Non Symmetrical Correspondence Analysis (NSCA): overview and recent developments

M. Gallo*, P. Sarnacchiaro**, L. D'Ambra*

*Dipartimento di Matematica e Statistica, Università degli Studi di Napoli

Via Cintia, Monte S. Angelo - 80126 Napoli, Italia

E-mail: dambra@unina.it. micgallo@unina.it

**Dipartimento di Progettazione Aeronautica, Università degli Studi di Napoli

P.le Tecchio, 80 - 80125 Napoli, Italia. E-mail: sarnacch@unina.it

1 Introduction

In this paper we propose an overview and recent developments of NSCA, considering both indicator matrix and two or three way contingency matrices. In the last section we discuss an approach for ordinal variables based on Partial Least Square (PLS).

2 Constraints Principal Component Analysis for qualitative variables (CPCA)

Let G, H_1 and H_2 be the binary indicator matrices related to the complete disjunctive coding of the qualitative variables G, H_1 and H_2 observed on n individuals with I, K and J categories respectively. And let the subspaces $\Re_G$, $\Re_{H_1}$ and $\Re_{H_2} \in \Re_n$ be spanned by the columns of G (criterion), H_1 and H_2 (predictors) respectively. Let $P_m^{\perp} = [I_n - P_m] = \left[I_n - \frac{1}{n}u_n u_n^{'}\right]$ be the projector operator orthogonal to the vector $u_n^{'} = [1, ..., 1]$ and I_n the identity matrix of order n. We consider the following decomposition of CPCA: $\frac{1}{n}tr\left(G^{'}G - G^{'}P_m G\right) = \frac{1}{n}tr\left(G^{'}P_{H_1}G - G^{'}P_m G\right) + \frac{1}{n}tr\left(G^{'}P_m^{\perp}P_{H_1}P_m^{\perp}G\right)$ (1) where $P_{H_1}^{\perp} = [I_n - P_{H_1}] = \left[I_n - H_1\left(H_1^{'}H_1\right)^{-1}H_1^{'}\right]$. Searching for the axes of maximal inertia we perform the eigen-analysis of $\frac{1}{n}tr\left(G^{'}P_{H_1}G - G^{'}P_m G\right)$ (2) whose trace, for the constant $C = \left[\frac{1}{n}tr\left(G^{'}G - G^{'}P_m G\right)\right]^{-1}$, is the Goodman-Kruskal's (τ) index. In case of three qualitative variables (with $\Re_{H_1}$ and $\Re_{H_2}$ disjoint) the generalization of CPCA for the indicator matrices can be found by the asymmetrical decomposition of $\frac{1}{n}G'\left[\sum_{k=1}^{2}\left(P_{H_k} - P_m\right)\right]G$ (3). In order to take into account the interaction among variables, we consider the product space $\Re_{H_{12}}$ spanned by the columns of H_{12} with K x Q categories. In this case CPCA is based on the diagonalization of $n^{-1}G'\left[P_{12} - P_m\right]G$ (4) with $P_{12} = H_{12}\left(H_{12}^{'}H_{12}\right)^{-1}H_{12}^{'}$. Its trace, unless the constant C, is coincident with the Gray-William's (τ) index. Furthermore, the analysis of the conditional effects for example H_1 lies on the diagonalization of $n^{-1}G^{'}P_{m}^{\perp}\left(P_{m}P_{12}\right)P_{m}^{\perp}G$ (5) whose trace, for the constant $\left[\frac{1}{n}tr\left(G^{'}P_{H_1}G\right)\right]^{-1}$, is the Gray-William's partial association index.

