From Plain to Sparse Correspondence Analysis: a Generalized SVD Approach

Hervé Abdi, Vincent Guillemot, Ruiping Liu, Ndèye Niang, Gilbert Saporta, Ju-chi Yu

Statistica Applicata - Italian Journal of Applied Statistics Vol. 35 (3)
doi.org/10.26398/IJAS.0035-014

FROM PLAIN TO SPARSE CORRESPONDENCE ANALYSIS: A GENERALIZED SVD APPROACH[†]

Hervé Abdi
ORCID: 0000-0002-9522-1978
School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Vincent Guillemot
ORCID: 0000-0002-7421-0655
Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France,

Ruiping Liu
ORCID: 0000-0001-8591-7712
School of Applied Science, Beijing Information Science and Technology University, Beijing, China,

Ndèye Niang
ORCID: 0000-0002-6109-9935
Cedric Lab, Conservatoire national des arts et métiers, Paris, France,

Gilbert Saporta[*]
ORCID: 0000-0002-3406-5887
Cedric Lab, Conservatoire national des arts et métiers, Paris, France,

Ju-Chi Yu
ORCID: 0000-0002-6360-1861
Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Canada,

[*]***Corresponding author:*** Gilbert Saporta, `gilbert.saporta@cnam.fr`

[†] The order of the authors reflects only the alphabetical order. All authors contributed equally to this paper.

Abstract

Correspondence Analysis (CA)—the method of choice to analyze contingency tables—is widely applied in text analysis, psychometrics, chemometrics, etc. But CA becomes difficult to interpret when items load on several dimensions, when dimensions comprise items whose loadings are of intermediate values, or when the number of rows or columns is large—a configuration routinely found in contemporary statistical practice. For principal component analysis (PCA), this interpretation problem has been traditionally handled with rotation and more recently with sparsification methods often inspired by the LASSO. Curiously, despite the strong connections between CA and PCA, sparsifying correspondence analysis remains essentially unexplored.

In this paper, we extend the Penalized Matrix Decomposition (a relatively recent method based on the singular value decomposition) to sparsify CA. We present some theoretical results and properties of the resulting sparse correspondence analysis and illustrate this method with the analysis of a large textual data set.

Keywords: Sparsity, Correspondence Analysis, Generalized singular value decomposition, LASSO, Penalized matrix decomposition.

From Plain to Sparse Correspondence Analysis: A Generalized Singular Value Decomposition Approach

Abstract
Correspondence Analysis (CA)—the method of choice to analyze contingency tables—is widely applied in text analysis, psychometrics, chemometrics, etc. But CA becomes difficult to interpret when items load on several dimensions, when dimensions comprise items whose loadings are of intermediate values, or when the number of rows or columns is large—a configuration routinely found in contemporary statistical practice. For principal component analysis (PCA), this interpretation problem has been traditionally handled with rotation and more recently with sparsification methods often inspired by the LASSO. Curiously, despite the strong connections between CA and PCA, sparsifying correspondence analysis remains essentially unexplored.
In this paper, we extend the Penalized Matrix Decomposition (a relatively recent method based on the singular value decomposition) to sparsify CA. We present some theoretical results and properties of the resulting sparse correspondence analysis and illustrate this method with the analysis of a large textual data set.

Keywords: Sparsity, Correspondence Analysis, Generalized singular value decomposition, LASSO, Penalized matrix decomposition.

1. Introduction

Correspondence analysis (CA)—the method of choice to analyze contingency tables—becomes difficult to interpret when 1) the data structure is complex as opposed to the *simple structure* (formalized by early psychometricians, such as, e.g., Thurstone, 1935, 1947) where each component is characterized by few items and each item contributes only to few—ideally one—components) or 2) when the number of rows or columns is large—a configuration routinely found in contemporary statistical practice. This interpretation problem, not specific to CA, also occurs in related multivariate methods such as principal component analysis (PCA) where it has been traditionally addressed with methods such as rotation and more recently with sparsification methods mostly derived from the LASSO (Hastie et al., 2001; Tibshirani, 1996). These sparsification methods are also commonly used in fields where the data comprise large numbers of variables (Jenatton et al., 2011) or observations that can include tens of thousands (e.g., in genomics, Chun and Keleş, 2010) to millions (as in neuroimaging, see, e.g., Le Floch et al., 2012; Silver et al., 2012).

But these recent sparsification methods have not yet been widely adapted for CA and its variants. In fact, so far, mostly multiple correspondence analysis (MCA)—which can be seen as an extension of PCA for qualitative variables, as well as an extension of CA to more than two qualitative variables—has benefited from such a (precious) few of these approaches (specifically, see Bernard et al., 2012; Guillemot et al., 2020; Mori et al., 2016).

It is only recently that sparsification for CA per se has been proposed (see, Liu et al., 2023). This approach uses the fact that CA can be interpreted 1) as a double weighted PCA of both rows and columns of the data matrix, or, equivalently, 2) as a generalized singular value decomposition (GSVD, see, e.g., Abdi, 2007; Greenacre, 1984) that incorporates metric constraints on the rows and columns of the data matrix. Within this framework, sparsification is implemented by adding additional constraints on the optimization problem solved by the singular value decomposition (SVD). This constrained SVD still decomposes the data matrix into ("pseudo") singular vectors and ("pseudo") singular values, but this decomposition seeks a compromise between concurrently maximizing explained variance and sparsity. Liu et al. (2023) distinguish two cases depending on whether sparsity is required for either rows or columns, or both.

This paper replicates and extends the approach of Liu et al. (2023) in particular by proposing in lieu of their sequential algorithm, a global algorithm for the simultaneous optimization of the dimensionality of the sparsified space and the

sparsification parameters of the rows and columns of the data.

Although the theory of sparsity-inducing constraints is well documented (especially for PCA), the extension to CA is not as straightforward given its special properties. In this paper, we introduce a general formulation of sparsification which can generalize PCA to other related multivariate methods and, specifically, to CA.

We begin with the definition and main properties of CA, followed by a short exposition of the relevant approaches to sparsify PCA. We then show how to extend the concepts from sparse PCA to obtain a sparse version of CA, and describe how sparsifying CA conflicts with some of its key properties that are therefore lost in the process. Finally, we illustrate sparse CA with an analysis of an example of textual analysis extracted from Project Gutenberg (Gerlach and Font-Clos, 2020).

2. Background

2.1. Notations

Matrices are denoted in upper case bold letters, vectors are denoted in lowercase bold letters, and their elements are denoted in lowercase italic letters (note that, by default, vectors are column vectors). Matrices, vectors and elements from the same matrix all use the same letter (e.g., $\mathbf{A}$, $\mathbf{a}$, a). The transpose operation is denoted by the superscript $^\mathsf{T}$, the inverse operation is denoted by $^{-1}$. The identity matrix is denoted $\mathbf{I}$, vectors or matrices of ones are denoted $\mathbf{1}$, matrices or vectors of zeros are denoted $\mathbf{0}$ (by default, $\mathbf{I}$, $\mathbf{0}$, and $\mathbf{1}$ are conformable with the other terms in a formula). The standard product between two matrices is indicated by juxtaposition (i.e., $\mathbf{XY}$ means $\mathbf{X}$ times $\mathbf{Y}$); the Hadamard product (i.e., element-wise) is denoted by $\odot$ (e.g., $\mathbf{X} \odot \mathbf{Y}$), note that the Hadamard product is defined only between matrices with the same dimensions.

When provided with a square matrix, the diag operator gives a vector that contains the diagonal elements of this matrix. When provided with a vector, the diag operator gives a diagonal matrix with the elements of the vector as the diagonal elements of this matrix. A diagonal matrix is denoted $\mathbf{D}$, when a subscript is attached, it denotes the vector that stores the diagonal elements ot he matrix; for example, $\mathbf{D_a} = \mathsf{diag}\,(\mathbf{a})$. When provided with a square matrix, the trace operator gives the sum of the diagonal elements of this matrix. For an I by J matrix $\mathbf{X}$ and for $\mathbf{M}$ being a J by J symmetric positive definite matrix, the squared $\mathbf{M}$-norm of $\mathbf{X}$ is denoted $\|\mathbf{X}\|^2_\mathbf{M}$ and is computed as:

$$\|\mathbf{X}\|^2_\mathbf{M} = \mathsf{trace}\left(\mathbf{XMX}^\mathsf{T}\right) . \tag{1}$$

When $\mathbf{M}$ is the identity matrix, the $\mathbf{M}$-norm is equal to the square root of the sum of squares of the entries of the matrix and is called the *Frobenius* norm denoted $L_2 = \|\mathbf{X}\|_2^2$. Another useful norm is the sum of the absolute values of the matrix called the L_1 norm.

A probabilistic matrix (i.e., a matrix with non-negative elements whose sum is equal to 1) is denoted $\mathbf{Z}$, its row (respectively column) sums are stored in vector $\mathbf{r}$ (respectively $\mathbf{c}$): $\mathbf{r} = \mathbf{Z1}$ (respectively $\mathbf{c} = \mathbf{Z}^\mathsf{T}\mathbf{1}$). The matrix of row (respectively column) profiles is denoted $\mathbf{R} = \mathbf{D}_\mathbf{r}^{-1}\mathbf{Z}$ (respectively $\mathbf{C} = \mathbf{Z}\mathbf{D}_\mathbf{c}^{-1}$).

When describing an optimization problem, the operator $\arg\min_{\mathbf{x}} f(\mathbf{x})$ [respectively $\arg\max_{\mathbf{x}} f(\mathbf{x})$] gives the value of $\mathbf{x}$ that minimizes (respectively maximizes) the function $f(\mathbf{x})$.

2.2. SVD and Generalized SVD

The singular value decomposition (SVD) and its extension the generalized singular value decomposition (GSVD, see, e.g., Abdi, 2007; Allen et al., 2014; Greenacre, 1984; Holmes, 2008; Takane, 2002) are the foundations of most contemporary multivariate statistical approaches.

The SVD of an $I \times J$ matrix $\mathbf{X}$ solves the following maximization problem (Eckart and Young, 1936): Find a matrix (denoted $\widehat{\mathbf{X}}_L$) of rank L [with $L \leq \min(I,J)$], computed as

$$\widehat{\mathbf{X}}_L = \sum_{\ell=1}^{L} \delta_\ell \mathbf{u}_\ell \mathbf{v}_\ell^\mathsf{T} = \mathbf{U}_L \mathbf{\Delta}_L \mathbf{V}_L^\mathsf{T} \text{ with } \mathbf{U}_L^\mathsf{T}\mathbf{U}_L = \mathbf{V}_L^\mathsf{T}\mathbf{V}_L = \mathbf{I} \text{ and } \mathbf{\Delta}_L = \mathsf{diag}\,(\boldsymbol{\delta}_L) \quad (2)$$

such that $\widehat{\mathbf{X}}_L$ is the matrix of L rank closest to $\mathbf{X}$ (in the metric defined by the L_2 norm):

$$\underset{\mathbf{U}_L, \mathbf{\Delta}_L, \mathbf{V}_L}{\arg\min} \|\mathbf{X} - \widehat{\mathbf{X}}_L\|_2^2 \quad (3)$$

The SVD of a matrix can be computed by first computing its rank one approximation [i.e., the singular triplet (δ_1, $\mathbf{u}_1$, $\mathbf{v}_1$)] and then subtracting this rank one approximation from $\mathbf{X}$—a procedure called *deflation*. The first singular triplet of the deflated matrix $\mathbf{X}$ is then the second singular triplet of $\mathbf{X}$, etc.

The generalized SVD (GSVD), differs from the plain SVD by incorporating different orthogonality constraints on the singular vectors. Specifically, with $\mathbf{M}$ being an $I \times I$ positive definite matrix (called the row *metric* matrix) and $\mathbf{W}$ a $J \times J$ positive definite matrix (called the column metric matrix), the GSVD of $\mathbf{X}$ solves

the following optimization problem (compare with Equation 3):

$$\underset{\mathbf{P}_L,\boldsymbol{\Delta}_L,\mathbf{Q}_L}{\arg\min} \|\mathbf{X}-\widehat{\mathbf{X}}_L\|_2^2 = \underset{\mathbf{P}_L,\boldsymbol{\Delta}_L,\mathbf{Q}_L}{\arg\min} \|\mathbf{X}-\mathbf{P}_L\boldsymbol{\Delta}_L\mathbf{Q}_L^{\mathsf{T}}\|_2^2 \tag{4}$$

with

$$\mathbf{P}_L^{\mathsf{T}}\mathbf{M}\mathbf{P}_L = \mathbf{Q}_L^{\mathsf{T}}\mathbf{W}\mathbf{Q}_L = \mathbf{I}, \text{ and } \boldsymbol{\Delta}_L = \mathsf{diag}\left(\boldsymbol{\delta}_L\right), \tag{5}$$

where $\mathbf{P}_L$ is the $I \times L$ matrix containing the *generalized* left singular vectors and $\mathbf{Q}_L$ the $J \times L$ matrix containing the generalized right singular vectors. In practice, the GSVD of a matrix $\mathbf{X}$ can be obtained from the plain SVD of a matrix denoted $\widetilde{\mathbf{X}}$ obtained by pre- and post-multiplying $\mathbf{X}$ by the square root of the row and column metric matrices:

$$\widetilde{\mathbf{X}} = \mathbf{M}^{\frac{1}{2}}\mathbf{X}\mathbf{W}^{\frac{1}{2}} \ . \tag{6}$$

More details are given in Appendix A.

