Electronic Journal of Applied Statistical Analysis
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 3, 413 – 417
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n3p413
© 2012 Università del Salento – http://siba-ese.unile.it/index.php/ejasa/index

BOOTSTRAP CONFIDENCE REGIONS IN NON-SYMMETRICAL CORRESPONDENCE ANALYSIS

Rosaria Lombardo[*(1)], Trevor Ringrose[(2)]

[(1)]*Economics Department, Second University of Naples, Italy*
[(2)]*Cranfield University, Defence Academy of the United Kingdom, UK*

Received 05 August 2012; Accepted 20 October 2012
Available online XX November 2012

Abstract: *Non-symmetric Correspondence analysis is a method increasingly used in place of classical correspondence analysis to portray the asymmetric association of two categorical variables. In this paper we investigate the reliability of graphical displays illustrating variable prediction, by looking at inferential aspects of the sampling variation of the configuration of points, using a bootstrap approach.*

Keywords: *Non-symmetric correspondence analysis, latent variables, bootstrap, elliptical confidence regions, biplot.*

1. Introduction

Non-symmetric correspondence analysis (NSCA) [3] is an exploratory technique for graphically representing the asymmetric association between two categorical variables. Conversely to correspondence analysis (CA), when there exists an asymmetric association between variables, i.e. one variable is logically antecedent to the other one, an informative analysis can be based on the partition of the asymmetric Goodman-Kruskal tau index [6]. Indeed, it is the decomposition of this measure of predictability that lies at the heart of non-symmetric correspondence analysis.
As the reliability of graphical displays illustrating the asymmetric variable association is of great interest to users, in this paper we aim to discuss inferential aspects in the sampling variation of the configuration of points, using a *complete* bootstrap approach. Our focus is on deriving confidence regions (CRs) for each data point in a low-dimensional plot or biplot via bootstrap resampling [15, 16]. The *complete* bootstrap approach for constructing CRs around row and column points is based on the difference, along latent (unobserved) variables or axes, between

[*] E-mail: rosaria.lombardo@unina2.it

sample and population points, coming from the NSCA of the original and re-sampled contingency tables.

2. Bootstrap sampling and Non-symmetrical correspondence analysis in brief

Consider an $I \times J$ two-way contingency table, where the (i, j)th relative frequency is denoted by p_{ij}. Define the i.th row relative marginal frequency by $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and the j.th column relative marginal frequency by $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$. Suppose now we treat the column variable as a predictor and the row variable as its response. For such an asymmetrically associated variable structure, non-symmetrical correspondence analysis can be used to provide a graphical summary of the row and column points. By decomposing the centered column profile (via the generalised singular value decomposition, GSVD) [9, 11] we get the singular values $(\lambda_1, \lambda_2, \ldots, \lambda_M)$ and the associated left and right latent variables or singular vectors a_{im} and b_{jm}, which are orthonormal in an unweighted and weighted metric, respectively.

In a typical asymmetric biplot display the row (response) and column (predictor) coordinates along the m.th axis are defined in terms of standard and principal coordinates (using Greenacre's terminology) $f_{im} = a_{im}$ and $g_{jm} = b_{jm}\lambda_m$, respectively. Another common plot shows both in principal coordinates, i.e. $f_{im} = a_{im}\lambda_m$ and $g_{jm} = b_{jm}\lambda_m$.

In practice the squared norm of the centered column profile is equal to the numerator of the Goodman-Kruskal tau index which, apart from multiplicative constants, follows a chi-squared distribution with (I–1)(J–1) degrees of freedom (C-statistic) [12]. For NSCA, the variation of the row and column categories can be measured such that this numerator is:

$$\tau_{num} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\right)^2 = \sum_{m=1}^{M}\lambda_m^2 = \sum_{m=1}^{M}\sum_{j=1}^{J} p_{\bullet j} g_{jm}^2 = \sum_{m=1}^{M}\sum_{i=1}^{I} f_{im}^2 .$$

In order to investigate the stability of behaviour of the response and predictor categories, we study the sampling variation and assume as a credible model the multinomial distribution. A large number (B=1000) of resampled contingency tables, constructed by bootstrapping with column margins equal to the original table, are taken. Balbi [1], following Greenacre's approach, describes a method, sometimes called a partial bootstrap, that ignores variation in the axes. The many bootstrap row and column points are projected onto the same sample axes, allowing convex hull to be drawn around them as ad hoc CRs. In contrast the approach here takes into consideration the variation in the axes, deriving elliptical CRs.