3 Non Symmetrical Correspondence Analysis for Contingency Table

The table 1 shows an overview of principal characteristics of the non symmetrical correspondence analysis for two and three way contingency tables. For each analysis the trace of matrix, for the constant term in the latter row, is the correspondent index. All property of the analysis are illustrated in the references. Here for NSTCA we take into account the possible interactions in three way contingency table. In order to show the interaction, we consider $f_{ikj} = f_{.kj}\left(f_{i..} + \sum_{\alpha}\lambda_{\alpha}^{-\frac{1}{2}}\pi_{i\alpha}\varphi_{kj\alpha}\right)$ (6) where $\pi_{i\alpha}$ and $\varphi_{kj\alpha}$ are coordinates of the I and KJ variable categories. The factor $\widehat{\varphi}_{kj\alpha}^{M}$ may be decomposed in the following way: $\widehat{\varphi}_{kj\alpha}^{M} = \widehat{\Theta}_{k\alpha} + \widehat{\Theta}_{j\alpha}$ (7). Reordering this vector in a two way matrix we perform the generalized Singular Value Decomposition we get the coordinates of variables with constraints $\sum_k f_{.k.}\widehat{\Theta}_{k\alpha} = \sum_j f_{..j}\widehat{\Theta}_{j\alpha} = 0$. Replacing these coordinates in (6), we have an estimation of f_{ikj}. We denote this quantity $\widehat{f}_{ikj}$. We can show the following decomposition $\sum_i\sum_k\sum_j f_{.kj}\left(\frac{f_{ikj}}{f_{.kj}} - f_{i..}\right)^2 = \sum_i\sum_k\sum_j f_{.kj}\left(\frac{\widehat{f}_{ikj}}{f_{.kj}} - f_{i..}\right)^2 + \sum_i\sum_k\sum_j f_{.kj}\left(\frac{f_{ikj}}{f_{.kj}} - \frac{\widehat{f}_{ikj}}{f_{.kj}}\right)^2$.

4 Multivariate Co-Inertia Analysis with Categorical and Ordinal Variables by PLS

Let $Z = [Z_1|...|Z_r|...|Z_R]$ be the normalized disjunctive complete matrices of the R predictor variables observed on the same n individuals ($Z = nH_r\left(H_r^{'}H_r\right)^{-1}$). Moreover let M_r and N be the diagonal metric with respect to the generic Z_r and G respectively. We maximize $\frac{1}{R}\sum_{r=1}^{R}cov^2\left(b, c_r\right) = \frac{1}{R}\sum_{r=1}^{R}cov^2\left(P_m^{\perp}GNu, P_m^{\perp}Z_rM_rv_r\right)$ (8) with the constraints v_r ($\|v_r\|_{M_r}^2 = 1$) and u ($\|u\|_N^2 = 1$) the coefficients vectors of the linear combinations for each Z_r and G respectively. We have $\frac{1}{R}\sum_{r=1}^{R}NG^{'}P_m^{\perp}Z_rM_rZ_r^{'}P_m^{\perp}GNu = \lambda u$ (9). To preserve the ordinal information of original data on the first axes we consider the principal column coordinate for the r^{th} table $\varphi_r = \sqrt{\frac{\mu_r}{M_r}}v_r$ (with $\mu_r = u^{'}NG^{'}P_m^{\perp}Z_rM_rZ_r^{'}P_m^{\perp}GNu$ and with $v_r = \mu_r^{-\frac{1}{2}}Z_r^{'}P_m^{\perp}GNu$) and the row standard coordinates $\psi_{(1)} = u_{(1)}$. Moreover let $\varphi_{(1)}^{'} = R^{-\frac{1}{2}}\left[\varphi_1^{'}, ..., \varphi_R^{'}\right]$ be the column principal coordinates of the first axes.