2.3. Basics of Plain Correspondence Analysis

Correspondence analysis was originally developed to analyze the pattern of deviations from independence (as measured by a χ^2 statistic) in a contingency table (see Abdi and Béra, 2018). CA provides, for both rows and columns, a set of factor scores whose total inertia is proportional to the independence χ^2 computed on the original contingency table. The factor scores are obtained from the following generalized singular value decomposition (cf. Equation 4) where $\mathbf{D}_\mathbf{r}^{-1}$ and $\mathbf{D}_\mathbf{c}^{-1}$ are called χ^2-metric matrices (Greenacre, 2010):

$$\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} \ \text{ with } \mathbf{P}^{\mathsf{T}}\mathbf{D}_\mathbf{r}^{-1}\mathbf{P} = \mathbf{Q}^{\mathsf{T}}\mathbf{D}_\mathbf{c}^{-1}\mathbf{Q} = \mathbf{I} \, . \tag{7}$$

Correspondence analysis can also be obtained from the plain SVD of :

$$\widetilde{\mathbf{Z}} = \mathbf{D}_\mathbf{r}^{-\frac{1}{2}}\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}\right)\mathbf{D}_\mathbf{c}^{-\frac{1}{2}} \tag{8}$$

(For further properties refer to Appendix C).

3. Sparse SVD with the Projected Penalized Matrix Decomposition

Because correspondence analysis is a particular PCA (and therefore a specific SVD, see Equations 8, above, as well as Equations 45 to 47 in Appendix B) a straightforward approach to the sparsification of CA is to adapt an already sparse version of PCA or SVD.

PCA being the oldest and most well-known multivariate method, it is no surprise that several sparse methods have been developed for PCA since the pioneering papers of Vines (2000) and Jolliffe et al. (2003). Case in point, in their—already old—review paper, Ning-min and Jing (2015) count about twenty algorithms for sparsifying PCA.

Recently, several authors have proposed sparse variants of the SVD (see, for reviews, e.g., Allen et al. 2014; Guillemot et al. 2019; Hastie et al. 2015; Jolliffe and Cadima 2016; Witten et al. 2009; Zou et al. 2006), or, specifically, of PCA (Benidis et al. 2016; Mattei et al. 2016). For most of these sparse variants, sparsification is obtained by adding sparsity constraints on both $\mathbf{P}$ and $\mathbf{Q}$, or on $\mathbf{Q}$ alone. We decided to use the *penalized matrix decomposition* method (PMD) developed by Witten et al. (2009) because it is well-known and is implemented in R (with the `PMA` package).

3.1. Penalized Matrix Decomposition: Background

The penalized matrix decomposition (PMD) method (Witten et al., 2009) generalizes the plain SVD by adding sparsification constraints on the right and left singular vectors. Specifically, the PMD methods solves the following optimization problem:

$$\underset{\substack{\delta_\ell,\mathbf{u}_\ell,\mathbf{v}_\ell \\ \ell=1,\ldots,L}}{\arg\min} \left\| \mathbf{X} - \sum_{\ell=1}^{L} \delta_\ell \mathbf{u}_\ell \mathbf{v}_\ell^\mathsf{T} \right\|_2^2 \text{ subject to } \begin{cases} \mathbf{u}_\ell^\mathsf{T}\mathbf{u}_\ell = 1 \\ \mathbf{v}_\ell^\mathsf{T}\mathbf{v}_\ell = 1 \\ \|\mathbf{u}_\ell\|_1 \leq s_{1,\ell} \\ \|\mathbf{v}_\ell\|_1 \leq s_{2,\ell} \end{cases} \tag{9}$$

where $s_{1,\ell}$ and $s_{2,\ell}$ are positive constants, provided by the user as two vectors of length L denoted $\mathbf{s}_1$ and $\mathbf{s}_2$ that will drive the sparsity of the solution. The solution to this optimization problem denoted $(\dot{\boldsymbol{\delta}}, \dot{\mathbf{U}}, \dot{\mathbf{V}})$ is called a pseudo-singular triplet (containing respectively the pseudo-singular values, left pseudo-singular vectors, and right pseudo-singular vectors).

In PMD, the first pseudo-singular triplet is estimated by solving Equation 9 for $\ell = 1$. The next pseudo-singular triplets are estimated by approximating each subsequent deflated matrix by a rank one matrix. At each iteration $\ell > 1$, the deflated matrix is equal to

$$\mathbf{X}_\ell = \mathbf{X}_{\ell-1} - \dot{\delta}_{\ell-1}\dot{\mathbf{u}}_{\ell-1}\dot{\mathbf{v}}_{\ell-1}^\mathsf{T}, \tag{10}$$

where, by convention, $\mathbf{X}_1 = \mathbf{X}$. This procedure is very similar to the standard (i.e., Hotelling's) deflation for the SVD, but in Equation 10, the deflated matrix is not

guaranteed to be orthogonal to the previous rank one optimal matrix (as noted, e.g., by Mackey, 2009).

To (partially) palliate this problem, Witten et al. (2009) and Mackey (2009) independently proposed a heuristic to handle the non-orthogonality of the row (i.e., left) factor scores (in the context of sparse PCA) where the left pseudo-singular vectors are not required to be sparse. In this case, Hotelling's deflation is replaced by the so-called *projection deflation*

$$\mathbf{X}_\ell = \left(\mathbf{I} - \dot{\mathbf{u}}_{\ell-1}\dot{\mathbf{u}}_{\ell-1}^\mathsf{T}\right)\mathbf{X}_{\ell-1}\ . \tag{11}$$

3.2. Projected Penalized Matrix Decomposition

In our case, we want to be able to obtain both sparse left and right singular vectors. To do so, we propose to extend the updating step from Equation 11 to the left and right pseudo-singular vectors. This way, for each Dimension ℓ, the projected deflated matrix is obtained as:

$$\mathbf{X}_\ell = \left(\mathbf{I} - \dot{\mathbf{u}}_{\ell-1}\dot{\mathbf{u}}_{\ell-1}^\mathsf{T}\right)\cdots\left(\mathbf{I} - \dot{\mathbf{u}}_1\dot{\mathbf{u}}_1^\mathsf{T}\right)\mathbf{X}\left(\mathbf{I} - \dot{\mathbf{v}}_1\dot{\mathbf{v}}_1^\mathsf{T}\right)\cdots\left(\mathbf{I} - \dot{\mathbf{v}}_{\ell-1}\dot{\mathbf{v}}_{\ell-1}^\mathsf{T}\right). \tag{12}$$

With this deflation scheme, PMD is applied iteratively to the data matrix after the projection deflation. Combining PMD and the projected deflation, gives the *projected Penalized Matrix Decomposition* (pPMD, see Liu et al. 2023). It should be noted, however, that pPMD does not yield perfect orthogonality but (according to Witten et al., 2009) as for projection deflation, the solutions are unlikely to be highly correlated.

4. Sparse Correspondence Analysis

In this section, we present a new way to select optimal values for the sparsity parameters, as well as choosing the optimal number of dimensions for sparse CA. Finally, we discuss the effect of introducing sparsity on the properties of CA.

4.1. Sparse CA with pPMD

Because CA is obtained from the plain SVD of $\widetilde{\mathbf{Z}}$ (see Equation 8), the current version of sparse CA is obtained by applying pPMD to $\widetilde{\mathbf{Z}}$. This procedure generates (see Equation 9) pseudo-singular values (denoted $\dot{\boldsymbol{\delta}}$), left pseudo-singular vectors (denoted $\dot{\mathbf{U}}$), and right pseudo-singular vectors (denoted $\dot{\mathbf{V}}$). These pseudo vectors and values are then used to compute sparse weight and contribution matrices for rows and columns as :

- row weight matrix $\dot{\mathbf{P}} = \mathbf{D_r}^{\frac{1}{2}}\dot{\mathbf{U}}$,
- column weight matrix $\dot{\mathbf{G}} = \mathbf{D_c}^{\frac{1}{2}}\dot{\mathbf{V}}$,

Contribution matrices are obtained from the weight matrices:

- row contributions $\dot{\mathbf{T}}_\mathbf{I} = \dot{\mathbf{U}} \odot \dot{\mathbf{U}}$,
- column contributions $\dot{\mathbf{T}}_\mathbf{J} = \dot{\mathbf{V}} \odot \dot{\mathbf{V}}$.

Note that a zero weight implies a null contribution.

In this paper, we define, (in a manner reminiscent of the transition formula from CA), the factor scores of sparse CA as linear combinations of the profiles with the (sparse) weights:

- row factor scores $\dot{\mathbf{F}} = \mathbf{R}\mathbf{D_c}^{-1}\dot{\mathbf{Q}}$,
- column factor scores $\dot{\mathbf{G}} = \mathbf{C}\mathbf{D_r}^{-1}\dot{\mathbf{P}}$,

and (just like in plain CA) for each dimension, the variance of the factor scores is equal to the squared pseudo-singular value. Note, also, that, while weights are sparse, factor scores may not be sparse.

Even though the computation of the factor scores for sparse CA bears some resemblance to the transition formula (and would be equivalent to the transition formula in plain CA), the transition formulas (from Equation 43 in the Appendix) no longer hold: row (respectively column) factor scores are not barycenters anymore of the column (respectively row) factor scores: Transition formulas hold only for plain CA.

4.2. Two Types of Sparsity

Using a sparse version of CA is especially useful when the data are high-dimensional. However, large data sets come in two types: 1) both row and column sets are high-dimensional and sparsifying both dimensions makes sense, or 2) only one of the row or column sets is high-dimensional—and the data table is a "flat" or a "tall" contingency table—and then, it makes more sense to sparsify only the larger set of items. Therefore, two types of sparsity need to be considered for the sparse solution of CA: 1) both-side (or double) sparse CA or 2) one-side (or simple) sparse CA.

Both-side sparsity looks for underlying dimensions that are explained by sparse combinations of both rows and columns. Sparse SVD provides this solution. The easiest way to implement both-side sparsity is to sparsify rows and

columns weights in the same proportion, an approach which leads to choose similar degrees of sparsity for $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$. Following Witten et al. (2009), for $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$ to have a similar level of sparsity, η is set constant (with $\eta < 1$) and the sparsity parameters are obtained as $s_{1,\ell} = \eta\sqrt{I}$, and $s_{2,\ell} = \eta\sqrt{J}$ (with I and J being the number of rows and columns of the data matrix). But, if the rows and columns contingency table correspond to essentially different types of variables, then it makes more sense to choose different degrees of sparsity for rows and columns. In this case, the parameter settings "grid" can be used, in order to restrict the L_1 norms of $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$ at different values: $s_{1,\ell}$ and $s_{2,\ell}$ will be restricted to take values (respectively) in the intervals $[1, \sqrt{I}]$ and $[1, \sqrt{J}]$.

One-side sparsity is suitable in asymmetrical situations when, for example, only the rows (or the columns) of the contingency table are relevant. In this case, only the relevant set needs to be sparsified. Interestingly, one-way sparsity is a special case of both-side sparsity when there is no penalty on one side (which is then left un-sparsified) and setting the sparsity parameter of the side to be sparsified equal to the square root of the cardinal of this set (e.g., to sparsify only the rows, set $s_{1,\ell} = \sqrt{I}$).

4.3. Choosing an Appropriate Value for the Sparsity Parameters: the Sparsity Index

An essential decision when using sparse CA is the choice of the values for the sparsity parameters $\mathbf{s}_1$ and $\mathbf{s}_2$ and the number of dimensions L. Various methods have been proposed: cross-validation (Witten et al., 2009), AIC or BIC (Shen et al. 2013; Zou et al. 2006), and compromise between the goodness of fit and sparsity (see, e.g., Trendafilov 2014; Trendafilov et al. 2017).

Among these procedures, we chose the sparsity index presented by Trendafilov et al. (2017). This sparsity index denoted, here, $\varsigma(\mathbf{s}_1, \mathbf{s}_2, L)$, is the product of a "fit ratio" and a "zero ratio."

