)$ and $K(k)$ respectively ($J, K = 1, \ldots, M$). Explicitly, this *independence-within-classes* model $H_0 = H_{\text{IWC}}$ assumes $n_{jk}^{\text{theo}} = \alpha_j \beta_k \gamma_{J(j)K(k)}$. Using notational conventions such as $n_{JK} := \sum_{j \in J\,;\, k \in K} n_{jk}$ and $n_{J\bullet} := \sum_{j \in J\,;\, k} n_{jk}$, ML-estimation (2) yields:

$$n_{jk}^{\text{theo}} = n f_{jk}^{\text{theo}} = \frac{n_{j\bullet}}{n_{J(j)\bullet}} \frac{n_{\bullet k}}{n_{\bullet K(k)}} n_{J(j)K(k)} \tag{7}$$

Therefore, the Kullback-Leibler expresses as

$$K(f||f^{\text{theo}}) = \sum_{jk} \frac{n_{jk}}{n} \ln \frac{n_{jk}}{n_{jk}^{\text{theo}}} = \sum_{jk} \frac{n_{jk}}{n} \ln \frac{n_{jk}}{n_{j\bullet}\, n_{\bullet k}} - \sum_{JK} \frac{n_{JK}}{n} \ln \frac{n_{JK}}{n_{J\bullet}\, n_{\bullet K}} \tag{8}$$

or equivalently $L^2(H_{\text{IWC}}) = L^2(H_{\text{IND, symbols}}) - L^2(H_{\text{IND, classes}})$. The corresponding degrees of freedom are readily found to be $\text{df} = (m-1)^2 - (M-1)^2$.

Considering in our text sample the three groups {blank} (J=1), "vowels"={a, e, i, o, u, y} (J=2) and "consonants" (J=3) comprising all the other symbols, one gets $L^2(H_{\text{IND, classes}}) = 3'934.8$, and thus $L^2(H_{\text{IWC}}) = 17'371.2 - 3'934.8 = 13'436.4$ with $\text{df} = (26-1)^2 - (3-1)^2 = 621$. While the proposed partitioning is too rough to withstand empirical confrontation, equation (8) can clearly serve at constructing a well-defined hierarchical classification scheme.

Example 3: symmetry model. ML-estimation of the expected frequencies under the symmetry model $H_0 = H_{\text{SYM}}$ are well known to be $f_{jk}^{\text{theo}} = f_{kj}^{\text{theo}} = (f_{jk} + f_{kj})/2$. One finds $K(f||f^{\text{theo}}) = .21$ and $L^2(H_{\text{SYM}}) = 6'337.9$ with $\text{df} = 26(26-1)/2 = 325$. Texts being not invariant by time-reversal, the rejection of the symmetry model hardly comes as a surprise.

3. Factorial data analysis

Linking model selection and factor analysis: the power divergence family. Factor analytic methods in data analysis consist in spectrally decomposing a sum of squares generally interpretable as a total variance or total inertia. The Kullback-Leibler dissimilarity $K(f||g)$ does not expresses as a sum of squares; however, it belongs to the *power divergence* family

$$I_s(f:g) := \tfrac{1}{s(s+1)} \textstyle\sum_i f_i((\frac{f_i}{g_i})^s - 1) \tag{9}$$

where s is a real parameter (Cressie and Read (1984)). Specifically, $I_0(f:g) = K(f||g)$ and $I_{-1}(f:g) = K(g||f)$ (more generally, $I_s(f:g) = I_{-s-1}(g:f)$). Moreover, other well-known functionals obtain for particular values of s, namely (in order) the (ordinary) khi-square, the Freeman-Tuckey statistic and the Neyman khi-square:

$$I_1(f:g) = \tfrac{1}{2} \textstyle\sum_i \frac{(f_i - g_i)^2}{g_i} \quad I_{-1/2}(f:g) = 2 \textstyle\sum_i (\sqrt{f_i} - \sqrt{g_i})^2 \quad I_{-2}(f:g) = \tfrac{1}{2} \textstyle\sum_i \frac{(f_i - g_i)^2}{f_i} \tag{10}$$

