Sparse Multiple Correspondence Analysis

Vincent Guillemot, Julie Le Borgne, Arnaud Gloaguen, Arthur Tenenhaus, Gilbert Saporta, Sylvie Chollet, Derek Beaton, Hervé Abdi

Sparse Multiple Correspondence Analysis

Vincent Guillemot[1,*], Julie Le Borgne[1], Arnaud Gloaguen[2], Arthur Tenenhaus[2], Gilbert Saporta[3], Sylvie Chollet[4], Derek Beaton[5], Hervé Abdi[6,*]

[1] *Bioinformatics/Biostatistics Hub, CBD, Institut Pasteur, USR 3756 CNRS, Paris, FR*
[2] *Laboratoire des Signaux et Systèmes, CentraleSupelec, Gif-Sur-Yvette, FR*
[3] *CEDRIC, Conservatoire National des Arts et Métiers, Paris, FR*
[4] *Institut Supérieur d'Agriculture de Lille, Lille, FR*
[5] *The Rotman Institute at Baycrest, Toronto, Canada*
[6] *The University of Texas at Dallas, Richardson, TX, USA*
[*] *E-mail: vincent.guillemot@pasteur.fr, herve@utdallas.edu*

Résumé. L'Analyse des Correspondances Multiples (ACM) est la méthode de choix pour l'analyse des données catégorielles multivariées. Basée sur la décomposition en valeurs singulières (SVD), l'ACM bénéficie naturellement des extensions de cette dernière, dont celles qui permettent de réaliser des analyses parcimonieuses. L'algorithme permettant de réaliser l'ACM parcimonieuse nécessite deux propriétés particulières additionnelles: l'inclusion des matrices de métriques (masses et poids) caractéristiques de l'ACM, et la possibilité de sélectionner des groupes entiers de variables (un groupe étant constitué du codage disjonctif complet d'une variable catégorielle). Nous proposons un algorithme pour l'ACM parcimonieuse basé sur la décomposition en valeurs singulières généralisée et une projection sur la boule $\ell_{1,2}$. Nous illustrons notre méthode avec les résultats d'une enquête par questionnaires sur les connaissances alimentaires et la perception de fromages.

Mots-clés. Parcimonie, Analyse Multivariée, Analyse des Correspondances Multiples

Abstract. Multiple Correspondence Analysis (MCA) is the method of choice for the multivariate analysis of categorical data. In MCA each qualitative variable is represented by a group of binary variables (with a coding scheme called "complete disjunctive coding") and each binary variable has a weight inversely proportional to its frequency. The data matrix concatenates all these binary variables, and once normalized and centered this data matrix is analyzed with a generalized singular value decomposition (GSVD) that incorporates the variable weights as constraints (or "metric"). The GSVD is, of course, based on the plain SVD and so MCA can be sparsified by extending algorithms designed to sparsify the SVD. To do so requires two additional features: to include weights and to be able to sparsify entire groups of variables at once. Another important feature of such a sparsification should be to preserve the orthogonality of the components, Here, we integrate all these constraints by using an exact projection scheme onto the intersection of subspaces (i.e., balls) where each ball represents a specific type of constraints. We illustrate our procedure with the data from a questionnaire survey on the perception of cheese in two French cities.

Keywords. Sparsity, Multivariate Analysis, Multiple Correspondence Analysis

Akin to principal component analysis (PCA), Multiple Correspondence Analysis (MCA, see for reviews [1, 6, 8]) is a multivariate analysis method that analyzes the structure of a set of I observations described by K qualitative variables each comprising J_k modalities. In MCA, the response of an observation to a qualitative variable is represented by a binary vector. The I by J data matrix to analyze (denoted $\mathbf{Y}$), called the disjunctive table, is the concatenation of K matrices each with I observations and J_k columns. In MCA, the matrix actually analyzed is the centered probability matrix denoted $\mathbf{X}$, and computed as

$$\mathbf{X} = \mathbf{Y} \times (IK)^{-1} - \mathbf{r}\mathbf{c}^\top \text{ with } \mathbf{r} = \mathbf{Y}\mathbf{1} \times (IK)^{-1} \text{ and } \mathbf{c} = \mathbf{Y}^\top\mathbf{1} \times (IK)^{-1} \ . \quad (1)$$

