9th International Workshop on Statistical Modelling
Exeter, 11-15 July 1994

Diabolic horseshoes

A. Baccini, H. Caussinus and A. de Falguerolles

Laboratoire de Statistique et Probabilités, URA-CNRS D0745
Université Paul Sabatier - 118, route de Narbonne
31062 TOULOUSE-CEDEX FRANCE

Abstract: Horseshoe and Guttman effects are investigated in the analysis of contingency tables by Correspondence Analysis and Row×Column association model. Attention is focused on tables obtained by discretization of continuous bivariate distributions. The emergence of these effects is explored by choosing some specific densities. In this context, neither Correspondence Analysis nor the Row×Column association model offers a general advantage in terms of parsimony.

Key words: Correspondence Analysis. Guttman effects. Horseshoe. Mehler expansion. Meixner classes. Row×Column association model.

1 Introduction

We investigate horseshoe type effects in the analysis of contingency tables by Correspondence Analysis (CA) and the Row×Column association model (RC). CA and RC produce scores which are currently used to draw plots. In some instances, a plot reveals a horseshoe or other elaborate patterns. The existence of such patterns is a frequent source of surprise to data-analysts who expect that all dimensions should be independent. Indeed, when this is present in the output of CA, it is often advocated to resort to RC since it has been frequently observed that RC successfully removes such a pattern and provides a more parsimonious description of the data. However, a variety of situations shows that the problem is more intricate. There are cases for which the converse is true, but also many cases where both analyses produce confusing patterns. Here we have chosen to focus our attention on tables obtained by discretization of continuous bivariate distributions. For some specific densities, we discuss the results of CA and RC from the point of view of the emergence of these patterns.

In section 2, following Schriever (1986) and Van Rijkevorsel (1987), we recall what is a standard horseshoe effect in CA and more generally what are Guttman effects. In section 3, we consider the case of an underlying bivariate normal distribution: in this particular case, RC is fully adapted whereas CA produces a whole series of polynomial effects (Goodman, 1991). In section 4, we investigate the case of bivariate distributions whose margins are obtained by convolution of univariate distributions in one of the Meixner classes (Eagleson, 1964; Lancaster, 1975, 1983).

Given that these bivariate distributions have also a Mehler type expansion, CA produces a series of polynomial effects as in the normal case. However, RC may lead to similar effects. In section 5, we consider a distribution for which Guttman effects appear for RC association model but not for CA, namely a 'dual' case to the normal one. Note that this distribution is studied in Lancaster (1958). In the last section, we consider some extensions of this simple distribution which are also briefly discussed in Lancaster (1958) and which lead to truly diabolic Guttman effects.

While this paper gives the mathematical framework for these diabolic patterns, the presentation will mainly illustrate them by graphical displays.

2 Horseshoes and Guttman effects

Horseshoe effects in CA have been investigated in deep by Schriever (1986) and Van Rijckevorsel (1987). Examples are observed when a seriation can be defined on rows and on columns of a contingency table (Hill, 1974). Striking examples can be found in archaeology (see e.g. Nielsen, 1991). Firstly we recall the structure of CA and RC models. Secondly we define some specific patterns which may appear in the corresponding plots.

2.1 CA and RC models

CA (formula 1) and RC (formula 2) models describe non independance between rows and columns of an $I \times J$ contingency table by specifying a bilinear model for the frequencies:

$$p_{ij} = p_{i+}\, p_{+j} \left(1 + \sum_{h=1}^{M} \lambda_h \mu_{ih} \nu_{jh}\right), \tag{1}$$

$$p_{ij} = \gamma\, \alpha_i\, \beta_j \exp\left(\sum_{h=1}^{M} \phi_h \xi_{ih} \eta_{jh}\right). \tag{2}$$

In the formulae above, M is the dimension of the model ($M \leq min(I, J) - 1$), the coefficients λ_h and ϕ_h are ordered so that $\lambda_1 \geq \ldots \geq \lambda_M > 0$ and $\phi_1 \geq \ldots \geq \phi_M > 0$, and the scores ($\mu_{ih}$ or ξ_{ih} for the rows, and ν_{jh} or η_{jh} for the columns), are subject to standard identification constraints (see e.g. Goodman, 1991).