In particular, those three expressions constitute sum of squares (and the only ones identified so far in the power divergence family) on which factor analysis can be performed. Power divergence functionals are "H_0-equivalent" in the sense that, irrespectively of the value of s, $2n\, I_s(f:g)$ asymptotically follows a khi-square distribution when g is the true distribution and f the empirical distribution. However, if data f sharply contradict g, then $I_s(f:g)=\infty$ holds for $s \geq 0$ only; similarly, $I_s(f:g)=\infty$ whenever model g sharply contradicts data f, provided $s \leq -1$: for that range of values, theories predicting unobserved outcomes are rejected.

Factor decomposition of the khi square and the "psi square". Let f_{jk} be the observed distribution, and f_{jk}^{theo} the associated theoretical distribution under some model H_0. Define the $(m_1 \times m_2)$ matrices

$$c_{jk} := (f_{jk} - f_{jk}^{\text{theo}})/\sqrt{f_{jk}^{\text{theo}}} \qquad\qquad \tilde{c}_{jk} := 2(\sqrt{f_{jk}} - \sqrt{f_{jk}^{\text{theo}}}) \tag{11}$$

as well as the $(m_1 \times m_1)$ matrices $B := CC'$ and $\tilde{B} := \tilde{C}\tilde{C}'$. By construction, B and $\tilde{B}$ are symmetric and positive definite, thus decomposable as $B = U\Lambda U'$ and $\tilde{B} = \tilde{U}\tilde{\Lambda}\tilde{U}'$. On the other hand, consider a set $\{X_j\}_{j=1,\ldots,m_1}$ (resp. $\{\tilde{X}_j\}_{j=1,\ldots,m_1}$) of, say, normally distributed vectors with variance-covariance matrix B (resp. $\tilde{B}$) and zero mean. Factor analysis of B and $\tilde{B}$ consists in spectrally decomposing the total variances, namely

$$\textstyle\sum_j \text{var}(X_j) = \text{trace}(B) = \sum_\alpha \lambda_\alpha = \sum_{j,k} \frac{(f_{jk}-f_{jk}^{\text{theo}})^2}{f_{jk}^{\text{theo}}} = 2\, I_1(f, f^{\text{theo}}) \quad \text{(phi square = khi square / } n\text{)}$$

$$\textstyle\sum_j \text{var}(\tilde{X}_j) = \text{trace}(\tilde{B}) = \sum_\alpha \tilde{\lambda}_\alpha = 4\sum_{j,k}(\sqrt{f_{jk}} - \sqrt{f_{jk}^{\text{theo}}})^2 = 2\, I_{-1/2}(f, f^{\text{theo}}) \quad \text{("psi square")}$$

Thus *any* model H_0 relative to a contingency table can be factor-analyzed by using one of the two decompositions above (corresponding to $s=1$ or $s=-1/2$ in (9): see Escofier (1978) for the latter case. The case $s=-2$ is not considered here, since any empty cell associated with a non-zero expected count would sharply reject the model). The procedure decomposes the deviations of f_{jk} from f_{jk}^{theo}, i.e. the deviations of the data from the model H_0, into independent components.

Usual computations and interpretation rules apply. The α-th factor scores column (of variance λ_α) obtains as $F_\alpha := \sum_j X_j\, u_{j\alpha}$, the cross-covariances as $\text{cov}(X_j, F_\alpha) = \lambda_\alpha u_{j\alpha}$ and the saturations (loadings) as

$$s_{j\alpha} = \text{corr}(X_j, F_\alpha) = \frac{\sqrt{\lambda_\alpha}}{\sqrt{b_{jj}}}\, u_{j\alpha} \qquad\qquad \sum_\alpha s_{j\alpha} s_{j'\alpha} = \text{corr}(X_j, X_{j'}) \tag{12}$$