The *fit ratio* is computed as the ratio of the sum of the pseudo-eigenvalues to the sum of the eigenvalues of the non-sparse solution, specifically

$$\text{fit ratio} = \frac{\sum_{\ell=1}^{L} \dot{\lambda}_\ell}{\sum_{\ell=1}^{L} \lambda_\ell} \,. \tag{13}$$

The fit ratio takes values between 0 and 1 with larger values indicating a better fit. The *zero ratio* is the ratio of the number of zero weights to the total number of

weights, specifically:

$$\text{zero ratio} = \frac{\#0(\dot{\mathbf{P}}) + \#0(\dot{\mathbf{Q}})}{(I+J)L}, \tag{14}$$

where $\#0(\dot{\mathbf{P}})$ [resp. $\#0(\dot{\mathbf{Q}})$] is the total number of zeros in $\dot{\mathbf{P}}$ (resp. $\dot{\mathbf{Q}}$). The zero ratio takes values between 0 and 1 with larger values indicating a sparser solution. The sparsity index $\varsigma(\mathbf{s}_1, \mathbf{s}_2, L)$ is obtained as the product of the fit and the zero ratios, namely:

$$\varsigma(\mathbf{s}_1, \mathbf{s}_2, L) = \underbrace{\frac{\sum_{\ell=1}^{L} \dot{\lambda}_\ell}{\sum_{\ell=1}^{L} \lambda_\ell}}_{\text{"fit ratio"}} \underbrace{\frac{\#0(\dot{\mathbf{P}}) + \#0(\dot{\mathbf{Q}})}{(I+J)L}}_{\text{"zero ratio"}}, \tag{15}$$

To sum up, the sparsity index is a compromise between maximizing the explained variance (i.e., the fit ratio) and sparsifying the results (i.e., the zero ratio). In our application of sparse CA, we will therefore seek for the value(s) of L, $\mathbf{s}_1$, and $\mathbf{s}_2$ that maximize $\varsigma(\mathbf{s}_1, \mathbf{s}_2, L)$.

Our global optimization algorithm differs from the sequential algorithm of Liu et al. (2023) which searches for the optimal sparsity level for each dimension conditional on the sparsity levels obtained for the previous dimensions, but without searching for an optimal value of L (i.e., the dimensionality of the space). In contrast, we obtain a global optimum in a space of Dimension L with the additional constraints that all dimensions have identical levels of sparsity : $s_{11} = s_{12} = \ldots = s_{1L}$.

4.4. Lost Properties and other Issues

In addition to the usual (but still open) question of "How many components to keep?" sparse exploratory methods raise new specific issues such as—among others—loss of orthogonality and choice of the sparsity level.

The simultaneous orthogonality of the weight vectors and of the factor scores characterizes PCA (and SVD) because weight vectors and factor scores are both true eigenvectors. But this simultaneous orthogonality is lost in sparse PCA and similar methods: One cannot have both orthogonality for the weights and for the factor scores. For example, if we force successive sparse weight vectors to be orthogonal, as in SCoTLaSS (Jolliffe et al., 2003), the associated factor scores are no longer orthogonal.

This lack of orthogonality makes the interpretation of the factor scores somewhat difficult (in a way reminiscent of the issues linked to oblique rotation in traditional factor analysis) because conclusions about one dimension involve all correlated dimensions and because the same information is explained (to different degrees) by all correlated dimensions. When interpreting the factor scores, one could erroneously find the same information in different dimensions. In addition, with non-orthogonal factor scores, the variances explained by different dimensions are no longer additive (i.e., the sum of the variances explained by a set of non-orthogonal dimensions will over-estimatethe variance of the sub-space spanned by these dimensions).

As we have noticed before, the simultaneous pseudo-barycentric transition formulas (from Equation 43 in Appendix B) do not hold anymore because these formulas are a characteristic property of plain CA. Here $\dot{\mathbf{F}}$ is not proportional to $\dot{\mathbf{P}}$ and $\dot{\mathbf{G}}$ is not proportional to $\dot{\mathbf{Q}}$: In other words, the relationship between the weights and the factor scores is not linear anymore. As a consequence, graphics should be drawn using the factor scores $\dot{\mathbf{F}}$ and $\dot{\mathbf{G}}$ rather than the weights $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$, because a graphic drawn from the weights is likely to have too many points stuck to the axes (these will be the items with zero weights). However, a graph drawn from the weights or even from the signed contributions could be of interest in some applications.

5. A Real Data Example

We applied sparse CA to a data set—obtained from the Project Gutenberg (Gerlach and Font-Clos, 2020)—compiling common words used in 100 books each from 5 book categories: Biographies, Love stories, Mystery, Philosophy, and Science Fiction. This created a contingency table (counting the number of occurrences of words per book) with 1502 rows (words) and 500 columns (books).

5.1. Plain CA Results

Factor scores maps for plain CA for Dimensions 1 and 2 are shown in Figures 1 (for the words) and 2 (for the books). The word factor map shows only a few words, whereas the book factor map does not show the names of the books but color them by genre and add, for each type of book, a 70% tolerance convex hulls (a *K% tolerance* interval comprises *K%* of a sample or a population, see, e.g., Abdi et al. 2009)[1]. Parallel to the partition of the vocabulary, Dimension 1

[1]The graph and convex hulls were created using functions `ggConvexHull` and `CreateFactorMaps4CA` from the `R-package PTCA4CATA`.

(see Figure 2) differentiates the Philosophical genre from Love stories, Mystery, and Science fiction.

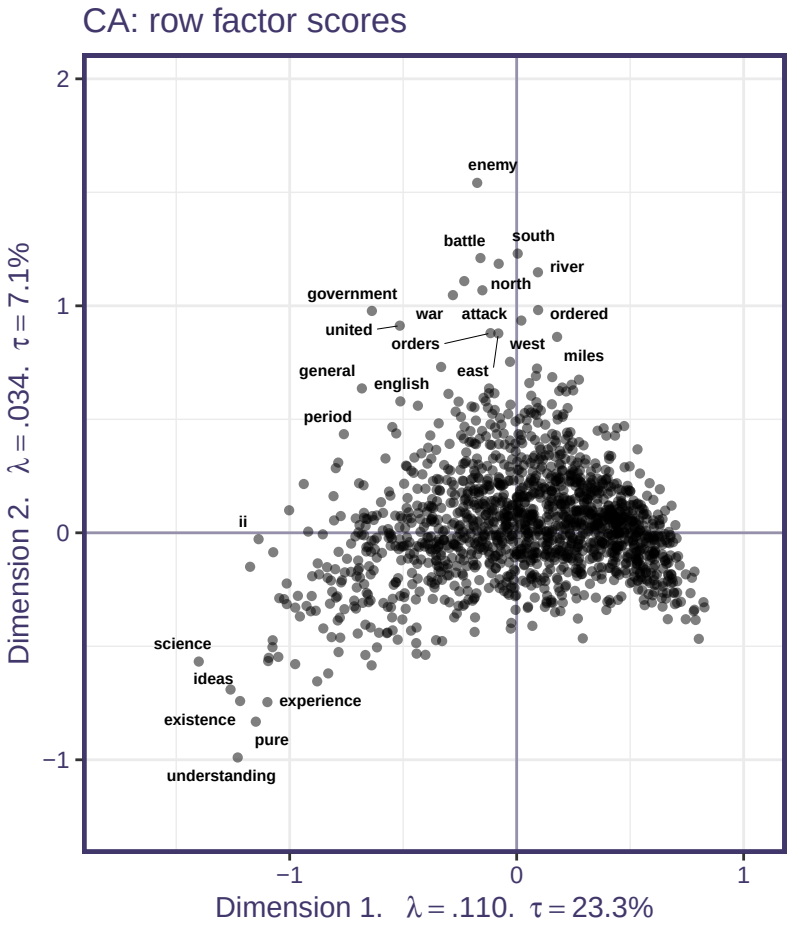

Figure 1: Plain CA: Row factor scores

The second dimension of plain CA explains 7% of the inertia. As shown in Figure 1 and Table 1b), the row factor scores are characterized by the opposition of (on the negative side) words related to war (enemy, battle, war, government) or geography (e.g., south, north, west, east, miles, city), and verbs in the past tense (e.g., ordered, united, received, sent, arrived) versus on the negative side, words related to thoughts (e.g., understanding, experience, reason, meaning, conscious) and verbs in the present tense (e.g., does, miss, mean, thinks, makes, is, can).

Parallel to the partition of the vocabulary, Dimension 2 (see vertical axis Figure 2) differentiates the Biography genre from books of Philosophy.

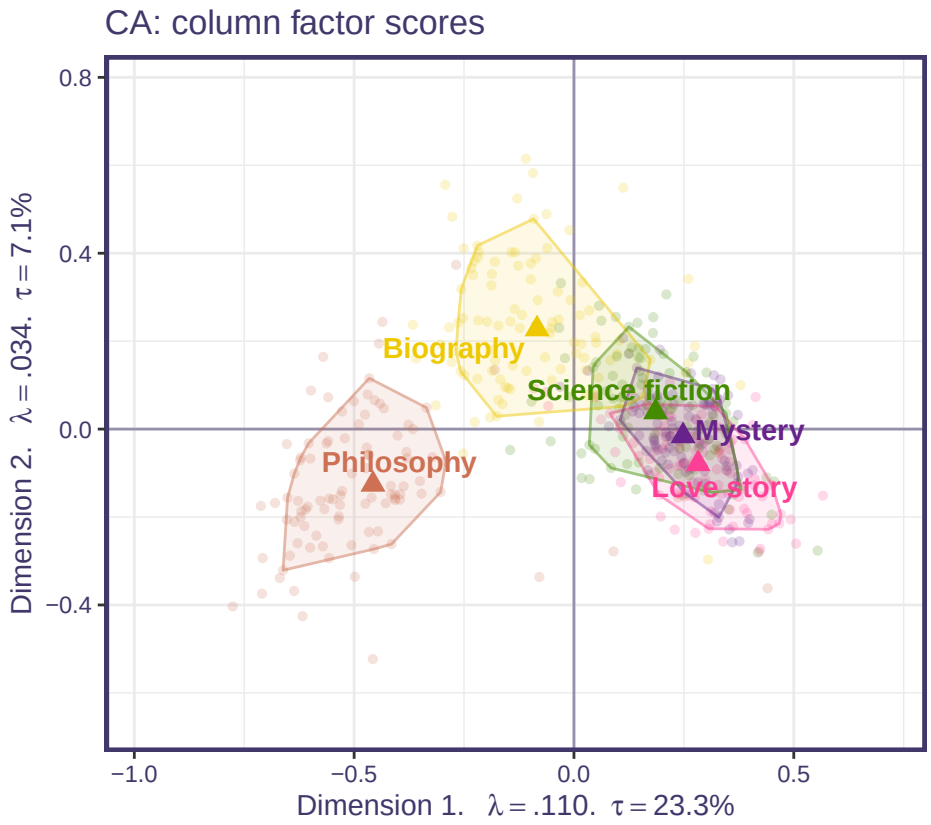

Figure 2: Plain CA: Column factor scores

In general, the Philosophy genre uses language closely aligned with scientific writing. In contrast, the Biography genre typically uses language focused on the events of a male character's past, often involving war. Meanwhile, the other three genres of fiction are more concerned with describing emotions and the experiences of female characters.

The Philosophy genre is particularly distinctive in its use of language, because it preferably uses highly specialized terminology not commonly found in

other genres. However, interpreting factor scores for this genre is challenging given the presence of many words with close-to-zero weights that are difficult to integrate in a coherent framework.

5.2. Choosing an Optimal Value of the Constraints Parameters with the Sparsity Index

Recall that the sparsity index $\varsigma(\mathbf{s}_1, \mathbf{s}_2, L)$ is a function of $\mathbf{s}_1$, $\mathbf{s}_2$, and L. To find the sparsity index optimal value, we explored the (3-dimensional) space spanned by these parameters. We chose L (i.e., the number of dimensions) from all integers between 2 and 20, and we chose $\mathbf{s}_1$ (resp $\mathbf{s}_2$) from the 20 possible values evenly distributed between 1 and $\sqrt{I}$ (resp. $\sqrt{J}$). To speed up computations and provide the same sparsity value for each dimension, we decided to have identical values for $\mathbf{s}_1$ and $\mathbf{s}_2$. We then iteratively applied sparse CA to the Gutenberg Project data set with a number of dimensions equal to L (with L varying from 1 to 20). Figure 3 shows the scatterplot of all combinations of the parameters with zero ratio (see Equation 15) on the horizontal axis and the fit ratio on the vertical axis. The points in the figure are colored according to the number of dimensions of these solutions. The isolines correspond to a fixed value of the sum of squared fit and zero ratios, for example, the thick isoline going from a fit ratio of 1 to the zero ratio of 1 is the locus of the sum of these two squared ratios equal to 1. The closest solutions to the upper right corner (which matches a fit and a zero ratio of 1) is the optimal one with the largest sparsity index. In Figure 3, this solution is indicated by the arrow and the value of its sparsity index. In this analysis, the optimal sparsity index is equal to .47 and occurs for an $L = 2$ factor solution with a fit ratio equal to .72, a zero ratio equals .66, and sparsity parameters for rows being $\mathbf{s}_1 \approx (10.94, 10.94)$ and for columns being $\mathbf{s}_2 \approx (13.37, 13.37)$.

Figure 4 shows the values taken by the sparsity parameter on a map where the horizontal axis corresponds to the 20 values chosen between $\frac{1}{I}$ and $\mathbf{s}_1$ and the vertical axis corresponds to the 20 values chosen between $\frac{1}{J}$ and $\mathbf{s}_2$. The values of $\mathbf{s}_1$ (resp. $\mathbf{s}_2$) are scaled by I (resp. J) so that the range of these possible sparsity parameters becomes between 0 and 1. In Figure 4, the optimal solution, which has the largest sparsity index is identified by the star.