2.2 Specific patterns

These scores are mainly used to draw plots for the levels of the two variables cross-classifying the contingency table. It may happen that these plots exhibit specific patterns. At its simplest, a horseshoe occurs if there exists a convex (or concave) function which relates the scores on dimension k to those on dimension h. Note that this may also arise in somewhat different context, for example multidimensional scaling (Mardia *et al.*, 1979). More generally, Guttman effects occur when a series of functions relate the scores associated with different dimensions to the scores of a given order (usually 1). In many cases the functions are standardized orthogonal polynomials (Tchebycheff-Hermite, Legendre ...).

3 Underlying bivariate normal

It is well known that CA of a contingency table obtained by discretization of a bivariate normal distribution gives rise to Guttman effects (Goodman, 1991). In this section, we recall how this phenomenon originates.

3.1 Discretization

Consider a continuous random vector (X, Y) having standardized bivariate normal distribution with correlation coefficient ρ. Then the probability density function for (X, Y) can be expressed as:

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right) \\ &= z(x)\ z(y)\left(1+\sum_{h=1}^{+\infty}\rho^h H_h(x)H_h(y)\right) \qquad (3) \\ &= \frac{1}{\sqrt{1-\rho^2}}\ z\left(\frac{x}{\sqrt{1-\rho^2}}\right)\ z\left(\frac{y}{\sqrt{1-\rho^2}}\right)\ \exp\left(\frac{\rho}{1-\rho^2}\ xy\right) \qquad (4) \end{aligned}$$

where z is the standardized univariate normal density function and H_h are the Tchebycheff-Hermite polynomials. Formula (3) is the well known Mehler expansion.

Any discretization of (X, Y) provides a contingency table with joint probability

$$p_{ij} = \int_{x_{i-1}}^{x_i}\int_{y_{j-1}}^{y_j} f(x,y)dxdy.$$

For sufficiently refined discretization, it appears from (3) and (4) that:

$$\begin{aligned} p_{ij} &\simeq p_{i+}\ p_{+j}\left(1+\sum_{h=1}^{M}\rho^h H_h(\mu_{i1})H_h(\nu_{j1})\right), \qquad (5) \\ p_{ij} &\simeq \gamma\ \alpha_i\ \beta_j\ \exp\left(\frac{\rho}{1-\rho^2}\ \xi_{i1}\ \eta_{j1}\right). \qquad (6) \end{aligned}$$

3.2 CA and RC analyses

It is clear from formulae (5) and (6) that CA produces artefacts (a whole series of polynomial effects including the horseshoe) whereas RC captures the true dimension of the table ($M = 1$). Theoretical aspects of this phenomenon have been treated by Lancaster (1958), Kendall & Stuart (1967), and Benzcri (1973). A numerical illustration can be found in Goodman (1991). Additionally, it appears that the fit is quite robust against the pattern of the cut-points defining the discretization: the estimated scores μ_{i1} and ξ_{i1} belong to $[x_{i-1}, x_i]$ and, correspondingly, ν_{j1} and η_{j1} belong to $[y_{j-1}, y_j]$. For a related discussion see Becker (1989). A generalization in which X and Y are p-dimensional is outlined in Benzécri (1973) and worked out in Dauxois & Pousse (1976). Practical implications are considered in Baccini *et al.* (1993). They show how the identification of patterns produced in the multidimensional case is quite intricate.

4 Some other underlying bivariate distributions

It was known beforehand that CA was not appropriate because of the Mehler expansion of the bivariate normal (see formula (3) above). Now it happens that other bivariate distributions present analogous expansions: the so-called polynomial biorthogonal property (see subsection 4.1 below). Many examples are provided by bivariate distributions whose margins are obtained by convolution of two univariate distributions in Meixner classes (subsection 4.2). Some features of their analysis by CA and RC are reported.