MCA is obtained from the generalized singular value decomposition (GSVD) of $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\top \text{ with } \mathbf{P}^\top\mathbf{M}\mathbf{P} = \mathbf{Q}^\top\mathbf{W}\mathbf{Q} = \mathbf{I} \text{ where } \mathbf{M} = \text{diag}\{\mathbf{r}\} \text{ and } \mathbf{W} = \text{diag}\{\mathbf{c}\} \quad (2)$$

with the diag operator transforming a vector into a diagonal matrix. In MCA, the interpretation of the dimensions is greatly facilitated when variables have either a very large or a very small (rather than a medium) contribution to a given dimension. In standard PCA such a pattern is obtained by rotation of the dimensions (e.g., with VARIMAX) or by sparsification (see [5]). To extend sparsification to MCA (for previous work on Sparse MCA see [2, 7]), the procedure needs to rely on extensions of sparsification that can incorporate 1) the constraints imposed by the matrices $\mathbf{M}$ and $\mathbf{W}$ and 2) a group constraint which imposes that the entire block of columns representing a variable is is selected or discarded by the sparsification procedure. In addition, as for standard sparsification of PCA, the interpretation of the results is improved when the sparsified dimensions are pairwise orthogonal, but previous sparsification methods for MCA do not implement the orthogonality constraint and sparsify only single variables. In this paper we present a generalization of the sparsified SVD (the SGSVD) that implements all the aforementioned constraints. We first present the method as an optimization problem, derive the algorithm for this optimization problem, and illustrate this new method with the analysis of a questionnaire on food preference.

1 Method

The algorithm that we propose solves the following optimization problem

$$\underset{\mathbf{P},\boldsymbol{\Delta},\mathbf{Q}}{\arg\min} \tfrac{1}{2} \left\|\mathbf{X} - \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\top\right\|_2^2 \text{ with } \begin{cases} \mathbf{P}^\top\mathbf{M}\mathbf{P} = \mathbf{I} \\ \mathbf{Q}^\top\mathbf{W}\mathbf{Q} = \mathbf{I} \end{cases}, \text{ and } \forall \ell = 1, ..., R \begin{cases} \|\mathbf{p}_\ell\|_{\mathcal{G}} \leq c_{1,\ell} \\ \|\mathbf{q}_\ell\|_{\mathcal{G}} \leq c_{2,\ell} \end{cases} \quad (3)$$

where $\|.\|_2$ is the Euclidean norm, $c_{1,\ell}$ and $c_{2,\ell}$ are positive scalars controlling the sparsity constraint for the pseudo-singular vectors, and where the group norm is defined as: $\|\mathbf{x}\|_{\mathcal{G}} = \sum_{g=1}^{G} \|\mathbf{x}_{\iota_g}\|_2$. In other words, it is the ℓ_1-norm of the vector containing the ℓ_2-norm of the sub-vectors defined by the groups. The $\ell_{1,2}$-ball associated with this norm is noted $\mathscr{B}_{1,2}(\cdot)$.