5.3. Sparse CA Results

The results from sparse CA are shown in Figures 5 to 8. With sparsification, the words that have small contributions in plain CA now have zero contribution and the words with large contributions now have even larger contributions—a

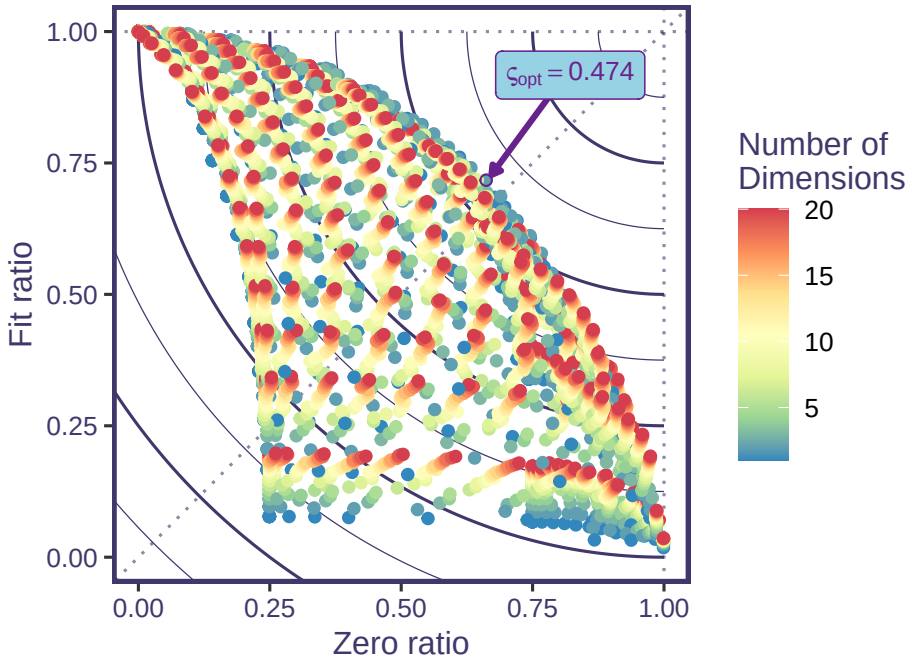

Figure 3: Fit ratio to zero ratio graph.

pattern shown in Figure 5 which plots, for the rows, the plain CA contributions (ordered from left to right by their factor scores) versus on the bottom the sparse CA.

A similar pattern, but with with a smaller effect of sparsity, is found for the contributions of the books (see Figure 6). The factor scores are shown in Figures 7 and 8 with words and books that have zero contributions on both dimensions indicated by hollow dots. The first dimension of sparse CA explains 17% of the inertia. Similar to plain CA, the first dimension differentiates neutral pronouns (e.g., itself, their, us, human, this) and words (e.g., understanding, pure, science, ideas) from

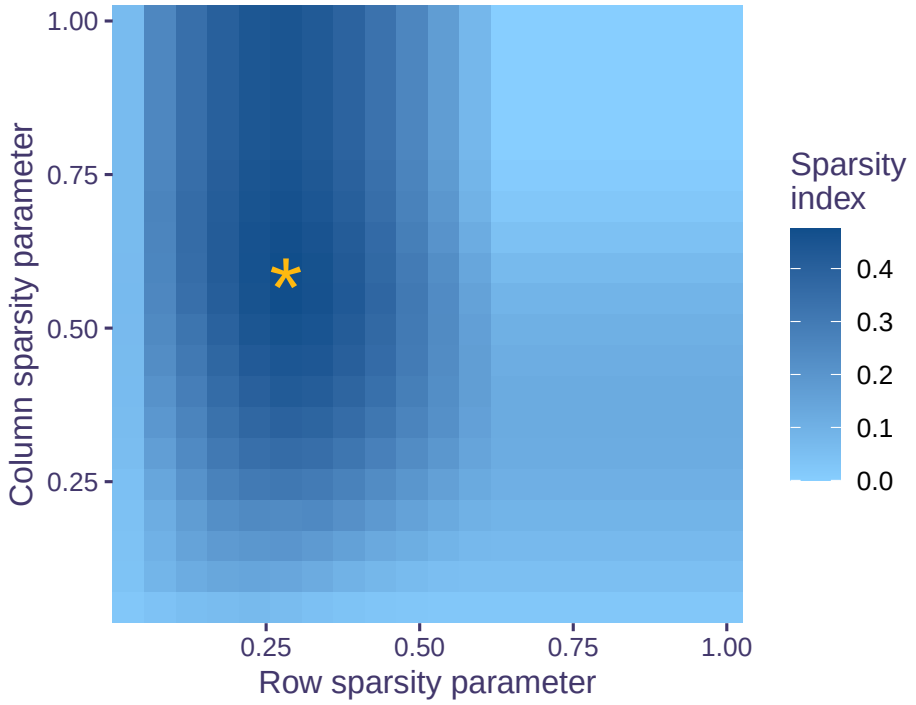

Figure 4: Sparsity Index map.

words of feminine figures (e.g., girl, she, herself, lady) and words describing emotions (e.g., miss, smiled, laughed, sorry; see Table 2a and the horizontal axis in Figure 7).

Patterns similar to plain CA are also found for the column factor scores of sparse CA, which differentiates the Philosophy genre from Love stories, Mystery, and Science fiction (see the horizontal axis in Figure 8). Books with a small contribution in plain CA have a zero contribution with sparse CA and books with a large contributions in CA now have even larger contributions in sparse CA—a pattern shown in Figure 6 which plots, for the books, the plain CA contributions

(ordered from left to right by their factor scores) versus, on the bottom, the sparse CA.

The second dimension of sparse CA explains 5% of the inertia. The row factor scores are characterized by the opposition (on the positive side the dimension) of words related to war (e.g., enemy, battle, attack, war), geography (e.g., south, north, east, west, miles, city), and verbs in the past tense (e.g., ordered, united, received) to (on the negative side the dimension) words related to thoughts (e.g., understanding, experience, reason, meaning, think) and feminine figures (e.g., she, girl, herself, lady, her), and verbs in the present tense (e.g., miss, is, does, say, be, can, am) (see the vertical axis in Figure 7 and Table 2b).

The column factor scores differentiate books from the Biography genre from books of Philosophy, Love stories, and Mystery (see the vertical axis in Figure 8 and bottom panel of Figure 6).

Compared to the results from plain CA, because sparse CA shrunk some words while emphasizing others, the pattern opposing neutral versus feminine pronouns along Dimension 1 becomes more noticeable. Compared to the results from plain CA, because the transition formula is no longer valid in sparse CA, the sparsity of the contributions (derived from loadings; see Figures 5 and 6) does not propagate to give sparse factor scores. But, as demonstrated, the sparsity of contributions can be integrated to facilitate the interpretation of factor scores. Moreover, although the results from sparse CA do not have the optimal proportion of explained inertia, sparse CA gives the solution with the optimal trade-off between the inertia explained and sparsity. Finally, it is worth noting that because the sparse CA factor scores of the two dimensions are not orthogonal, their percentages of explained inertia are not additive and need to be considered separately. However, here although the components from sparse CA are not orthogonal, the two selected dimensions are close-to-orthogonal with a correlation coefficient of .02.

words	loadings	contributions*	factor scores
science	−4.22	60	−1.40
ideas	−3.80	38	−1.26
understanding	−3.71	21	−1.23
existence	−3.68	39	−1.22
system	−3.54	28	−1.17
pure	−3.47	20	−1.15
ii	−3.43	27	−1.14
experience	−3.31	36	−1.10
physical	−3.30	16	−1.09
material	−3.30	15	−1.09
nature	−3.25	79	−1.08
knowledge	−3.24	49	−1.07
iii	−3.24	15	−1.07
itself	−3.17	64	−1.05
according	−3.16	22	−1.05
example	−3.09	14	−1.02
parts	−3.06	17	−1.01
series	−3.05	10	−1.01
progress	−3.02	15	−1.00
object	−2.94	26	−0.98
stepped	2.11	3	0.70
sat	2.12	14	0.70
box	2.13	5	0.70
sorry	2.13	6	0.71
smile	2.13	9	0.71
cried	2.13	16	0.71
door	2.15	27	0.71
window	2.17	10	0.72
yes	2.18	30	0.72
guess	2.25	5	0.74
nice	2.25	4	0.74
laughed	2.31	9	0.77
shook	2.33	8	0.77
she	2.35	380	0.78
ca	2.36	14	0.78
whispered	2.36	6	0.78
oh	2.38	31	0.79
miss	2.42	36	0.80
girl	2.48	31	0.82
smiled	2.50	9	0.83

(a) Dimension 1

words	loadings	contributions*	factor scores
understanding	−5.40	4	−0.99
pure	−4.54	3	−0.83
experience	−4.07	5	−0.75
existence	−4.05	4	−0.74
ideas	−3.77	3	−0.69
space	−3.57	2	−0.65
reason	−3.38	6	−0.62
sense	−3.19	4	−0.58
object	−3.16	2	−0.58
science	−3.10	3	−0.57
physical	−3.08	1	−0.56
material	−3.01	1	−0.55
itself	−2.99	5	−0.55
meaning	−2.94	1	−0.54
conscious	−2.94	0	−0.54
does	−2.91	5	−0.53
merely	−2.87	2	−0.53
soul	−2.76	2	−0.50
nature	−2.75	5	−0.50
absolutely	−2.65	0	−0.49
post	3.74	0	0.69
hundred	3.77	4	0.69
report	3.95	1	0.72
english	3.99	4	0.73
city	4.11	4	0.75
miles	4.71	4	0.86
east	4.79	2	0.88
orders	4.80	2	0.88
united	4.98	2	0.91
west	5.11	2	0.94
government	5.34	5	0.98
ordered	5.36	2	0.98
war	5.72	9	1.05
attack	5.83	3	1.07
command	6.05	4	1.11
river	6.27	8	1.15
north	6.47	5	1.19
battle	6.61	5	1.21
south	6.72	5	1.23
enemy	8.42	13	1.54

(b) Dimension 2

Table 1: The 20 most extreme words from each dimension of Plain CA.

Note: The contributions shown as 0 in 1b were too small to be displayed as integers; it is worth noting that this value does not indicate zero contributions. * indicates that the original values were multiplied by 10,000 and rounded to the nearest integer for display purposes.

words	loadings*	contributions*	factor scores
understanding	−0.67	2	−1.45
science	−1.75	9	−1.44
ideas	−1.16	5	−1.32
pure	−0.63	2	−1.30
existence	−1.24	5	−1.29
experience	−1.29	5	−1.19
system	−0.79	2	−1.18
ii	−0.81	2	−1.16
physical	−0.42	1	−1.13
knowledge	−1.82	7	−1.12
material	−0.36	0	−1.11
series	−0.25	0	−1.11
itself	−2.58	10	−1.11
nature	−3.09	12	−1.10
iii	−0.37	0	−1.09
object	−0.99	3	−1.07
space	−0.59	1	−1.06
according	−0.65	1	−1.05
parts	−0.49	1	−1.04
example	−0.33	0	−1.02
sat	0.44	0	0.64
lips	0.18	0	0.64
dear	0.71	1	0.65
smiling	0.00	0	0.65
sorry	0.04	0	0.65
yes	1.17	2	0.65
shook	0.07	0	0.65
ca	0.30	0	0.65
whispered	0.02	0	0.66
smile	0.20	0	0.66
lady	1.01	1	0.67
nice	0.00	0	0.68
laughed	0.15	0	0.68
her	22.60	67	0.70
herself	0.86	1	0.70
smiled	0.13	0	0.73
oh	1.17	2	0.73
girl	1.09	2	0.75
she	23.05	77	0.78
miss	1.65	4	0.82

(a) Dimension 1

words	loadings*	contributions*	factor scores
understanding	−0.95	5	−0.93
pure	−0.78	3	−0.76
experience	−1.58	7	−0.67
space	−0.75	2	−0.64
existence	−1.28	5	−0.63
ideas	−1.01	3	−0.57
reason	−2.28	9	−0.57
object	−0.97	3	−0.53
sense	−1.31	4	−0.52
miss	−1.46	3	−0.51
does	−2.27	7	−0.49
mean	−0.93	2	−0.48
meaning	−0.17	0	−0.48
conscious	−0.05	0	−0.46
physical	−0.23	0	−0.46
merely	−0.57	1	−0.45
material	−0.17	0	−0.45
things	−2.88	8	−0.44
nice	−0.00	0	−0.44
absolutely	−0.05	0	−0.43
advance	0.17	0	0.67
report	0.16	0	0.72
city	0.97	3	0.73
general	3.26	16	0.76
english	1.48	7	0.77
miles	0.92	4	0.80
east	0.36	1	0.88
orders	0.57	2	0.92
west	0.54	2	0.94
united	0.56	2	0.95
ordered	0.51	2	1.01
government	1.43	9	1.05
war	2.34	18	1.12
attack	0.69	4	1.13
river	1.65	13	1.17
command	1.05	8	1.22
north	1.04	8	1.23
south	1.11	9	1.29
battle	0.93	7	1.29
enemy	2.32	28	1.67

(b) Dimension 2

Table 2: The 20 most extreme words from each dimension of Sparse CA.