(analogous results hold for the psi square decomposition (12)). The sum rules $\sum_\alpha s_{j\alpha}^2 = 1$ and $\sum_j b_{jj}\, s_{j\alpha}^2 = \lambda_\alpha$ permit to define contributions of the factors or dimensions to the variance of the variables and vice-versa. In particular, $\lambda_\alpha / \sum_\beta \lambda_\beta$ (resp. $\tilde{\lambda}_\alpha / \sum_\beta \tilde{\lambda}_\beta$) is the proportion of the total divergence $I_1(f:f^{\text{theo}})$ (resp. $I_{-1/2}(f:f^{\text{theo}})$) explained by dimension α.

Considering column instead of row profiles would lead to define m_2 variables $Y_1, \ldots, Y_{m_2}$ of variance-covariance $(m_2 \times m_2)$ matrix $B_Y := C'C$ (or $\tilde{B}_Y := \tilde{C}'\tilde{C}$). As $B_X := B = CC'$, normalized eigenvalues $v_{k\alpha}$ of B_Y are related to normalized eigenvalues $u_{j\alpha}$ of B_X by

$$v_\alpha = \tfrac{1}{\sqrt{\lambda_\alpha}} C' u_\alpha \qquad\qquad u_\alpha = \tfrac{1}{\sqrt{\lambda_\alpha}} C v_\alpha \tag{13}$$

for the same eigenvalue λ_α. Corresponding saturations obtain as $s_{k\alpha}^* = \frac{\sqrt{\lambda_\alpha}}{\sqrt{b_{Y\ kk}}}\, v_{k\alpha}$.

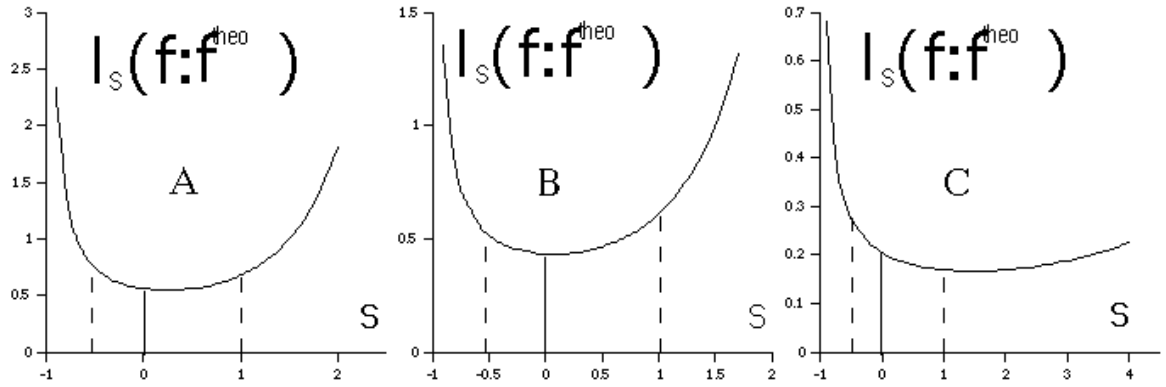

Figure 1: *power divergence $I_s(f : f^{theo})$ where f^{theo} is the ML-estimate $s = 0$ corresponding to models H_{IND} (A), H_{IWC} (B) and H_{SYM} (C). The value at $s = 1$ gives the phi square; $s = -1/2$ gives the psi square.*

This generalizes to arbitrary models H_0 the well-known results of FCA, the latter covering the case $H_0 = H_{\text{IND}}$ only. Note that f^{theo} above is the ML-estimate under H_0, minimizing $I_0(f : f^{\text{theo}}) = K(f||f^{\text{theo}})$. It is *not* the minimizer of $I_s(f : f^{\text{theo}})$ for $s = -1/2$ or $s = 1$, although such a specification would have been perfectly possible also, with still another resulting factorial representation; see Bavaud (2000 b) for an example bearing upon the model of quasi-symmetry. Besides computational convenience, our choice simply matches the usual practice in khi square testing or FCA. Figure 1 depicts the near constancy of $I_s(f : f^{\text{theo}})$ in the range $-1/2 \leq s \leq 1$.