3. Elliptical Confidence Regions

Among several definitions of stability [7, 14], here we consider external stability, the degree of sensitivity of correspondence analysis to changes in the data.
Various strategies have previously been proposed that discuss a variety of different ways to obtain CRs for classical and multiple, symmetric and non-symmetric correspondence analysis. Without going into construction details, we can distinguish between parametric and non-parametric (or semi-parametric) procedures. Parametric regions are derived in a purely algebraic way, which may be circular [10] or elliptical [2, 4]. Non-parametric regions are based on asymptotic statistics, such as confidence ellipses calculated by the delta method [5; pp. 408-415], and bootstrap-based regions such as CRs by Markus [14], convex hulls by Greenacre [8; pp. 194-197] and elliptical CRs by Ringrose [16]. Apart from Ringrose, on which the present work is based, all previous regions have looked directly at the variability in only the sample points projected onto the sample axes, and so have, implicitly, compared sample and population points projected onto different sets of axes. This can be seen by considering the sample points $(\boldsymbol{x})$ that are projected onto the sample axis or latent variable $(\boldsymbol{a})$ and the population points $(\boldsymbol{\theta})$ onto the population axis or latent variable $(\boldsymbol{\alpha})$, so that they are evaluating the variation of sample and population points projected onto different axes, i.e. $\text{var}(\boldsymbol{a}'\boldsymbol{x} - \boldsymbol{\alpha}'\boldsymbol{\theta})$. In this paper we consider the variation between the sample $(\boldsymbol{x})$ and population points $(\boldsymbol{\theta})$ on the same axes, so that CRs are based on the variance in the difference between the sample point and the population point, when both are projected onto the sample axes, i.e. $\text{var}(\boldsymbol{a}'(\boldsymbol{x} - \boldsymbol{\theta}))$. This is as in Ringrose [16], which describes bootstrap CRs for classical correspondence analysis. In synthesis, we require that a 100(1-α)% CR should contain, 100(1-α)% of the time, the population point projected onto the sample axes and not onto the population axes, which represent a different set of axes to the ones we are looking at and are of scarce interest. The study of point configuration stability is particularly important in assessing the reliability of the analysis results, which in the case of NSCA means the reliability of the column categories in predicting the row categories.

4. Example

Consider the well-known two-way contingency table that cross-classifies a mother's attachment to her child, and the child's response to their mother's level of attachment, based on an extensive study of mother–child attachment, that has already been presented in literature in symmetrical [2; 16] and asymmetrical analysis [9]. The column variable is defined as *Mother Attachment Classification* and the row variable is defined as *Infant Response*. The four column categories are a result of the adult attachment interview (*DISMISSING, AUTONOMOUS, PREOCCUPIED, UNRESOLVED*), while the four row categories (*avoidant, secure, resistant*, *disorganized*) organize infant behavior. It is apparent that the association structure between *Mother Attachment Level* and *Infant Response* may be treated asymmetrically, and as a consequence the nature of the asymmetric association can be portrayed by considering NSCA. The Goodman-Kruskal tau numerator is $\tau_{\text{num}} = 0.1259$ and the *C*–statistic is 326.51, therefore there is a statistically significant (p-value < 0.0001) prediction of the row (response) variable (*infant response*) by the column (predictor) variable (*mother attachment*). In order to portray the asymmetric association

and the CRs around the row and column categories, we can consider different graphical displays like classical plot or biplot which enhance the inertia representation of the point cloud or the asymmetric relationship, respectively. Here we portray a two-dimensional, column metric preserving, biplot of the asymmetric association, in Figures 1 and 2. The origin of the plot represents the row marginal distribution, i.e. independence of rows from columns. In Figures 1 and 2, the bootstrap confidence ellipses of Ringrose are superimposed on the predictor and response categories, respectively.

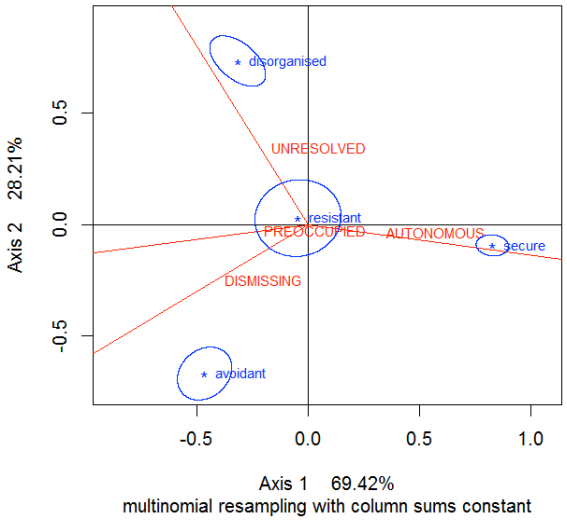

Figure 1. Biplot and CRs on rows.