Note: The loadings and contributions shown as 0 in these tables were too small to be displayed as integers; it is worth noting that these values do not indicate sparsity. * indicates that the original values were multiplied by 10,000 and rounded to the nearest integer for display purposes.

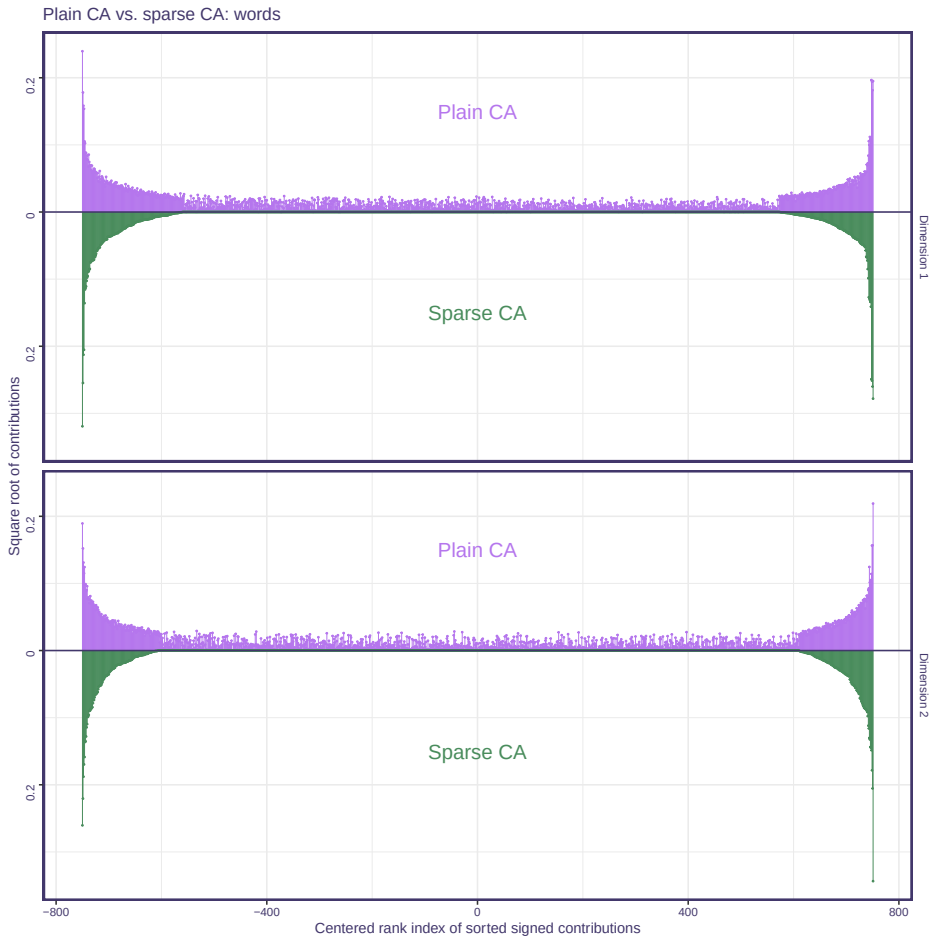

Figure 5: Plain CA vs. Sparse CA: Row contributions

6. Conclusion and Perspectives

In this paper, we extended sparse correspondence analysis developed by Liu et al. (2023) by adding a new global algorithm that identifies the optimal sparsity solution by determining both the optimal sparsity tuning parameters and the optimal number of kept dimensions. Specifically, by integrating this global algorithm, this new version of sparse CA estimates the optimal solution in a more analytic and objective way. Sparse correspondence analysis simplifies the interpretation in the analysis of large tables by highlighting important categories and obtaining

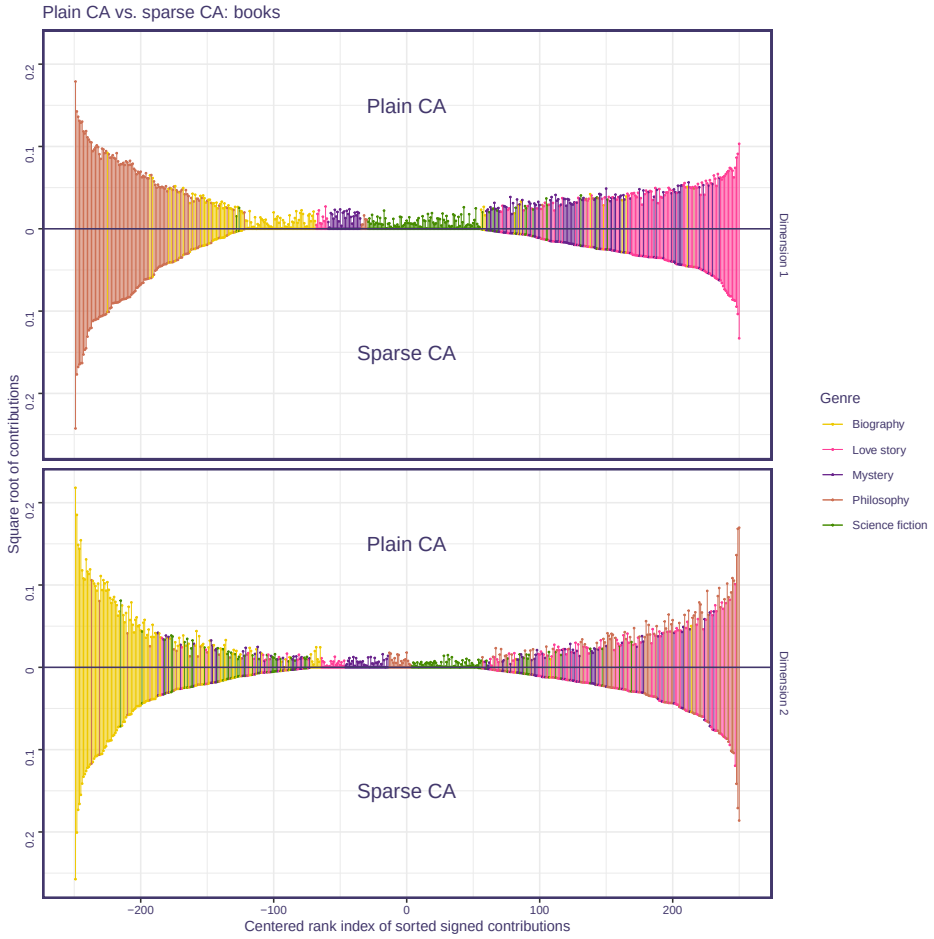

Figure 6: Plain CA vs. Sparse CA: Column contributions

simple successive dimensions in the spirit of the simple structure of factor analysis. Its practical application raises new problems such as the choice of the optimal level of sparsity for rows and or columns, which could be different according to each dimension.

Another concern is the loss of orthogonality of successive dimensions—An issue that should be explored in future work.

Sparse CA remains basically a symmetrical method where rows and columns play the same role. In future work, we also plan to develop sparse variants of the non symmetric correspondence analysis introduced by Lauro and D'Ambra

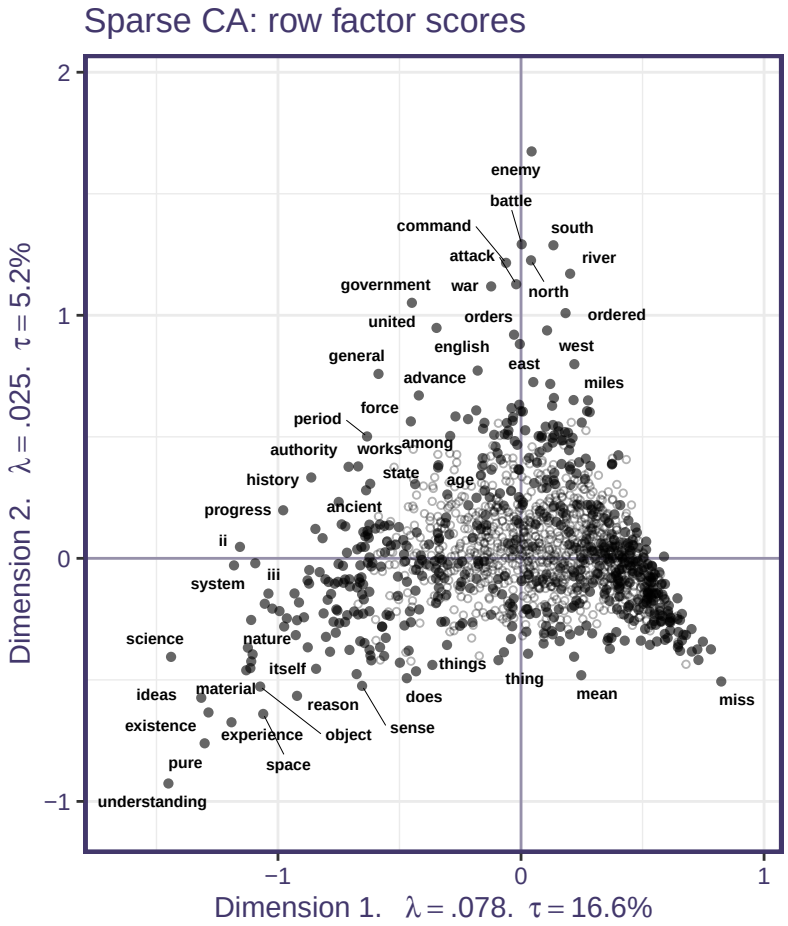

Figure 7: Sparse CA: Row factor scores

(1984) and explored by Balbi (1998).

Code and data are available at:
`https://github.com/vguillemot/sparseCorrespondenceAnalysis`.

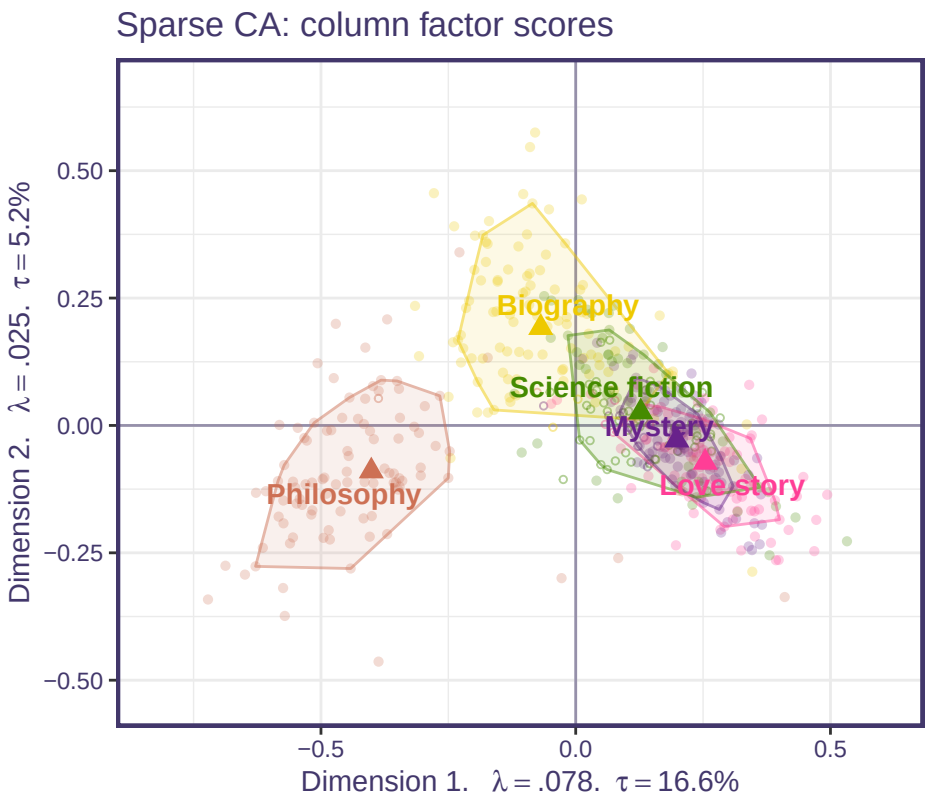

Figure 8: Sparse CA: Column factor scores

References

Abdi, H. (2007). Singular value decomposition (SVD) and generalized singular value decomposition (GSVD). In N. Salkind, ed., *Encyclopedia of Measurement and Statistics*, 907–912. Sage Publications, Thousand Oaks.

Abdi, H. and Béra, M. (2018). Correspondence analysis. In R. Alhajj and J. Rokne, eds., *Encyclopedia of Social Networks and Mining (2nd Edition)*, 1–12. Springer, New York. doi:10.1007/978-1-4614-7163-9_140-2.

Abdi, H., Dunlop, J., and Williams, L.J. (2009). How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the bootstrap and 3-way multidimensional scaling (DISTATIS). In *NeuroImage*, 45 (1): 89–95. doi:10.1016/j.neuroimage.2008.11.008.

Abdi, H. and Williams, L.J. (2010). Principal component analysis. In *WIREs Computational Statistics*, 2 (4): 433–459. doi:https://doi.org/10.1002/wics.101.