4.1 Polynomial biorthogonal property

Let $F(x, y)$ be a bivariate distribution function with marginals $G(x)$ and $H(x)$. It is said that F possesses the polynomial biorthogonal (PB) property (Lancaster, 1975) if and only if:

$$dF(x,y) = \left(1 + \sum_{h=1}^{+\infty} \rho_h P_h(x) Q_h(y)\right) dG(x)\ dH(y) \tag{7}$$

where $\{P_h(x)\}_{h\in\mathbb{N}}$ and $\{Q_h(y)\}_{h\in\mathbb{N}}$ are complete orthogonal systems of standardized polynomials on the respective marginal distributions ($P_h(x)$ and $Q_h(y)$ are polynomials of degree h in x and in y respectively). Note that $\rho_h = E[P_h(x)Q_h(y)]$ and that $E[P_h(x)Q_k(y)] = 0$ if $h \neq k$. Formula (7) is nothing but a generalization of the Mehler expansion. Clearly, discretization of such bivariate distributions will give contingency tables for which

$$p_{ij} \simeq p_{i+}\ p_{+j} \left(1 + \sum_{h=1}^{M} \rho_h P_h(\mu_{i1}) Q_h(\nu_{j1})\right). \tag{8}$$

Once again, CA is not appropriate. As already seen, CA generates a whole series of polynomial scores. Moreover, it might be difficult to identify these Guttman effects since the ρ_h in formulae (7) and (8) may not be ordered. For example, the scores on dimensions 1 and 2 may be related to scores on a given higher dimension.

4.2 Examples with Meixner classes margins

Meixner classes include many univariate statistical distributions among which the normal. As for the normal case, bivariate extensions can be defined. We consider those obtained by additive trivariate reduction also called variables in common method, or common components (Lancaster, 1975, 1983; Mardia *et al.*, 1979; Hutchinson & Lai, 1990). We investigate the case of contingency tables obtained by their discretization.

Meixner classes of distribution functions

They have been introduced by Meixner (1934) and can be defined from two equivalent manners: the first one from specific generating functions (Eagleson, 1964; Lancaster, 1975, 1983; Lai, 1982) and the second one from the exponential family with a quadratic variance function (Morris, 1982). There exist six classes of such distribution functions: normal, Poisson, gamma, binomial, negative binomial and generalized hyperbolic secant.

Bivariate extensions

They are obtained as follows. Let W_1, W_2 and W_3 be three independent random variables with additively compatible univariate distribution functions in the same Meixner class. Consider the so called additive trivariate reduction of W_1, W_2 and W_3: $X = W_1 + W_2$ and $Y = W_2 + W_3$. The distribution of (X, Y) is a bivariate generalization of the univariate distribution in the corresponding Meixner class.

Proposition

A bivariate distribution (X, Y) defined as above has the PB property and the corresponding correlation coefficients ρ_h are decreasing (Eagleson, 1964; Lancaster, 1975).

CA and RC analyses

Given (8) it is clear that, for these bivariate distributions (X, Y), CA produces Guttman effects as in the normal case. However, as opposed to the normal case, RC is not necessarily in dimension one. Even more, in some instances, RC gives rise to Guttman effects.

5 An underlying bivariate distribution associated with CA

In a dual vein to the normal case, we study the discretization of a bivariate distribution for which CA is fully adapted. This distribution, which is studied in Lancaster (1958), is also known as the Eyraud-Farlie-Gumbel-Morgenstern copula (Scarsini & Venetoulias, 1993). Discretizations of this joint distribution are of interest: CA fits exactly in dimension one whereas RC exhibits Guttman effects.

5.1 Definition

Let us consider the continuous bivariate distribution (X, Y) whose density function is defined by:

$$f(x, y) = \frac{1}{4}\ (1 + \alpha\ x\ y)\ \mathbb{1}_{[-1,+1]}(x)\ \mathbb{1}_{[-1,+1]}(y) \text{ with } \alpha\ \in\ [-1, +1]. \tag{9}$$

Note that this expression is a slightly altered version of the Eyraud-Farlie-Gumbel-Morgenstern copula (Hutchinson & Lai, 1990; Scarsini & Venetoulias, 1993), introduced to have uniform marginals on $[-1, +1]$. Then it is easy to verify that $E[X^2] = E[Y^2] = \frac{1}{3}$, that $E[XY] = \frac{\alpha}{9}$ and that the correlation coefficient, $\frac{\alpha}{3}$, belongs to $[-\frac{1}{3}, +\frac{1}{3}]$. Whitout loss of generality, α can be chosen to be positive.