4.1 MCOICAT Algorithm to preserve the variable ordering on the first axes

Step 1 Compute the new sub-vector φ_r^{+}, by means of the theoretical values of a weighted least squares monotone regression

Step 2 Normalize the quantity $\varphi_r^+ = \frac{\varphi_r^+}{\sqrt{\varphi_r^{+'} M_r \varphi_r^+}}$ with $\varphi_{(1)}^+ = \left[\varphi_1^{+'}, \ldots, \varphi_R^{+'}\right]$, $(r = 1, \ldots, R)$

Step 3 By using the transition formula compute $\psi_{(1)}^+ = NG' P_m^\perp Z_r M_r \varphi_r^+$, $\varphi_r^+ = (\mu_r^+)^{-\frac{1}{2}} M_r Z_r' P_m^\perp GN \psi_1^+$ with $\mu_r^+ = \psi_1^+ NG' P_m^\perp Z_r M_r Z_r' P_m^\perp \psi_1^+$

Step 4 Go to step 1 until the inertia μ_r^+ does not increase and the vector $\varphi_{(r)}^+$ does not change much

Step 5 Set $\varphi_{(1)} = \varphi_{(1)}^+$, $v_{(1)} = M^{\frac{1}{2}} \varphi_{(1)}$ and $\psi_{(1)} = \psi_{(1)}^+$ so that the principal column coordinates satisfy ordinal compliance.

4.2 MCOICAT Algorithm to preserve the variable ordering on the other axis.

For the determination of the remaining co-inertia $s > 1$ we maximize the covariance between the component b^s and c_r^s by PLS algorithm, under the orthonormality constraints on the eigenvectors. Define the residual matrix Z_r^{s-1} as the orthogonal projection Z_r of onto subspaces spanned by the components $c^1, ..., c^{s-1}$, the PLS algorithm follows.

Step 1 Let $P_{c^{(1)}} = c^{(1)} \left(c^{(1)\prime} c^{(1)}\right)^{-1} c^{(1)\prime}$ be the orthogonal projector with $c^{(1)} = P_m^\perp Z M v_{(1)}$ and set $s = 2$

Step 2 Compute the residual matrix $Z^{(s)} = P_m^\perp Z^{(s-1)} M - P_{c^{(s-1)}} P_m^\perp Z^{(s-1)} M$

Step 3 Maximize the covariance criterion (9) using the residual matrix computing the first eigenvector associated to the greatest eigenvalue $\sum_{r=1}^R NG' P_m^\perp Z^{(s)} M_r Z^{(s)\prime} P_m^\perp GN$

Step 4 Compute the components scores $c^{(s)} = Z^{(s)} M v_{(s)}$ and the column component loading $\varphi^{s+1} = \sqrt{\frac{\mu_r}{\lambda_s}} v_{(s)}$ by the eigenvector $v_{(s)}$, respectively

Step 5 Increase s by one and go to step 2, repeat for $s = 3, \ldots, T$ where T is the number of interactions so that all the elements of $Z^{(s)}$ are almost zero.

A modified algorithm can be used for contingency tables (D'Ambra and ot., 2000).

5 References

D'Ambra, L. Lauro N. (1989) Non Symmetrical Analysis of Three-Way Contingency Tables - Multiway Data Analysis.

D'Ambra, L. Lombardo, R. Amenta, P. (2000) Multivariate Co-Inertia Analysis for nominal and ordinal variables by PLS - In Press.

Greenacre, M. (1984) Theory and Application of Correspondance Analysis - Academic Press.

Lombardo, R. Carlier, A. D'Ambra, L. (1997) NSCA for Three-Way Contingency Table - Methodologica, 6.

Acknowledgements: *This paper has been supported with a grant from CNR 1999, (L. D'Ambra)*