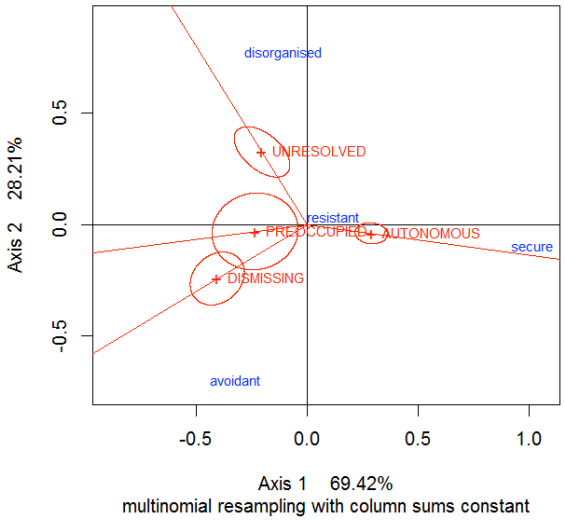

Figure 2. Biplot and CRs on columns.

The relative size of column and row regions is consistent with their relative stability. That is, the regions for the categories *Resistent* and *PREOCCUPIED* in both figures dominate. However, by taking into consideration the unequal role of these two categories, we observe that the region for *PREOCCUPIED* is very large and close to the origin which indicates that this column category does not play an important role in the prediction of infant behavior. Indeed, the strict sizes associated with elliptically generated regions, such as *AUTONOMOUS*, reflect the predictive usefulness of column categories for the accuracy of prediction of infant behavior. On the other side, looking at the CRs sizes of rows, we can argue the reliability of the prediction when the regions are smaller, it is the case of *secure* infants predicted by *AUTONOMOUS* mothers, and to a lesser extent of *disorganized* infants by *UNRESOLVED* mothers and of *avoidant* children by *DISMISSING* mothers.

5. Conclusion

Among different variants of correspondence analysis, NSCA is a suitable method of graphically exploring the asymmetric association of two categorical variables in contingency tables. The study of stability in graphical representation of NSCA via bootstrapping is of particular interest for cases when we cannot make strong distributional assumptions. We have drawn attention to assess the sampling variation and the derivation of elliptical CRs in graphical displays, in order

to make a *complete* bootstrap approach. Further research can be made to derive CRs in multiple correspondence analysis and related strategies [13].

References

[1]. Balbi, S. (1992). On Stability in Non Symmetrical Correspondence Analysis Using Bootstrap. Statistica Applicata, 4, 544–552.
[2]. Beh, E. J. (2010). Elliptical confidence regions for simple correspondence analysis. Journal of Statistical Planning and Inference, 140, 2582–2588.
[3]. D'Ambra, L. and Lauro, N. C. (1989). Non-symmetrical correspondence analysis for three-way contingency tables. In Multiway Data Analysis, eds. R. Coppi, S. Bolasco, Amsterdam: North-Holland, 301–315.
[4]. D'Ambra A., Crisci A. (2012). The confidence ellipses in decomposition Multiple Non-Symmetrical Correspondence Analysis. Communications in Statistics. In press.
[5]. Gifi, A. (1990). Nonlinear Multivariate Anaysis. Wiley, Chichester.
[6]. Goodman, L.A. and Kruskal, W.H. (1954). Measures of association for cross classifications. Journal of the American Statistical Association, 49, 732–764.
[7]. Greenacre, M. (1984). Theory and Applications of Correspondence Analysis. Academic Press, London.
[8]. Greenacre, M. (2007). Correspondence Analysis in Practice (2nd ed). Chapman & Hall/CRC, London.
[9]. Kroonenberg, P.M. and Lombardo, R. (1999). Non-symmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. Multivar. Behav. Res. Journal, 34, 367–397.
[10]. Lebart, L., Morineau, A. and Warwick, K. M. (1984). Multivariate Descriptive Statistical Analysis. New York, Wiley.
[11]. Lombardo, R., Kroonenberg, P. and D'Ambra, L. (2000). Non-symmetric correspondence analysis: a simple tool in market share distribution. J. of Italian Statist. Soc., 9, 107–126.
[12]. Lombardo, R., Beh, E. J. and D'Ambra, L., 2007. Non-symmetric correspondence analysis with ordinal variables. Computational Statistics and Data Analysis, 52, 566–577.
[13]. Lombardo, R. and Meulman, J. (2010). Multiple correspondence analysis via polynomial transformations of ordered categorical variables. Journal of Classification, 27, 191–210.
[14]. Markus, M.T. (1994). Bootstrap Confidence Regions in Non-Linear Multivariate Analysis. DSWO Press, Leiden.
[15]. Ringrose, T. J. (1996). Alternative confidence regions for canonical variate analysis. Biometrika, 83, 575–587.
[16]. Ringrose, T. J. (2011). Bootstrap confidence regions for correspondence analysis. Journal of Statistical Computation and Simulation, 82, 10, 1397–1413.

This paper is an open access article distributed under the terms and conditions of the Creative Commons Attribuzione - Non commerciale - Non opere derivate 3.0 Italia License.