Data: $\mathbf{X}$, ε, R
Result: Sparse MCA of $\mathbf{X}$
Define $\mathbf{P} = \mathbf{0}$;
Define $\mathbf{Q} = \mathbf{0}$;
Apply weights and masses to $\mathbf{X}$;
for $\ell = 1, \ldots, R$ **do**
  $\mathbf{p}^{(0)}$ and $\mathbf{q}^{(0)}$ are randomly initialized;
  $\delta^{(0)} \leftarrow 0$;
  $\delta^{(1)} \leftarrow \mathbf{p}^{(0)\top}\mathbf{X}\mathbf{q}^{(0)}$;
  $s \leftarrow 0$;
  while $\left|\delta^{(s+1)} - \delta^{(s)}\right| \geq \varepsilon$ **do**
    $\mathbf{p}^{(s+1)} \leftarrow \text{proj}(\mathbf{X}\mathbf{q}^{(s)}, \mathscr{B}_{1,2}(c_{1,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{P}^{\perp})$;
    $\mathbf{q}^{(s+1)} \leftarrow \text{proj}(\mathbf{X}^{\top}\mathbf{p}^{(s+1)}, \mathscr{B}_{1,2}(c_{2,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{Q}^{\perp})$;
    $\delta^{(s+1)} \leftarrow \mathbf{p}^{(s+1)\top}\mathbf{X}\mathbf{q}^{(s+1)}$;
    $s \leftarrow s + 1$;
  end
  $\delta_\ell \leftarrow \delta^{(s+1)}$;
  $\mathbf{P} \leftarrow \text{vec}\left(\mathbf{P}, \mathbf{p}^{(s+1)}\right)$;
  $\mathbf{Q} \leftarrow \text{vec}\left(\mathbf{Q}, \mathbf{q}^{(s+1)}\right)$;
end
Apply inverse weights and masses to $\mathbf{P}$ and $\mathbf{Q}$;

Algorithm 1: General algorithm of the sparse MCA.

1.1 The sparse MCA algorithm

The sparse MCA algorithm is presented in Algorithm 1. It is essentially an alternate projection algorithm. Its key component is the projection onto the intersection between the ball defined by the group constraint, the ℓ_2-ball, and the space orthogonal to the already estimated singular triplets (left or right).

To achieve the projections on $\mathscr{B}_{1,2}(c_{1,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{P}^{\perp}$ and $\mathscr{B}_{1,2}(c_{2,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{Q}^{\perp}$, we perform a Projection Onto Convex Sets (POCS) [3] with two components: the projection onto the intersection of the group ball and the ℓ_2-ball, and the projection onto the orthogonal spaces defined by the already estimated pseudo-singular vectors. We detail the first projection in the next section.

1.2 Projection onto the intersection of the group and the ℓ_2-balls

Here, we present a fast and exact algorithm for the projection of $\mathbf{x}$, a fixed vector of $\mathbb{R}^n$ that comprises K non-overlapping groups, onto the intersection of an $\ell_{1,2}$-ball of radius c and the ℓ_2-ball of radius 1. This generalizes the projection onto $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$ [4].

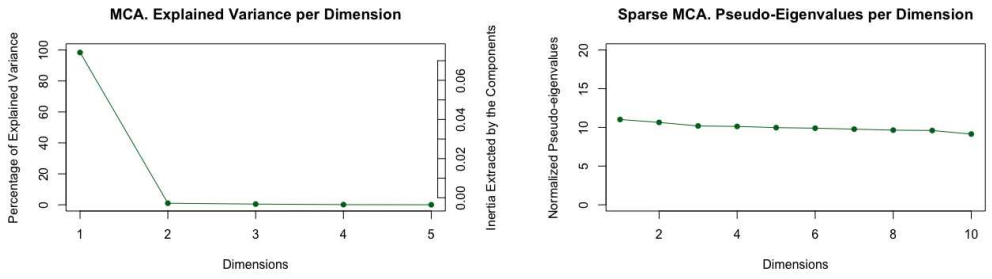
Denote by $\mathcal{G}$ the set of the indices defining the groups: $\mathcal{G} = \{\iota_k, k = 1, ..., K\}$, where ι_k indicates the variables contained in group k and K is the number of groups. Let $\mathbf{v}$ be the vector containing all the ℓ_2-norms of the sub-vectors $\mathbf{x}_{\iota_k}$, $k = 1, ..., K$. The real and positive value λ^* is such that $\| \operatorname{prox}_{\lambda^* \|\|\|_{\mathcal{G}}}(\mathbf{x})\|_{\mathcal{G}} = c$ if and only if $\| \operatorname{prox}_{\lambda^* \|\|\|_1}(\mathbf{v})\|_1 = c$, where prox_f is the proximal operator of a convex function f. Recall that the projection onto a ball is intimately linked to the proximal operator of the ball's norm. So, projecting $\mathbf{x}$ onto $\mathscr{B}_{1,2}(c) \cap \mathscr{B}_2(1)$ is equivalent to projecting $\mathbf{v}$ onto $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$, which can be achieved with the algorithm presented in [4].