Figure 2 shows the row (or column) saturations $s_{j\alpha}$ associated to the three models, in the phi square or psi square version. Another representation, generalizing the usual practice in FCA (see e.g. Saporta (1990) or Lebart et al. (1995)), consists in defining factorial coordinates for row j by $\psi_{j\alpha} := \sqrt{\alpha_j}\, s_{j\alpha}$ or $\tilde{\psi}_{j\alpha} := \sqrt{\tilde{\alpha}_j}\, \tilde{s}_{j\alpha}$ where α_j and $\tilde{\alpha}_j$ are the *atypicities* defined as

$$\alpha_j := \sum_k \frac{(w_{jk} - w_{jk}^{\text{theo}})^2}{w_{jk}^{\text{theo}}} \quad \text{(phi square)} \qquad \tilde{\alpha}_j := 4 \sum_k (\sqrt{w_{jk}} - \sqrt{w_{jk}^{\text{theo}}})^2 \quad \text{(psi square)}$$

and $w_{jk}^{\text{theo}} := f_{jk}^{\text{theo}} / \rho_j^*$. One can check marginal homogeneity of our data to insure $w_{j\bullet}^{\text{theo}} = 1$ in the three models, although $w_{j\bullet}^{\text{theo}} \neq 1$ in general. ψ-coordinates permit to express total divergence as an *inertia*, i.e. as a weighted origin-row squared euclidean distance:

$$\sum_j \rho_j^* \sum_\alpha \psi_{j\alpha}^2 = 2\, I_1(f : f^{\text{theo}}) \qquad \sum_j \rho_j^* \sum_\alpha \tilde{\psi}_{j\alpha}^2 = 2\, I_{-1/2}(f : f^{\text{theo}})$$

$\psi_{j\alpha}$ and $\tilde{\psi}_{j\alpha}$ represent *residuals* with respect to the model under consideration: $\psi_j = 0$ or $\tilde{\psi}_j = 0$ iff $w_{jk} = w_{jk}^{\text{theo}}$ for all k. More on inertia (in particular on aggregation invariance, scaling properties and Huygens' principle for dissimilarities) can be found in Bavaud (2000 a).

The phi square decomposition of the symmetry model H_{SYM} produces an noticeable phenomenon, namely an *eigenvalues doubling*: one finds indeed that $\lambda_1 = \lambda_2 \geq \lambda_3 = \lambda_4 \geq \lambda_5 = \lambda_6 \geq \ldots$ (where the last eigenvalue is zero in case of an odd number of categories m). The explanation is the following: in the phi square version, $c_{jk} = (f_{jk} - f_{jk})/\sqrt{2(f_{jk} + f_{jk})}$ and thus $C' = -C$. Then if u_α is an eigenvalue of $B_X = CC'$ for the value λ_α, so is Cu_α since

$$B_X C u_\alpha = -CCCu_\alpha = CCC'u_\alpha = C\lambda_\alpha u_\alpha = \lambda_\alpha C u_\alpha$$

On the other hand, Cu_α, proportional to v_α by (13), is generally distinct from u_α, whence the doubling of eigenvalues.

