e-ISSN 2320 –7876 www.ijfans.org

Research Paper

Vol.11, Iss.9, Dec 2022

A Study on Spares Data Analysis of Ixj Contingence Tables

*1 JEYASEELI. S, MPhil Scholar, Department of Mathematics Bharath Institute of Higher Education and Research, Chennai,-73, India.
*2 Dr. R. Ishwariya , Associate Professor, Department of Mathematics, Bharath Institute of Higher Education and Research, Chennai,-73, India.
jeyaseelibrte@gmail.com ishwariyarose@gmail.com

Address for Correspondence

*1 JEYASEELI. S, MPhil Scholar, Department of Mathematics Bharath Institute of Higher Education and Research, Chennai,-73, India.
*2 Dr. R. Ishwariya , Associate Professor, Department of Mathematics, Bharath Institute of Higher Education and Research, Chennai,-73, India.
jeyaseelibrte@gmail.com ishwariyarose@gmail.com

Abstract

Since the introduction of the lasso in regression, various sparse methods have been developed in an unsupervised context like sparse principal component analysis (s-PCA), sparse canonical correlation analysis (s-CCA) and sparse singular value decomposition (s-SVD). These sparse methods combine feature selection and dimension reduction. One advantage of s-PCA is to simplify the interpretation of the (pseudo) principal components since each one is expressed as a linear combination of a small number of variables. The disadvantages lie on the one hand in the difficulty of choosing the number of non-zero coefficients in the absence of a well established criterion and on the other hand in the loss of orthogonality for the components and/or the loadings.

Keywords: sparse methods, correspondence analysis, textual data, sparse correspondence analysis.

1.1 Introduction

When the number of columns (or rows) in a contingency table is very high, as in textual data (rows = documents, columns = terms), the interpretation of CA is difficult. Hence the idea

of a sparse version of CA need to be considered, where the factor scores are sparse, i.e. contain many zero coecients. CA of a contingency table is a double weighted PCA of rows and columns proles with the chi-square distance, or a weighted SVD. CA is also a canonical correlation analysis (CCA) between the two groups of indicator variables of the row and column categories and lastly CA is a particular Multiple Correspondence Analysis (MCA) with p = 2 variables. Therefore applying sparse versions of PCA, SVD, and CCA leads to sparse versions of CA but sparse CA cannot be a particular case of sparse Multiple Correspondence Analysis as it will be shown later. As CA treats symmetrically rows and columns, it is natural to develop what we call a double sparse CA where both rows and columns weights include many zeros. However, when one is not interested in sparsifying both rows and columns, a column sparse only CA comes down to a sparse PCA of the row prole matrix with weighted rows and chi-square metric: only columns loadings are sparse. This could be the case in some documentsterms analysis where one looks for components which are explained each by a small number of words, while one keeps all the documents. Since there are several versions of sparse PCA, CCA, and SVD with dierent properties, a choice should be done. Like in other sparse methods some properties may be lost such as orthogonality and barycentric relations in CA. The rest of the paper is organized as follows. Section 2 gives a reminder on CA. The various sparse versions of PCA, CCA, MCA, SVD as well as general related issues are presented in section 3 and their main properties are compared. The proposed sparse CA method is detailed in section 4 together with a toy example. Applications of sparse CA to textual data are presented in section 5. Conclusion and perspectives for future work are given in section 6.

A detailed presentation can be found in correspondence analysis (MCA). Saporta (2014) propose an historical sketch. One can also refer to the comprehensive book of Beh & Lombardo (2014).

e-ISSN 2320 –7876 www.ijfans.org
Research Paper
Vol.11, Iss.9, Dec 2022

Notations are as follows: bold uppercase letters like X denote matrices, bold lowercase letters like **x** denote vectors. Italic lower case letters denote scalars or elements. In PCA n is the number of observations, p the number of centered variables, **v** a norm 1 eigenvector of principal coefficients or a right singular vector of **X**, **z** a principal component such as $\mathbf{z} = \mathbf{Xv}$, **u** a left singular vector of **X**, **Σ** the variance covariance matrix associated to **X** : $\Sigma = \frac{1}{n}\mathrm{X}'\mathrm{X}$ where superscript ' denotes matrix transposition. In correspondence analysis principal components (rows and columns coordinates) will be denoted **a** and **b**. The operation transforming a vector into a diagonal matrix is denoted diag. The symbol $\propto$ menns proportional to.

2.1 CA standard equations

Given two categorical variables with I and J categories, the contingency table N is first transformed into the frequency table $\mathbf{P} = \mathrm{N}/n$, where n is the total of the $I \times J$ elements of **N**. Let **r** and **c** be the vectors containing respectively the rows and columns totals of P(sum $= 1$), in other terms the marginal probabilities p_i and p_j – D_r and D_c are the associated diagonal matrices $\mathrm{D}_r = \mathrm{diag}\,(\mathbf{r})$ and $\mathrm{D}_c = \mathrm{diag}\,(\mathbf{c})$. $\mathbf{D}_r^{-1}\mathbf{P}$ is the matrix of row profiles (row conditional distributions) and $\mathrm{D}_c^{-1}\mathrm{P}'$ is the transposed matrix of the column profiles.