2 Results

We analyzed the answers to two sets of questions of a survey on cheese answered by a sample of French participants from the two French cities of Angers and Lille. The 8 questions from the first set evaluate knowledge with answers coded as either correct or incorrect. The 23 questions from the second set evaluate the behaviors, opinions, or attitudes of the respondents toward cheese that are either farm-made or industrial. These questions are answered with a 4 point Likert scale (from 1 meaning "I totally agree" to 4 meaning "I totally disagree"). In addition we have information about the respondents: Sex, Age (coded in 4 categories), and the city where they live (Angers or Lille).

A regular MCA was applied to this data, as well as a sparse MCA with the constraints that each dimension be based on only one group of variables (i.e., a unique categorical variable) and only one group of observations (i.e., city of origin).

The Scree-plots on Figures 1a and 1b show that the first dimension of the regular MCA captures most of the variability of the data, whereas for the sparse MCA the pseudo-eigenvalues (normalized by the total inertia) are almost all equal.



(a) MCA scree-plot.

(b) Sparse MCA scree plot.

Figure 1: Scree-plots of the regular (left) and sparse (right) MCA.

The first two dimensions obtained with the regular MCA are shown in Figures 2a and 2b that confirm that city of origin is the main source of variability from this data set.

The results of the sparse MCA (see Figures 2c, 2d, 2e, and 2f) reveal that: Age is an important component on the structure of the Angers group, followed by "knowledge" and gender whereas the second city of origin (Lille) is associated with the fourth dimension.

3 Conclusion

We developed a sparse version of the MCA that incorporates into the GSVD, the group constraints imposed on the different modalities of a qualitative variable, and illustrated with real data that this new approach can simplify the interpretation of the factorial dimensions as well as reveal deeper insights. Future directions will include taking into account a hierarchical structure of either variables or observations such as overlapping groups of grouped variables as can be found, for example, for SNP data structured into pathways.

References

[1] H. Abdi and D. Valentin. Multiple Correspondence Analysis. In *Encyclopedia of Measurement and Statistics*. SAGE Publications, Inc., Thousand Oaks, 2007.

[2] A. Bernard, C. Guinot, and G. Saporta. Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis. In *Proceedings of 20th International Conference on Computational Statistics (COMPSTAT 2012)*, pages 99–106, 2012.

[3] P. Combettes. The foundations of set theoretic estimation. *Proceedings of the IEEE*, 81(2):182–208, 1993.

[4] A. Gloaguen, V. Guillemot, and A. Tenenhaus. An efficient algorithm to satisfy ℓ_1 and ℓ_2 constraints. In *49èmes Journées de statistique*, Avignon, France, 2017.

[5] V. Guillemot, D. Beaton, A. Gloaguen, T. Löfstedt, B. Levine, N. Raymond, A. Tenenhaus, and H. Abdi. A constrained singular value decomposition method that integrates sparsity and orthogonality. *PLOS ONE*, 14(3):e0211463, mar 2019.

[6] B. Le Roux and H. Rouanet. *Multiple Correspondence Analysis*. Sage, Thousand Oaks, CA, 2010.

[7] Y. Mori, M. Kuroda, and N. Makino. Sparse Multiple Correspondence Analysis. In *Nonlinear Principal Component Analysis and Its Applications*, chapter 5, pages 47–56. Springer Singapore, 2016.

[8] G. Saporta. *Probabilités, Analyse des Données et Statistique*. Technip, Paris, France, 3rd edition, 2011.