5.2 Discretization

Any discretization of (X, Y) provides a contingency table such that

$$p_{ij} = p_{i+}\ p_{+j}\ \left(1 + \alpha\ \frac{x_{i-1} + x_i}{2}\ \frac{y_{j-1} + y_j}{2}\right)$$

where $p_{i+} = \frac{x_i - x_{i-1}}{2}$ and $p_{+j} = \frac{y_j - y_{j-1}}{2}$. Note that $\sum_{i=1}^{I} p_{i+} \ \frac{x_{i-1}+x_i}{2} = 0$ but that $\sum_{i=1}^{I} p_{i+} \ \left(\frac{x_{i-1}+x_i}{2}\right)^2$ depends on the discretization scheme.

5.3 CA and RC analyses

It is clear that CA in dimension 1 fits exactly tables built under any discretization scheme. The scores are proportional to the midpoints $\frac{x_{i-1}+x_i}{2}$ and $\frac{y_{j-1}+y_j}{2}$, with a proportionality constant which depends on the discretization scheme. As opposed, RC gives rise to Guttman effects. However, the theoretical reasons of this phenomenon are not quite clear and remain to be investigated.

5.4 Multivariate extension

Baccini *et al.* (1993) have proposed a $(2 \times M)$-dimensional normal distribution (with adapted correlation structure) as an underlying model for RC in dimension M $(M > 1)$. In a similar fashion, the density defined by formula (9) can be easily extended as

$$\frac{1}{4^M}\left(1 + \sum_{h=1}^{M} \alpha_h \ x_h \ y_h\right) \prod_{h=1}^{M} \ \mathbb{1}_{[-1,+1]}(x_h) \ \mathbb{1}_{[-1,+1]}(y_h) \text{ with } \sum_{h=1}^{M} \ | \ \alpha_h \ | \ \in \ [0,+1],$$

to underly CA in any dimension M.

6 Two examples from H.O. Lancaster

We revisit two interesting examples considered in Lancaster (1958). They question the problem of dimensionality: at least two dimensions are needed for a good fit, but only one set of scores is relevant. They may also lead to diabolic patterns: rotated horseshoe and permuted Guttman effects.

6.1 Example 1

We consider the natural orthonormal sets of functions on $[-1,+1]$ provided by the Legendre polynomials. In particular, $L_1(x) = \sqrt{3}\ x$ and $L_2(x) = \frac{\sqrt{5}}{2}(3x^2 - 1)$. We can now assign coefficients α_1 and α_2 subject to the condition that the following density becomes nowhere negative:

$$f(x,y) = \frac{1}{4}\left(1 + \alpha_1 L_1(x)L_1(y) + \alpha_2 L_2(x)L_2(y)\right) \ \mathbb{1}_{[-1,+1]}(x) \ \mathbb{1}_{[-1,+1]}(y).$$

An interesting case is when $\alpha_2 > \alpha_1$: the scores on dimension 1 are function of the relevant scores on dimension 2. As expected, empirical investigations show a rotated horseshoe both in CA and RC. An exact fit is obtained for CA in dimension 2, but RC shows, in further dimensions, Guttman effects associated with the scores on dimension 2.

6.2 Example 2

The cosine series are taken as the orthonormal sets on $[-\frac{1}{2}, +\frac{1}{2}]$. In particular, $C_1(x) = \sqrt{2}\ cos(2\pi x)$ and $C_2(x) = \sqrt{2}\ cos(4\pi x)$. We consider the bivariate distribution defined by

$$f(x,y) = (1 + \alpha_1 C_1(x)C_1(y) + \alpha_2 C_2(x)C_2(y))\ \mathbb{1}_{[-\frac{1}{2},+\frac{1}{2}]}(x)\ \mathbb{1}_{[-\frac{1}{2},+\frac{1}{2}]}(y)$$

with α_1 and α_2 subject to the condition that the density becomes nowhere negative. Empirical investigations show that CA is well adapted in dimension 2 whereas RC gives rise to Guttman effects with trigonometric patterns. According to the relative values of α_1 and α_2, permuted Guttman effects turn out.

7 Concluding remarks

We have compared CA and RC applied to contingency tables obtained by dicretization of underlying continuous bivariate distributions. It is clear that no model offers a general advantage so that both analyses should be run in parallel and compared. Moreover, the nature of the functions involved in the Guttman effects in RC remains to be properly studied.