	Non Symmetrical Correspondence Analysis $\Re^I$	Non Symmetrical Multiple-way Correspondence Analysis $\Re^I$	Non Symmetrical Partial Correspondence Analysis $\Re^I$	Normalized Non Symmetrical Correspondence Analysis $\Re^{IK}$	Three-way Correspondence Analysis $\Re^I$
Aim: analysis, in the factorial context, of the dependence structure between two or more qualitative variables	G respect H_1 (with I, K categories)	G respect H_1 and H_2 (with I, KJ categories)	G respect H_1 considering the effects of H_2 (with I, K, J categories)	G and H_1 respect H_2 (with I, K, J categories)	G respect H_1 and H_2 (with I, K, J categories)
Matrix (general term)	$\mathrm{F}_{IK}(f_{ik})$ D_K diagonal matrix $(f_{.k})$ $\mathrm{F}_{I/K} = \mathrm{F}_{IK} D_K^{-1}$ $\overline{F}_{I/K} \left(\frac{f_{ik}}{f_{.k}} - f_{i.}\right)$	$\mathrm{F}_{IKJ}\left(f_{ikj}\right)$ D_{KJ} diagonal matrix $\left(f_{.kj}\right)$ $\mathrm{F}_{I/KJ} \left(\frac{f_{ikj}}{f_{.kj}}\right)$ $\overline{F}_{I/KJ} \left(\frac{f_{ikj}}{f_{.kj}} - f_{i..}\right)$	$\mathrm{F}_{IKJ}\left(f_{ikj}\right)$ D_{KJ} diagonal matrix $\left(\frac{f_{.kj}}{f_{.k.}}\right)$ $D = \begin{bmatrix} D_{K1} & 0 & 0 \\ & \ddots & \\ 0 & & 0 \\ 0 & 0 & D_{KJ} \end{bmatrix}$ $\overline{F}_{IK/j} \left(\frac{f_{ikj}}{f_{.kj}} - \frac{f_{ij.}}{f_{..j}}\right)$ $M = \begin{bmatrix} \overline{F}_{IK/1} & 0 & 0 \\ & \ddots & \\ 0 & & 0 \\ 0 & 0 & \overline{F}_{IK/J} \end{bmatrix}$	$\mathrm{F}_{IKJ}\left(f_{ikj}\right)$ D_J diagonal matrix $\left(f_{..j}\right)$ D_A diagonal matrix $\left(\sum_j f_{..j}\left(\frac{f_{ikj}}{f_{..j}} - f_{ik.}\right)^2\right)$ $\overline{F}_{IK/J}$ $\left(\sum_j \left(\frac{f_{ikj}}{f_{..j}} - f_{ik.}\right)\right)$	$\mathrm{F}_{IKJ}\left(f_{ikj}\right)$ D_J diagonal matrix $\left(f_{..j}\right)$ D_K diagonal $matrix\,(f_{.k.})$ $\overline{F}_{IK/j}$ $\left(\frac{f_{ikj}}{f_{.k.}f_{..j} - f_{i..}}\right)$
Column profile	$\frac{f_{ik}}{f_{.k}}$	$\frac{f_{ikj}}{f_{.kj}}$	$\frac{f_{ikj}}{f_{.kj}}$	$\frac{f_{ikj}}{f_{..j}}$	$\frac{f_{ikj}}{f_{.k.}f_{..j}}$
Weight	f_k	$f_{.kj}$	$f_{.kj}$	$f_{..j}$	$f_{.k.}f_{..j}$
Centered	$f_{i.}$	$f_{i..}$	$\frac{f_{.kj}}{f_{..j}}$	$f_{ik.}$	$f_{i..}$
Eigen-analysis (PCA)	$\overline{F}_{I/K} D_K \overline{F}'_{I/K}$	$\overline{F}_{I/KJ} D_{KJ} \overline{F}'_{I/KJ}$	MDM'	$D_A^{-\frac{1}{2}} \overline{F}_{IK/J} D_k \overline{F}'_{IK/J} D_A^{-\frac{1}{2}}$	Nipals Parafac/Candecom $\overline{F}_{IK/j}$ with $f_{.k.}f_{..j}$
Constant	$\left(1 - \sum_i f_{i.}^2\right)^{-1}$	$\left(1 - \sum_i f_{i..}^2\right)^{-1}$	$\left(1 - \sum_i \sum_j \frac{f_{i.j}^2}{f_{..j}}\right)^{-1}$	-	-
Trace of matrix (product with constant)	Goodman-Kruskal's τ	Multiple association Gray-William's index	Partial Gray-William's index $\tau_{IJ/K}$	Tallur's index (cluster analysis)	Marcotorchino's index

Table 1: Non symmetrical correspondence analysis for two-three way contingence tables.