Allen, G.I., Grosenick, L., and Taylor, J. (2014). A generalized least-square matrix decomposition. In *Journal of the American Statistical Association*, 109 (505): 145–159. doi:10.1080/01621459.2013.852978.

Balbi, S. (1998). Graphical displays in non-symmetrical correspondence analysis. In J. Blasius and M.J. Greenacre, eds., *Visualization of Categorical Data*, 297–309. Academic Press, San Diego. doi:10.1016/B978-012299045-8/50023-1.

Beaton, D. (2020). Generalized eigen, singular value, and partial least squares decompositions: The GSVD package. doi:10.48550/ARXIV.2010.14734.

Beh, E.J. and Lombardo, R. (2021). *An Introduction to Correspondence Analysis*. Wiley, London. doi:10.1002/9781119044482.

Benidis, K., Sun, Y., Babu, P., and Palomar, D.P. (2016). Orthogonal sparse PCA and covariance estimation via procrustes reformulation. In *IEEE Transactions on Signal Processing*, 64 (23): 6211–6226. doi:10.1109/TSP.2016.2605073.

Bernard, A., Guinot, C., and Saporta, G. (2012). Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis. In A. Colubi and al., eds., *Proceedings of the 20th International Conference on Computational Statistics (COMPSTAT 2012)*, 99–106. International Association for Statistical Computing.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. In *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 72 (1): 3–25. doi:10.1111/j.1467-9868.2009.00723.x.

Dray, S. (2014). Analysing a pair of tables: coinertia analysis and duality. In J. Blasius and M.J. Greenacre, eds., *Visualization and Verbalization of Data*, 289–300. CRC Press, London.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. In *Psychometrika*, 1 (3): 211–218. doi:10.1007/BF02288367.

Escoufier, Y. (2006). Operators related to a data matrix: a survey. In A. Rizzi and M. Vichi, eds., *Proceedings of the 17th International Conference on Computational Statistics (COMPSTAT 2006)*, 285–297. Physica-Verlag, Heidelberg. doi:10.1007/978-3-7908-1709-6_22.

Gerlach, M. and Font-Clos, F. (2020). A standardized project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. In *Entropy*, 22 (1). doi:10.3390/e22010126.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, New York.

Greenacre, M.J. (2010). Correspondence analysis. In *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (5): 613–619.

Guillemot, V., Beaton, D., Gloaguen, A., Löfstedt, T., Levine, B., Raymond, N., Tenenhaus, A., and Abdi, H. (2019). A constrained singular value decomposition method that integrates sparsity and orthogonality. In *PLOS ONE*, 14 (3): e0211463. doi:10.1371/journal.pone.0211463.

Guillemot, V., Le Borgne, J., Gloaguen, A., Tenenhaus, A., Saporta, G., Chollet, S., Beaton, D., and Abdi, H. (2020). Sparse multiple correspondence analysis. In *52èmes Journées de Statistique*, 830–835. URL `https://pasteur.hal.science/pasteur-03037346/`.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York. doi:10.1007/978-0-387-84858-7.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press, Boca Raton. doi:10.1201/b18401.

Holmes, S. (2008). Multivariate data analysis: The French way. In *Probability and Statistics: Essays in Honor of David A. Freedman*, vol. 2, 219–234. Institute of Mathematical Statistics. doi:10.1214/193940307000000455.

Jenatton, R., Audibert, J.Y., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. In *The Journal of Machine Learning Research*, 12: 2777–2824. doi:10.5555/1953048.2078194.

Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. In *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 374 (2065). doi: 10.1098/rsta.2015.0202.

Jolliffe, I.T., Trendafilov, N.T., and Uddin, M. (2003). A modified principal component technique based on the LASSO. In *Journal of Computational and Graphical Statistics*, 12 (3): 531–547. doi:10.1198/1061860032148.

Lauro, N. and D'Ambra, L. (1984). L'analyse non symétrique des correspondances. In E. Diday, ed., *Data Analysis and Informatics III*, 433–446. Elsevier, North–Holland.

Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J.B., and Duchesnay, E. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. In *NeuroImage*, 63 (1): 11–24. doi:10.1016/j.neuroimage.2012.06.061.

Lebart, L., Morineau, A., and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. Wiley, New York.

Liu, R., Niang, N., Saporta, G., and Wang, H. (2023). Sparse correspondence analysis for large contingency tables. In *Advances in Data Analysis and Classification*, 17: 1–20. doi:10.1007/s11634-022-00531-5.

Mackey, L. (2009). Deflation methods for sparse PCA. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds., *Advances in Neural Information Processing Systems*, vol. 21, 1017–1024. Curran Associates, Inc.

Mattei, P.A., Bouveyron, C., and Latouche, P. (2016). Globally sparse probabilistic pca. In A. Gretton and C.C. Robert, eds., *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, vol. 51 of *Proceedings of Machine Learning Research*, 976–984. PMLR, Cadiz, Spain.

Mori, Y., Kuroda, M., and Makino, N. (2016). Sparse multiple correspondence analysis. In Y. Mori, M. Kuroda, and N. Makino, eds., *Nonlinear Principal Component Analysis and its Applications*, 47–56. Springer, New York. doi: 10.1007/978-981-10-0159-8_5.

Ning-min, S. and Jing, L. (2015). A literature survey on high-dimensional sparse principal component analysis. In *International Journal of Database Theory and Application*, 8 (6): 57–74. doi:10.14257/ijdta.2015.8.6.06.

Saporta, G. and Niang-Keita, N. (2006). Correspondence analysis and classification. In M.J. Greenacre and J. Blasius, eds., *Multiple Correspondence Analysis and Related Methods*, 371–392. CRC Press, London. doi:10.1201/9781420011319-19.

Shen, D., Shen, H., and Marron, J.S. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. In *Journal of Multivariate Analysis*, 115: 317–333. doi:10.1016/j.jmva.2012.10.007.

Silver, M., Janousova, E., Hua, X., Thompson, P.M., Montana, G., and Alzheimer's Disease Neuroimaging Initiative, T.A.D.N. (2012). Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. In *NeuroImage*, 63 (3): 1681–1694. doi:10.1016/j.neuroimage.2012.08.002.

Takane, Y. (2002). Relationships among various kinds of eigenvalue and singular value decompositions. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J. Meulman, eds., *New Developments in Psychometrics*, 45–56. Springer Verlag, Tokyo.

Thurstone, L.L. (1935). *The Vectors of Mind: Multiple Factor Analysis for the Isolation of Primary Traits.* University of Chicago Press. doi:10.1037/10018-000.

Thurstone, L.L. (1947). *Multiple Factor Analysis.* University of Chicago Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. In *Journal of the Royal Statistical Society. Series B (Methodological)*, 58 (1): 267–288. URL `http://www.jstor.org/stable/2346178`.

Trendafilov, N.T. (2014). From simple structure to sparse components: a review. In *Computational Statistics*, 29 (3–4): 431–454. doi:10.1007/s00180-013-0434-5.

Trendafilov, N.T., Fontanella, S., and Adachi, K. (2017). Sparse exploratory factor analysis. In *Psychometrika*, 82 (3): 778–794. doi:10.1007/s11336-017-9575-8.

Vines, S. (2000). Simple principal components. In *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49 (4): 441–451. doi:10.1111/1467-9876.00204.

Witten, D.M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. In *Biostatistics*, 10 (3): 515–534. doi:10.1093/biostatistics/kxp008.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. In *Journal of Computational and Graphical Statistics*, 15 (2): 265–286. doi:10.1198/106186006X113430.

Appendix

A. Notations

Matrices are denoted in upper case bold letters, vectors are denoted in lowercase bold letters, and their elements are denoted in lowercase italic letters (note that, by default, vectors are column vectors). Matrices, vectors and elements from the same matrix all use the same letter (e.g., $\mathbf{A}$, $\mathbf{a}$, a). The transpose operation is denoted by the superscript $^\mathsf{T}$, the inverse operation is denoted by $^{-1}$. The identity matrix is denoted $\mathbf{I}$, vectors or matrices of ones are denoted $\mathbf{1}$, matrices or vectors of zeros are denoted $\mathbf{0}$ (by default, $\mathbf{I}$, $\mathbf{0}$, and $\mathbf{1}$ are conformable with the other terms in a formula). The standard product between two matrices is indicated by juxtaposition (i.e., $\mathbf{XY}$ means $\mathbf{X}$ times $\mathbf{Y}$); the Hadamard product (i.e., element-wise) is denoted by $\odot$ (e.g., $\mathbf{X} \odot \mathbf{Y}$), note that the Hadamard product is defined only between matrices with the same dimensions.

When provided with a square matrix, the diag operator gives a vector that contains the diagonal elements of this matrix. When provided with a vector, the diag operator gives a diagonal matrix with the elements of the vector as the diagonal elements of this matrix. A diagonal matrix is denoted $\mathbf{D}$, and the subscript denotes the vector that stores the diagonal elements, for example, $\mathbf{D}_\mathbf{a} = \mathsf{diag}\left(\mathbf{a}\right)$. When provided with a square matrix, the trace operator gives the sum of the diagonal elements of this matrix. For an I by J matrix $\mathbf{X}$ and for $\mathbf{M}$ being a J by J symmetric positive definite matrix, the squared $\mathbf{M}$-norm of $\mathbf{X}$ is denoted $\|\mathbf{X}\|^2_\mathbf{M}$ and is computed as:

$$\|\mathbf{X}\|^2_\mathbf{M} = \mathsf{trace}\left(\mathbf{XMX}^\mathsf{T}\right) \ . \tag{16}$$

When $\mathbf{M}$ is the identity matrix, the $\mathbf{M}$-norm is equal to the square root of the sum of squares of the entries of the matrix and is called the *Frobenius* norm denoted $L_2 = \|\mathbf{X}\|^2_2$. Another useful norm is the sum of the absolute values of the matrix called the L_1 norm.

When describing an optimization problem, the operator $\underset{\mathbf{x}}{\arg\min} f(\mathbf{x})$ searches for the value of $\mathbf{x}$ that minimizes the function $f(\mathbf{x})$, and the operator $\underset{\mathbf{x}}{\arg\max} f(\mathbf{x})$ searches for the value of $\mathbf{x}$ that maximizes the function $f(\mathbf{x})$.

B. The Plain and Generalized Singular Value Decompositions

The singular value decomposition (SVD) and its extension—the generalized singular value decomposition (GSVD, for details on the generalized singular value

decomposition see Abdi 2007; Greenacre 1984; Takane 2002)—are the foundations of most contemporary multivariate statistical approaches.

B.1. The (Plain) Singular Value Decomposition

The SVD of an $I \times J$ matrix $\mathbf{X}$ solves the following maximization problem (Eckart and Young, 1936): Find a matrix, denoted $\widehat{\mathbf{X}}_L$, of rank L [with $L < \min(I,J)$] equal to

$$\widehat{\mathbf{X}}_L = \sum_{\ell=1}^{L} \delta_\ell \mathbf{u}_\ell \mathbf{v}_\ell^\mathsf{T} = \mathbf{U}_L \mathbf{\Delta}_L \mathbf{V}_L^\mathsf{T} \text{ with } \mathbf{U}_L^\mathsf{T}\mathbf{U}_L = \mathbf{V}_L^\mathsf{T}\mathbf{V}_L = \mathbf{I} \text{ and } \mathbf{\Delta}_L = \mathsf{diag}\,(\boldsymbol{\delta}_L) \quad (17)$$

where $\mathbf{U}$ is the $I \times L$ matrix containing the left singular vectors, $\mathbf{V}$ is the $J \times L$ matrix containing the right singular vectors, and $\mathbf{\Delta}$ the $L \times L$ diagonal matrix containing the singular values $\delta_1 \geq \cdots \geq \delta_L \geq 0$, and such that $\widehat{\mathbf{X}}_L$ is the L rank matrix closest to $\mathbf{X}$. Specifically, $\widehat{\mathbf{X}}_L$ solves the following minimization problem:

$$\underset{\mathbf{U}_L,\mathbf{\Delta}_L,\mathbf{V}_L}{\arg\min} \|\mathbf{X} - \widehat{\mathbf{X}}_L\|_2^2 = \underset{\mathbf{U}_L,\mathbf{\Delta}_L,\mathbf{V}_L}{\arg\min} \|\mathbf{X} - \mathbf{U}_L\mathbf{\Delta}_L\mathbf{V}_L^\mathsf{T}\|_2^2 \quad \text{with} \quad \mathbf{U}_L^\mathsf{T}\mathbf{U}_L = \mathbf{V}_L^\mathsf{T}\mathbf{V}_L = \mathbf{I}, \quad (18)$$

When L is equal to the rank of $\mathbf{X}$, the SVD of $\mathbf{X}$ is called the *complete* SVD (when unspecified, the SVD is the complete SVD), in this case, matrices $\mathbf{U}$ and $\mathbf{V}$ are written without their L index. When L is smaller than the rank of $\mathbf{X}$, its SVD is called the *truncated* SVD of $\mathbf{X}$.