Noting that RC and CA differ by their link function (Baccini *et al.*, 1993; Goodman, 1993; Falguerolles & Francis, 1994), the theoretical examples above illustrate the need to identify, in practical situations, the link function which leads to the most parsimonious model. To this aim, the plot of fitted values versus observed values may offer some guidance.

When tables result from the discretization of more general $(2 \times M)$-distribution, the problem becomes a lot more intricate. Preliminary investigations reveal that the discretization scheme is quite influential in the ordering of truly significant dimensions and Guttman type artefacts.

References

Baccini, A., Caussinus, H., & Falguerolles, A. de (1993), Analysing dependence in large contingency tables: Dimensionality and patterns in scatter-plots, in *Multivariate Analysis: Future Directions 2*, (C.M. Cuadras and C.R. Rao Eds.), 245-263, North-Holland, Amsterdam.

Becker, M.P. (1989), On the bivariate normal distribution and association models for ordinal categorical data, *Statistics and Probability Letters*, 8, 435-440.

Benzécri, J.P. (1973), *L'Analyse des Données, vol. 2 : l'Analyse des Correspondances*, Dunod, Paris.

Dauxois, J., & Pousse, A. (1976), *Les Analyses Factorielles en Calcul des Probabilités et en Statistique: Essai d'Étude Synthétique*, Thèse, Université Paul Sabatier, Toulouse.

Eagleson, G.K. (1964), Polynomial expansions of bivariate distributions, *The Annals of Mathematical Statistics*, 35, 1208-1215.

Falguerolles, A. de, & Francis, B. (1994), Algorithmic approach to bilinear models for two-way contingency tables, in *Proceedings of the* 4^{th} *Conference of the International Federation of Classification Societies*, to appear.

Goodman, L.A. (1991), Measures, models and graphical displays in the analysis of cross-classified data (with discussion), *Journal of the American Statistical Association*, 86, 1085-1138.

Goodman, L.A. (1993), Correspondence analysis, association analysis, and generalized nonindependence analysis of contingency tables: Saturated and unsaturated models, and appropriate graphical displays, in *Multivariate Analysis: Future Directions 2*, (C.M. Cuadras and C.R. Rao Eds.), 265-294, North-Holland, Amsterdam.

Hill, M.O. (1974), Correspondence analysis: a neglected multivariate method, *Applied Statistics*, 23, 340-354.

Hutchinson, T.P., & Lai, C.D. (1990), *Continuous Bivariate Distributions, Emphasising Applications*, Rumsby Scientific Publishing, Adelaide.

Kendall, M.G., & Stuart, A. (1967), *The Advanced Theory of Statistics*, vol. 2, Griffin, London.

Lai, C.D. (1982), Meixner classes and Meixner hypergeometric distributions, *Australian Journal of Statistics*, 24, 221-233.

Lancaster, H.O. (1958), The structure of bivariate distributions, *The Annals of Mathematical Statistics*, 29, 719-736.

Lancaster, H.O. (1975), Joint probability distributions in the Meixner classes, *Journal of the Royal Statistical Society*, Ser. B, 37, 434-443.

Lancaster, H.O. (1983), Special joint distributions of Meixner variables, *Australian Journal of Statistics*, 25, 298-309.

Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979), *Multivariate Analysis*, Academic Press, London.

Meixner, J. (1934), Orthogonale polynomsysteme mit einer besonderen gestalt der erzeugenden funktion, *Journal of the London Mathematical Society*, 9, 6-13.

Morris, C.N. (1982), Natural exponential families with quadratic variance functions, *The Annals of Statistics*, 10, 65-80.

Nielsen, K.H. (1991), The application of correspondence analysis: Some examples in archaeology, in *Classification, Data Analysis, and Knowledge Organization*, (H.H. Bock and P. Ihm Eds.), 343-351, Springer-Verlag, Berlin.

Scarsini, M., & Venetoulias, A. (1993), Bivariate distributions with nonmonotone dependence structure, *Journal of the American Statistical Association*, 88, 338-344.

Schriever, B.F. (1986), *Order Dependence*, CWI-tract 20, Amsterdam: Centre for Mathematics and Computer Science.

Van Rijckevorsel, J. (1987), *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*, DSWO Press, Leiden.