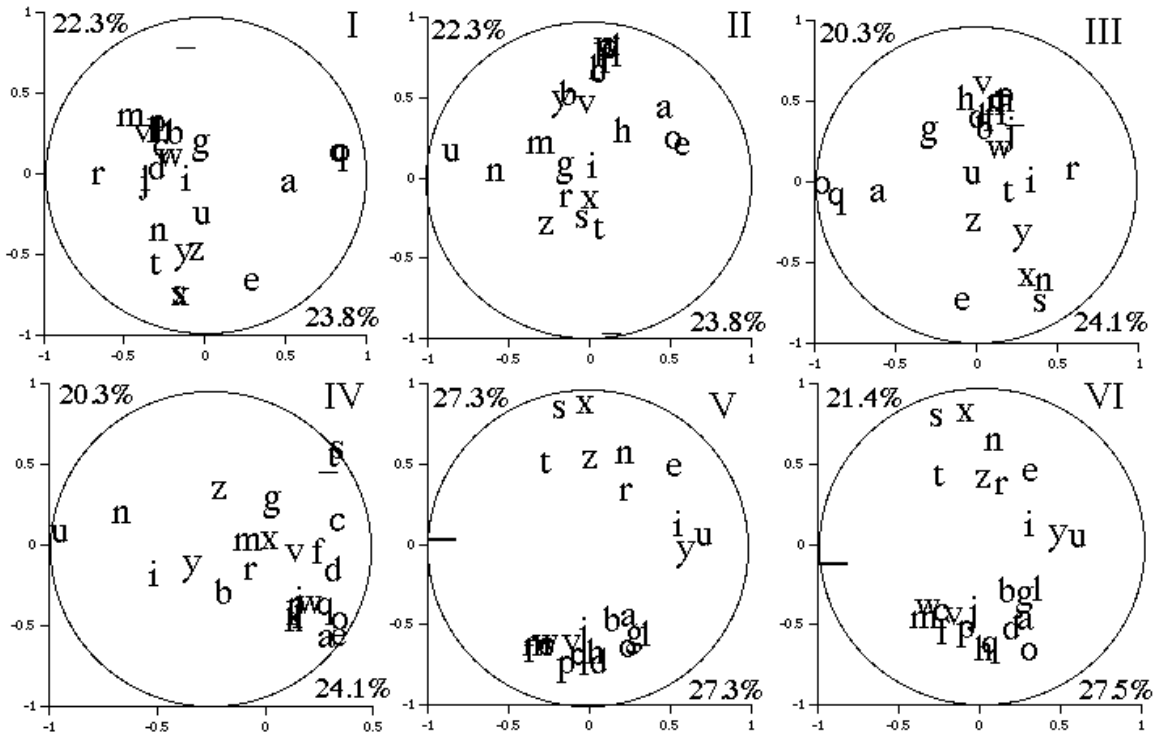

Figure 2: *circles of correlations (saturations) in the factorial plane spanned by the two first dimensions. I) row profiles (coordinates of X_j) under H_{IND} (phi square). II) column profiles (coordinates of Y_j) under H_{IND} (phi square). III) row profiles under H_{IWC} (phi square). IV) column profiles under H_{IWC} (phi square). V) row profiles under H_{SYM} (phi square). VI) row profiles under H_{SYM} (psi square).*