The SVD of a matrix can be computed by first computing its rank one approximation [i.e., the singular triplet (δ_1, $\mathbf{u}_1$, $\mathbf{v}_1$)] and then subtracting this rank one approximation from $\mathbf{X}$—a procedure called a *deflation*. The first singular triplet of the deflated matrix $\mathbf{X}$ is obtained then the second singular triplet of $\mathbf{X}$. These procedure can then be continued till completion of the SVD of $\mathbf{X}$.

B.2. Generalized Singular Value Decomposition

The generalized SVD (GSVD), differs from the plain SVD by incorporating different orthogonality constraints on the singular vectors. Specifically, with $\mathbf{M}$ being an $I \times I$ positive definite matrix (called the row *metric* matrix) and $\mathbf{W}$ a $J \times J$ positive definite matrix (called the column metric matrix); the GSVD of $\mathbf{X}$ solves the following problem (compare with Equation 18): Specifically, $\widehat{\mathbf{X}}_L$ solves

$$\underset{\mathbf{P}_L,\mathbf{\Delta}_L,\mathbf{Q}_L}{\arg\min} \|\mathbf{X} - \widehat{\mathbf{X}}_L\|_2^2 = \underset{\mathbf{P}_L,\mathbf{\Delta}_L,\mathbf{Q}_L}{\arg\min} \|\mathbf{X} - \mathbf{P}_L\mathbf{\Delta}_L\mathbf{Q}_L^\mathsf{T}\|_2^2 \quad (19)$$

with

$$\mathbf{P}_L^\mathsf{T}\mathbf{M}\mathbf{P}_L = \mathbf{Q}_L^\mathsf{T}\mathbf{W}\mathbf{Q}_L = \mathbf{I}, \boldsymbol{\Delta}_L = \mathsf{diag}\left(\boldsymbol{\delta}_L\right). \tag{20}$$

where $\mathbf{P}_L$ is the $I \times L$ matrix containing the *generalized* left singular vectors, $\mathbf{Q}_L$ is the $J \times L$ matrix containing the generalized right singular vectors, and $\boldsymbol{\Delta}_L$ is the diagonal matrix of the generalized singular values.

Similarly to the plain SVD, the optimal rank-L approximation of $\mathbf{X}$ is obtained by $\widehat{\mathbf{X}}_L$ (i.e., the L-truncated GSVD of $\mathbf{X}$) as:

$$\widehat{\mathbf{X}}_L = \sum_{\ell=1}^{L} \delta_\ell \mathbf{p}_\ell \mathbf{q}_\ell^\mathsf{T} = \mathbf{P}_L \boldsymbol{\Delta}_L \mathbf{Q}_L^\mathsf{T}. \tag{21}$$

B.3. Generalized SVD from Plain SVD

In practice, the GSVD matrix $\mathbf{X}$ can be obtained from a plain SVD of a matrix denoted $\widetilde{\mathbf{X}}$ obtained by first pre- and post-multiplying $\mathbf{X}$ by the square root of the row and column metric matrices:

$$\widetilde{\mathbf{X}} = \mathbf{M}^{\frac{1}{2}}\mathbf{X}\mathbf{W}^{\frac{1}{2}} \ . \tag{22}$$

Matrix $\widetilde{\mathbf{X}}$ is then decomposed with a plain SVD as:

$$\widetilde{\mathbf{X}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\mathsf{T} \quad \text{such that} \quad \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}. \tag{23}$$

The generalized singular vectors of $\mathbf{X}$ are then obtained from the (plain) singular vectors of $\widetilde{\mathbf{X}}$ as

$$\mathbf{P} = \mathbf{M}^{-\frac{1}{2}}\mathbf{U} \quad \text{and} \quad \mathbf{Q} = \mathbf{W}^{-\frac{1}{2}}\mathbf{V}. \tag{24}$$

The constraints from Equations 18 and 19 are equivalent because

$$\begin{aligned} \mathbf{P}^\mathsf{T}\mathbf{M}\mathbf{P} &= \mathbf{U}^\mathsf{T}\mathbf{M}^{-\frac{1}{2}}\mathbf{M}\mathbf{M}^{-\frac{1}{2}}\mathbf{U} \\ &= \mathbf{U}^\mathsf{T}\mathbf{U} \\ &= \mathbf{I}, \end{aligned} \quad \text{and} \quad \begin{aligned} \mathbf{Q}^\mathsf{T}\mathbf{W}\mathbf{Q} &= \mathbf{V}^\mathsf{T}\mathbf{W}^{-\frac{1}{2}}\mathbf{W}\mathbf{W}^{-\frac{1}{2}}\mathbf{V} \\ &= \mathbf{V}^\mathsf{T}\mathbf{V} \\ &= \mathbf{I}. \end{aligned} \tag{25}$$

Finally, the decomposition of $\mathbf{X}$ from Equations 18 and 19 are also equivalent because

$$\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\mathsf{T} = \mathbf{M}^{-\frac{1}{2}}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\mathsf{T}\mathbf{W}^{-\frac{1}{2}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{M}^{\frac{1}{2}}\mathbf{X}\mathbf{W}^{\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}} = \mathbf{X} \ . \tag{26}$$

C. Plain Correspondence Analysis

Just like most multivariate methods, CA can be interpreted as an optimization problem (actually, as several equivalent optimization problems). But, in order to sparsify CA, we will need to add more constraints to its standard GSVD optimization problem. These new constraints can, in some cases, conflict with the original optimization problem and therefore, as a trade-off, some of the essential properties of CA could be relaxed or even lost. To facilitate the evaluation of this trade-off, we list below the relevant basic equations for CA along with its essential properties (for more details see, e.g., Abdi and Béra, 2018; Abdi and Williams, 2010; Beh and Lombardo, 2021; Greenacre, 1984; Lebart et al. 1984; or Saporta and Niang-Keita, 2006).

C.1. The Basic Equations of Correspondence Analysis

Correspondence analysis was originally developed to analyze the pattern of deviations from independence (as measured by a χ^2 statistic) in a contingency table (see Abdi and Béra, 2018). CA provides, for both rows and columns, a set of factor scores whose total inertia is proportional to the independence χ^2 computed on the original contingency table.

The contingency table to be analyzed is stored in an I rows by J columns matrix denoted $\mathbf{X}$, whose generic element $x_{i,j}$ gives the number of observations that belongs to the ith level of the first nominal variable (i.e., the rows) andthe jth level of the second nominal variable (i.e., the columns). The grand total of the table is denoted N.

The matrix $\mathbf{X}$ is first transformed into a probability matrix (i.e., a matrix comprising non-negative numbers and whose sum is equal to one) denoted $\mathbf{Z}$ and computed as $\mathbf{Z} = N^{-1}\mathbf{X}$. We denote: $\mathbf{r}$ the I by 1 vector of the row totals of $\mathbf{Z}$ and by r_i the ith element of $\mathbf{r}$ (i.e., $\mathbf{r} = \mathbf{Z1}$, with $\mathbf{1}$ being a conformable vector of 1's); $\mathbf{c}$ the J by 1 vector of the columns totals, by c_j the jth element of $\mathbf{c}$ (i.e., $\mathbf{c} = \mathbf{Z}^{\mathsf{T}}\mathbf{1}$); and $\mathbf{D_c} = \mathsf{diag}\,(\mathbf{c})$, $\mathbf{D_r} = \mathsf{diag}\,(\mathbf{r})$ the diagonal matrices obtained from (respectively) $\mathbf{r}$ and $\mathbf{c}$; these two diagonal matrices are called (respectively) row and column *mass* matrices. We denote by $\mathbf{R} = \mathbf{D_r}^{-1}\mathbf{Z}$ (respectively $\mathbf{C} = \mathbf{D_c}^{-1}\mathbf{Z}^{\mathsf{T}}$) the row (respectively column) profile matrix (i.e., all elements are not negative, rows of $\mathbf{R}$ and columns of $\mathbf{C}$ sum to 1).

The factor scores are obtained from the following generalized singular value decomposition (cf. Equation 19) where the metric matrices $\mathbf{D_r}^{-1}$ and $\mathbf{D_c}^{-1}$ are

called χ^2-metric matrices (Greenacre, 2010)

$$\mathbf{Z} - \mathbf{r}\mathbf{c}^\mathsf{T} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\mathsf{T} \text{ with } \mathbf{P}^\mathsf{T}\mathbf{D}_\mathbf{r}^{-1}\mathbf{P} = \mathbf{Q}^\mathsf{T}\mathbf{D}_\mathbf{c}^{-1}\mathbf{Q} = \mathbf{I}\,. \tag{27}$$

Here, by subtracting the (rank one) matrix $\mathbf{r}\mathbf{c}^\mathsf{T}$ from the probability matrix $\mathbf{Z}$, the decomposed matrix: $\left(\mathbf{Z} - \mathbf{r}\mathbf{c}^\mathsf{T}\right)$ is *double-centered* because now all rows and columns have zero means. In addition, this double centering of matrix $\mathbf{Z} - \mathbf{r}\mathbf{c}^\mathsf{T}$ propagates to the singular vectors.

For example, to show that matrix $\mathbf{Q}$ has zero mean, we compute the column means of $\mathbf{Z} - \mathbf{r}\mathbf{c}^\mathsf{T}$ and replace it by its GSVD from Equation 27 to get

$$\left(\mathbf{Z} - \mathbf{r}\mathbf{c}^\mathsf{T}\right)\mathbf{1} = \mathbf{0} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\mathsf{T}\mathbf{1} \Longrightarrow \mathbf{P}^\mathsf{T}\mathbf{D}_\mathbf{r}^{-1}\mathbf{P}\Delta\mathbf{Q}^\mathsf{T}\mathbf{1} = \mathbf{0} \Longrightarrow \Delta\mathbf{Q}^\mathsf{T}\mathbf{1} = \mathbf{0} \Longrightarrow \mathbf{Q}^\mathsf{T}\mathbf{1} = \mathbf{0}\,, \tag{28}$$

where $\mathbf{1}$ is a J by 1 vector of 1s.

The squared singular values are called *eigenvalues* (denoted λ_k) and are stored into the diagonal matrix $\boldsymbol{\Lambda}$. The sum of the eigenvalues gives the total inertia (denoted $\mathscr{I}$ or φ^2 and equal to χ^2/N) of $\left(\mathbf{Z} - \mathbf{r}\mathbf{c}^\mathsf{T}\right)$. With the so-called "triplet notation," (Dray, 2014; Escoufier, 2006; Holmes, 2008)—sometimes used as a general framework to formalize multivariate techniques—CA is equivalent to the analysis of the triplet $\left(\mathbf{Z} - \mathbf{r}\mathbf{c}^\mathsf{T}, \mathbf{D}_\mathbf{c}^{-1}, \mathbf{D}_\mathbf{r}^{-1}\right)$. From this GSVD, the principal row and (respectively) column factor scores are obtained as

$$\mathbf{F} = \mathbf{D}_\mathbf{r}^{-1}\mathbf{P}\boldsymbol{\Delta} \qquad \text{and} \qquad \mathbf{G} = \mathbf{D}_\mathbf{c}^{-1}\mathbf{Q}\boldsymbol{\Delta}\,. \tag{29}$$

Note that the inertia of each dimension (i.e., each column of $\mathbf{F}$ and $\mathbf{G}$) is equal to its eigenvalue and that factor scores corresponding to different eigenvalues are orthogonal (under the constraints imposed by their masses). Specifically:

$$\lambda_\ell = \sum_{i=1}^{I} r_i f_{i,\ell}^2 = \sum_{j=1}^{J} c_j g_{j,\ell}^2 \text{ and } \sum_{i=1}^{I} r_i f_{i,\ell} f_{i,\ell'} = \sum_{j=1}^{J} c_j g_{j,\ell} g_{j,\ell'} = 0 \quad \forall \ell \neq \ell' \tag{30}$$

or, in matrix notations:

$$\mathbf{F}^\mathsf{T}\mathbf{D}_\mathbf{r}\mathbf{F} = \boldsymbol{\Lambda} \text{ and } \mathbf{G}^\mathsf{T}\mathbf{D}_\mathbf{c}\mathbf{G} = \boldsymbol{\Lambda}\,, \tag{31}$$

where $\mathbf{D}_\mathbf{r}$ and $\mathbf{D}_\mathbf{c}$ are called (respectively) the row and column *mass* matrices. This equality can be directly derived from Equations 27 and 29 (here illustrated for $\mathbf{F}$):

$$\mathbf{F}^\mathsf{T}\mathbf{D}_\mathbf{r}\mathbf{F} = \boldsymbol{\Delta}\mathbf{P}^\mathsf{T}\mathbf{D}_\mathbf{r}^{-1}\mathbf{D}_\mathbf{r}\mathbf{D}_\mathbf{r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \boldsymbol{\Delta}\mathbf{P}^\mathsf{T}\mathbf{D}_\mathbf{r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \boldsymbol{\Delta}^2 = \boldsymbol{\Lambda}\,. \tag{32}$$