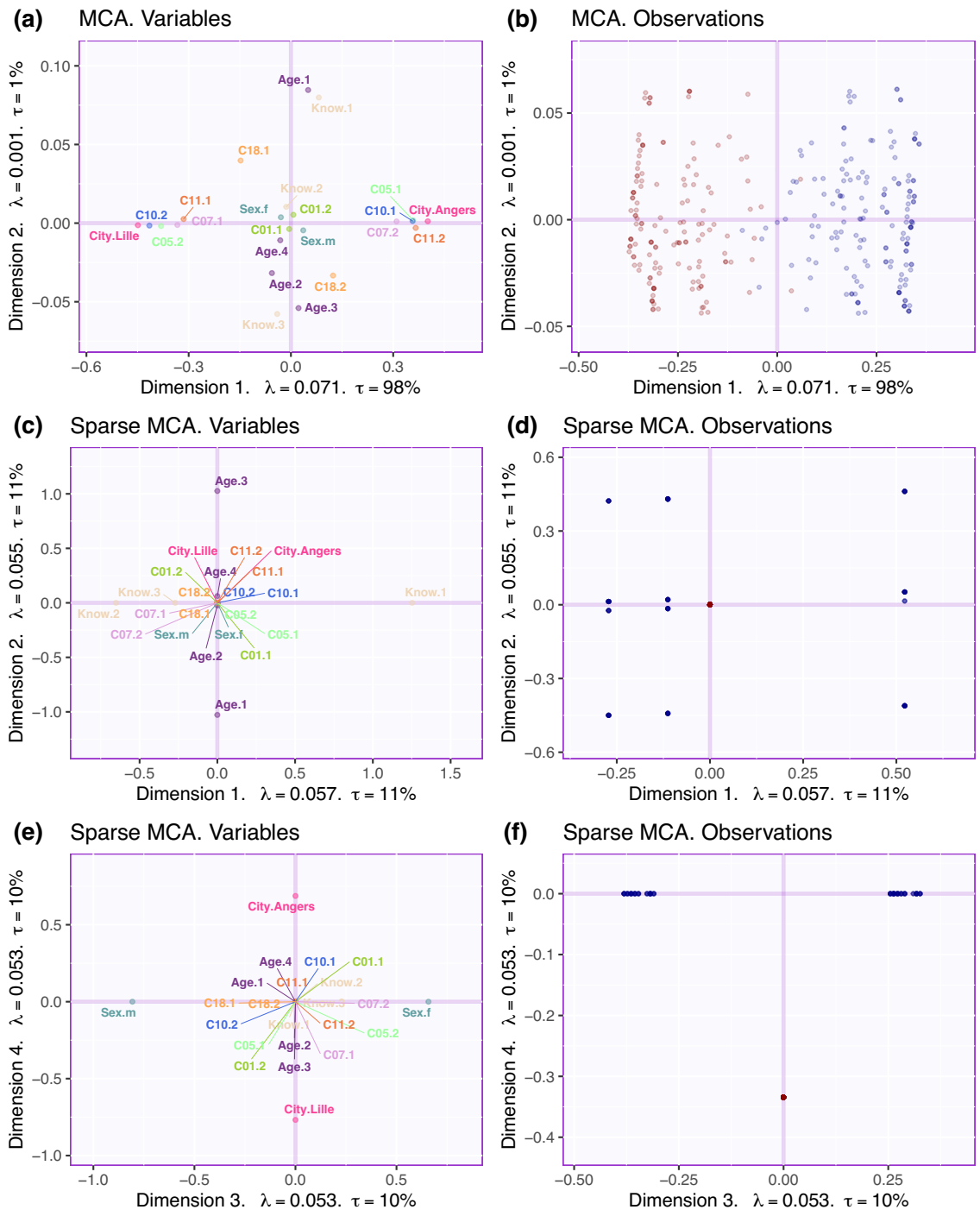


Figure 2: Variable and observation maps for MCA and sparse MCA. Only Dimensions 1 and 2 are shown for the regular MCA. For the sparse MCA, we show Dimensions 1 and 2 (middle figures) and Dimensions 3 and 4 (lower figures).