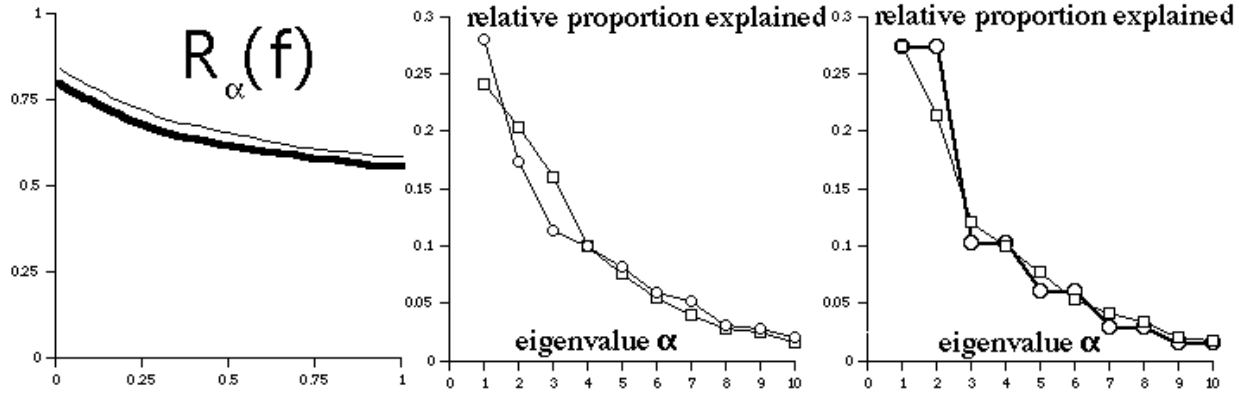

Figure 3: *left: Rényi-like index $R_\alpha(f)$ (6) (thick line; the thin line represents $R_\alpha(f)$ for the same text where all repetitions of the same letter have been suppressed). The only necessarily coinciding value with the graph of figure 1A is $I_0(f : f^{theo}) = R_1(f) = K(f||f^{theo}) = 0.56$. Were $f = f^{theo}$, then $I_s(f : f^{theo}) \equiv R_\alpha(f) \equiv 0$ for all s and all α. $R_\alpha(f)$ (thick line) decreases from $R_0(f) = 2\ln 26 - \ln 301 = 0.81$ (among the $26^2 = 676$ possible bigrams, 301 only did actually occur) to $R_1(f) = 0.56$. Middle: scree graphs for the phi square (circles) and psi square (squares) decompositions for H_{IWC}. Right: scree graphs for the phi square (circles) and psi square (squares) decompositions for H_{SYM}. Note the eigenvalue doubling phenomenon associated to the former.*

References

Bavaud, F. (1998). Models for spatial weights: a systematic look. *Geographical Analysis*, vol 30: 153-171

Bavaud, F. (2000 a). On a class of aggregation-invariant dissimilarities obeying the weak Huygens' principle. Submitted for publication.

Bavaud, F. (2000 b). The quasi-symmetric side of gravity modelling. Submitted for publication.

Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory*. Wiley, New York.

Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *J.R.Statist.Soc.B*, vol 46: 440-464

Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statist.Appl.*, vol 26: 29-37

Greenacre, M. (1984). *Theory and Applications of Correspondance Analysis*. Academic Press, London.

Lebart, L. and Salem, A. (1994). *Statistique textuelle*. Dunod, Paris.

Lebart, L. et al. (1995). *Statistique exploratoire multidimensionnelle*. Dunod, Paris.

Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT-Press, Cambridge.

Markov, A.A. (1913). An example of statistical enquiry on the text "Eugène Onéguine", illustrating tests on chains (in Russian). *Bulletin de l'Académie Impériale des Sciences de St-Pétersbourg*.

Ramuz, C.-F. (1937). *La pensée remonte les fleuves*. Mermod, Lausanne.

Saporta, G. (1990). *Probabilités, analyse des données et statistique*. Editions Technip, Paris.

Shannon, C.E. (1951). Prediction and entropy of printed English. *Bell Sys.Tech. Journal*, vol 30: 50-64

Xanthos, A. (2000). Entropizer 1.1: un outil informatique pour l'analyse séquentielle. *Proceedings of the 5th International Conference on the Statistical Analysis of Textual Data (JADT 2000).*

ERRATUM (August 2001)

Inequality (6) is referred to as the sub-additivity property by Alfred Rényi. Although verified for the data considered in this paper, inequality (6) does not hold in general (unless $\alpha = 0$ or $\alpha = 1$), as pointed out by Rényi himself (1962). That is to say, inequality $\sum_{jk} f_{jk}^{\alpha} \leq (\sum_j f_{j\bullet}^{\alpha})(\sum_k f_{\bullet k}^{\alpha})$ is not valid in general for $\alpha \in (0, 1)$; indeed, with a bit of numerical exploration, a counter-example can be found. My apologies for this.

Rényi, A. (1962). *Wahrscheinlichkeitsrechnung : mit einem Anhang über Informationstheorie*. Deutscher Verlag der Wissenschaften, Berlin.