By contrast with the *principal* factor scores whose $\mathbf{D}_\mathbf{r}$ and $\mathbf{D}_\mathbf{c}$ norms are equal to the singular values, the *standard* factor scores (indicated by a superscript *) have $\mathbf{D}_\mathbf{r}$ and $\mathbf{D}_\mathbf{c}$ norms equal to one, and are computed as

$$\mathbf{F}^* = \mathbf{D}_\mathbf{r}^{-1}\mathbf{P} \qquad \text{and} \qquad \mathbf{G}^* = \mathbf{D}_\mathbf{c}^{-1}\mathbf{Q}\,. \tag{33}$$

C.2. Correspondence Analysis from a Plain SVD

Correspondence analysis can also be obtained from both the plain SVD and the GSVD (for details, see, e.g., Abdi 2007, and Beaton 2020). Specifically, generalized singular vectors and values and factor scores can be obtained by the following plain SVD:

$$\widetilde{\mathbf{Z}} = \mathbf{D_r}^{-\frac{1}{2}} \left(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\mathsf{T}}\right) \mathbf{D_c}^{-\frac{1}{2}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathsf{T}} \tag{34}$$

which, in turn, gives the generalized singular vectors as

$$\mathbf{P} = \mathbf{D_r}^{\frac{1}{2}}\mathbf{U} \quad \text{and} \quad \mathbf{Q} = \mathbf{D_c}^{\frac{1}{2}}\mathbf{V}\,. \tag{35}$$

Finally, transposing this last equation in Equation 29 gives:

$$\mathbf{F} = \mathbf{D_r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \mathbf{D_r}^{-\frac{1}{2}}\mathbf{U}\boldsymbol{\Delta} \qquad \text{and} \qquad \mathbf{G} = \mathbf{D_c}^{-1}\mathbf{Q}\boldsymbol{\Delta} = \mathbf{D_c}^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Delta}\,. \tag{36}$$

As indicated by Equation 30, the inertia of a dimension is the sum of the inertia of either all the rows or all the columns, therefore a convenient way of evaluating the importance of a row (respectively a column) is to compute the proportion accounted by a given row (respectively column) into this total. This index, called the *contribution* of a row (respectively a column) is denoted $t_{i,\ell}$ (respectively $t_{j,\ell}$ for a column) and is computed as

$$t_{i,\ell} = \frac{r_i f_{i,\ell}^2}{\sum_{i'=1}^{I} r_{i'} f_{i',\ell}^2} = r_i f_{i,\ell}^2 \lambda_\ell^{-1} \quad \text{and} \quad t_{j,\ell} = \frac{c_j g_{j,\ell}^2}{\sum_{j'=1}^{J} c_{j'} g_{j',\ell}^2} = c_j g_{j,\ell}^2 \lambda_\ell^{-1} \tag{37}$$

In matrix notations, the row (respectively columns) contributions are stored in the matrix $\mathbf{T_I}$ (respectively $\mathbf{T_J}$) computed as:

$$\mathbf{T_I} = \mathbf{D_r}\left(\mathbf{F} \odot \mathbf{F}\right)\boldsymbol{\Lambda}^{-1} = \mathbf{D_r}^{-1}\left(\mathbf{P} \odot \mathbf{P}\right) \text{ and } \mathbf{T_J} = \mathbf{D_c}\left(\mathbf{G} \odot \mathbf{G}\right)\boldsymbol{\Lambda}^{-1} = \mathbf{D_c}^{-1}\left(\mathbf{Q} \odot \mathbf{Q}\right). \tag{38}$$

Note that contributions can be obtained in two equivalent ways: from the factor scores or from the generalized singular vectors.

To facilitate the interpretation of a given dimension, to interpret a dimension we, traditionally, use only the items whose contribution is larger than their mass (i.e., r_i or c_j). The contributions are also often plotted according to the sign of their corresponding factor scores and are then called *signed* contributions.

C.3. Important Properties of Correspondence Analysis

In this section we list some important properties of correspondence analysis relevant for sparsification.

C.3.1. Inertia and χ^2

The inertia (i.e., $\mathscr{I}$ or equivalently φ^2) of the centered matrix $\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^\mathsf{T}\right)$—as obtained from Equation 27—is equal to the independence χ^2 divided by N. Recall that, with the present notations, $\chi^2/N = \varphi^2$ is computed as

$$\varphi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(z_{i,j}-r_i c_j\right)^2}{r_i c_j} = \mathsf{trace}\left(\mathbf{D_c}^{-\frac{1}{2}}\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^\mathsf{T}\right)^\mathsf{T}\mathbf{D_r}^{-1}\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^\mathsf{T}\right)\mathbf{D_c}^{-\frac{1}{2}}\right). \quad (39)$$

To show that φ^2 is equal to the sum of the eigenvalues from Equation 27, suffice to plug the singular values decomposition from Equation 27 into Equation 39 to get:

$$\varphi^2 = \mathsf{trace}\left(\mathbf{D_c}^{-\frac{1}{2}}\left(\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\mathsf{T}\right)^\mathsf{T}\mathbf{D_r}^{-1}\left(\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\mathsf{T}\right)\mathbf{D_c}^{-\frac{1}{2}}\right). \quad (40)$$

Using the properties of the trace operator and re-arranging shows that φ^2 is equal to the sum of the eigenvalues of $\mathbf{Z}-\mathbf{r}\mathbf{c}^\mathsf{T}$, namely that:

$$\varphi^2 = \mathsf{trace}\left(\boldsymbol{\Delta}\mathbf{P}^\mathsf{T}\mathbf{D_c}^{-1}\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\mathsf{T}\mathbf{D_c}^{-1}\mathbf{Q}\right) = \mathsf{trace}\left(\boldsymbol{\Delta}^2\right) = \mathsf{trace}\left(\boldsymbol{\Lambda}\right) = \sum_{\ell=1}^{L}\lambda_\ell\,. \quad (41)$$

which shows, as stated, that $\varphi^2 = \sum\lambda_\ell$.

C.3.2. Factors are Centered

The centering of the singular vectors propagates to the factor scores when the means are computed using the masses stored in the diagonal matrices $\mathbf{D_r}$ (for the rows) and $\mathbf{D_c}$ (for the columns). So, $\overline{\mathbf{F}}$ (respectively $\overline{\mathbf{G}}$), denoting the average row (respectively column) factor scores, is computed as

$$\overline{\mathbf{F}} = \mathbf{1}\mathbf{D_r}\mathbf{F} = \mathbf{1}\mathbf{D_r}\mathbf{D_r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \mathbf{0} \text{ and } \overline{\mathbf{G}} = \mathbf{1}\mathbf{D_c}\mathbf{G} = \mathbf{1}\mathbf{D_c}\mathbf{D_c}^{-1}\mathbf{Q}\boldsymbol{\Delta} = \mathbf{0} \quad (42)$$

(where $\mathbf{1}$ and $\mathbf{0}$ are conformable vectors of 1s and 0s).

C.3.3. Transition Formulas: from Row to Column Factor Scores and Back

In CA the factor scores of one set (e.g., the rows) can be obtained from the profiles of this set and the factor scores of the other set (e.g., the columns). Specifically we have

$$\mathbf{F} = \mathbf{D_r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \mathbf{R}\mathbf{G}\boldsymbol{\Delta}^{-1} \text{ and } \mathbf{G} = \mathbf{D_c}^{-1}\mathbf{Q}\boldsymbol{\Delta} = \mathbf{C}\mathbf{F}\boldsymbol{\Delta}^{-1} \ . \tag{43}$$

These formulas called *transition formulas* are obtained from Equations 27 and 29; for example, the transition formula for the row factor scores (i.e., from the column factor scores) is derived as (taking into account that $\mathbf{Q}$ is centered)

$$\begin{aligned}
\mathbf{F} &= \mathbf{D_r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \mathbf{D_r}^{-1}\left(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\mathsf{T}}\right)\mathbf{D_c}^{-1}\mathbf{Q} \quad \left[\text{because } \mathbf{P}\boldsymbol{\Delta} = \left(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\mathsf{T}}\right)\mathbf{D_c}^{-1}\mathbf{Q}\right] \\
&= \mathbf{D_r}^{-1}\mathbf{Z}\mathbf{D_c}^{-1}\mathbf{Q} - \mathbf{D_r}^{-1}\mathbf{r}\mathbf{c}^{\mathsf{T}}\mathbf{D_c}^{-1}\mathbf{Q} \\
&= \mathbf{D_r}^{-1}\mathbf{Z}\mathbf{D_c}^{-1}\mathbf{Q} \quad \left[\text{because } \mathbf{c}^{\mathsf{T}}\mathbf{D_c}^{-1}\mathbf{Q} = \mathbf{1}^{\mathsf{T}}\mathbf{Q} = \mathbf{0}\right] \\
&= \mathbf{D_r}^{-1}\mathbf{Z}\mathbf{D_c}^{-1}\mathbf{D_c}\mathbf{G}\boldsymbol{\Delta}^{-1} \quad \left[\text{because } \mathbf{Q} = \mathbf{D_c}\mathbf{G}\boldsymbol{\Delta}^{-1}\right] \\
&= \mathbf{D_r}^{-1}\mathbf{Z}\mathbf{G}\boldsymbol{\Delta}^{-1} \\
&= \mathbf{R}\mathbf{G}\boldsymbol{\Delta}^{-1}.
\end{aligned} \tag{44}$$

Note that, together, the two transition formulas imply that the eigenvalues in CA cannot be larger than 1.

The transition formulas can be interpreted as a two step process. Using the formula above (for computing $\mathbf{F}$ from $\mathbf{G}$) The first step corresponds to the term $\mathbf{RG}$ and computes the row factor scores as the weighted average (i.e., the *barycenter*) of the column factor scores; the second step corresponds to the term $\boldsymbol{\Delta}^{-1}$ and is an expansion that is inversely proportional to the singular value of each factor (this is an expansion because the singular values being no larger than 1, their inverse is no smaller than 1).

C.3.4. Correspondence Analysis as a Double Principal Component Analysis

The row and column factor scores of CA can also be obtained from two different GSVD (or equivalently two weighted PCA), one performed on the row profiles (i.e., the matrix $\mathbf{R}$) and the other one on the column profiles (i.e., the matrix $\mathbf{C}$).

This way, the factor scores are obtained from the GSVD of the matrix of the row profiles matrix (i.e., $\mathbf{R}$) as:

$$\mathbf{D_r}^{-1}\left(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\mathsf{T}}\right) = \left(\mathbf{R} - \mathbf{1}\mathbf{c}^{\mathsf{T}}\right) = \mathbf{P_R}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} \text{ with } \mathbf{P_R}^{\mathsf{T}}\mathbf{D_r}\mathbf{P_R} = \mathbf{Q}^{\mathsf{T}}\mathbf{D_c}^{-1}\mathbf{Q} = \mathbf{I} \tag{45}$$

where $\mathbf{P_R}$ contains the left generalized singular vectors of the row profile matrix $\mathbf{R}$. We can link the decomposition of the row profile matrix to the centered data as:

$$\mathbf{P_R} = \mathbf{D_r^{-1}P}, \quad \mathbf{F} = \mathbf{P_R\Delta} = \mathbf{D_r^{-1}P\Delta}, \quad \text{and} \quad \mathbf{G} = \mathbf{D_c^{-1}Q\Delta}\ ; \tag{46}$$

But these factor scores can also be obtained from the GSVD of the matrix of the column profiles (i.e., matrix $\mathbf{C}$) as:

$$\left(\mathbf{Z} - \mathbf{rc}^\mathsf{T}\right)\mathbf{D_c^{-1}} = \left(\mathbf{C} - \mathbf{1c}^\mathsf{T}\right) = \mathbf{P\Delta Q_C^\mathsf{T}} \text{ with } \mathbf{P^\mathsf{T}D_r^{-1}P} = \mathbf{Q_C^\mathsf{T}D_cQ_C} = \mathbf{I} \tag{47}$$

where $\mathbf{Q_C}$ contains the right generalized singular vectors of the column profile matrix $\mathbf{C}$. We can link the decomposition of the column profile matrix to the centered data as:

$$\mathbf{F} = \mathbf{D_c^{-1}P\Delta}, \qquad \mathbf{Q_C} = \mathbf{D_c^{-1}Q} \quad \text{and} \quad \mathbf{G} = \mathbf{Q_C\Delta} = \mathbf{D_c^{-1}Q\Delta}. \tag{48}$$

Within the framework of generalized PCA, Equations 44, 46, and 48 show, together, that the matrices of the principal factor scores can be obtained as linear combinations of the row (respectively column) profile matrix as (respectively):

$$\mathbf{F} = \mathbf{RD_c^{-1}Q} \quad \text{and} \quad \mathbf{G} = \mathbf{C^\mathsf{T}D_r^{-1}P}\ . \tag{49}$$

In this framework, the matrix $\mathbf{D_c^{-1}Q}$ (respectively $\mathbf{D_r^{-1}P}$) that stores the coefficients of the linear combinations of the columns of $\mathbf{R}$ (respectively $\mathbf{C^\mathsf{T}}$) is called the matrix of the *row-weights* (respectively *column-weights*). Note that the weight matrix for one set (e.g., matrix $\mathbf{D_c^{-1}Q}$ for the rows) is the matrix of the standard coordinates for the other set (i.e., $\mathbf{G^*} = \mathbf{D_c^{-1}Q}$, see Equation 33).