The row coordinates (principal component or row scores) are the eigenvectors a of $\mathrm{D}_r^{-1}\mathrm{P}\mathrm{D}_c^{-1}\mathrm{P}'$ with sealing $\mathrm{a}'\mathrm{D}_r\mathrm{a} = \lambda$, while the column coordinates are the eigenvectors **b** of $\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}$ with the same eigenvalue $\mathrm{D}_r^{-1}\mathrm{P}\mathrm{D}_c^{-1}\mathrm{P}'\mathrm{a} = \lambda\mathrm{a}$ $\mathrm{D}_c^{-1}\mathrm{P}'\mathrm{D}_r^{-1}\,\mathrm{Pb} = \lambda\mathrm{b}$ Equations (1) and (2) have a trivial extraneous solution $\lambda = 1$ which has to, be discarded. The number of components is $\min(I-1, J-1)$.

Note that equations (1) and (2) are not solved independently, which could lead to undesired reversions of principal axis. The so-called transition formulas link a and b for a given eigenvalue :

e-ISSN 2320 –7876 www.ijfans.org
Research Paper
Vol.11, Iss.9, Dec 2022

$$\mathbf{a} = \frac{1}{\sqrt{\lambda}}\mathbf{D}_r^{-1}\mathbf{P}\mathbf{b} \text{ and } \mathbf{b} = \frac{1}{\sqrt{\lambda}}\mathbf{D}_c^{-1}\mathbf{P}_\mathbf{a}'$$

Vectors a and b contains rows and columns coordinates on a common direction. This allows simultaneous mappings of the $I + J$ categories of both nominal variables, which is one of the major attractions of CA. We use here the so-called symmetric map (Greenacre, 2010). Transition formulas (3) are often interpreted as (pseudo) barycentric properties: the coordinate of a row (a column) is proportional to the weighted mean of the coordinates of all columns (rows), with weights equal to the conditional frequencies of the columns (rows) given that row (column).

3.1 Sparse PCA

Due to its two main properties: maximal variance and orthogonality of the components, PCA is widely and sucoessfully used in many applications. However, when the number of variables is large the interpretation of the principal components may become difficult, since ench component is a linear combination of all variables. When the number of variables is much larger than the number of observations, an other less known problem occurs: inconsistency which means that the sample principal components may not converge towards the population principal components (Hall et al., 2005). Sparse PCA where components would be linear combinations of a small number of variables is thus appealing, but there is no unique definition of sparse PCA. There is a trade-off between sparsity and variance: a large mumber of zeros makes the interpretation easier but leads to a poor approximation, cf later part 3.5.1

Over the past fifteen years, a large number of variants of sparse PCA has been proposed. In their review paper Shen Ning-min & Li Jing (2015) count about twenty algorithms that they divide into 3 classes related to different viewpoints on PCA:

- data-variance-maximization
- minimal-reconstruction-error

- probabilistic modeling viewpoint

The sparse PCA variants are mostly based on the first two classes.

a. Since the principal components are the linear combination of the input variables with maximal variance, a first set of methods consist in imposing sparsity constraints to the maximisation problem : max($\mathbf{v}'\mathbf{\Sigma}\mathbf{v}$).

b. The second viewpoint starts from finding a low rank k approximation of X denoted $\hat{\mathrm{X}}_k$ that minimizes the total squared reconstruction error $\|\mathbf{X} - \hat{\mathbf{X}}_k\|^2$ which is the Frobenius norm of the difference between both matrices. The Eckart-Young theorem states that $\hat{\mathbf{X}}_k$ is obtained by the truncated singular value decomposition (SVD) of X of order :

$$\widehat{\mathbf{X}_k} = \sum_{j=1}^{k} \alpha_j \mathbf{u}_j \mathbf{v}_j'$$

Sparse SVD has been proposed with sparsity constraints on both **u** and v, or v alone.

3.1.1 PCA with additional L1 constraints

a. Using L1 constraints is one of the simplest ways to obtain sparse weights. SC oT L ASS (Joliffe *et* al., 2003) which stands for Simplified Component Technique for Least Absolute Shrinkage and Selection is considered as the first true algorithmie method to achieve sparsity. It consists in adding the extra constraint $\| \mathbf{v} \|_1 \leqslant \tau$ where $\| \mathbf{v} \|_1 = \sum_{j=1}^{p} |v_j|$ to the classical maximization of the variance: max$\mathbf{v}'\mathbf{\Sigma}\mathbf{v}$ with $\| \mathbf{v} \|^2 = \mathbf{v}'\mathbf{v} = 1$. It is necessary that $1 < \tau < \sqrt{p}$ since if $\tau \geqslant \sqrt{p}$ we have usual PCA, and if $\tau = 1$ there is only one non zero weight. SCoTLASS is a non convex and computationally costly algorithm which generally prevents to test many values of the sparsification (or tuning) parameter τ.

b. Zou & Hastie (2006) proposed SPCA, a more efficient algorithm based upon a ridge regression property of PCA: they observe that each PC being a linear combination of the p

variables, its weights can be recovered by a ridge regression of the principal component onto the p variables: a principal component z and the associated weight vector β are the solution of

$$\beta_{\text{ridge}} = \arg\min_{\beta} [\| \mathbf{z} - \mathbf{X}\beta \|^2 + \lambda \| \beta \|^2]$$

Sparse weights are produced by adding a L1 constraint:

$$\hat{\beta} = \arg\min_{\beta} [\| \mathbf{z} - \mathbf{X}\beta \|^2 + \lambda \| \beta \|^2 + \lambda_1 \| \beta \|_1]$$

λ_1 is the sparsity parameter, the larger it is, the more weights are null. The solution in obtained by iterating elastic net and SVD. In their applications, it was found that SPCA account for a larger amount of variance with a much sparser structure than the SCoTLASS.

3.1.2 PR specification of the number of zeros

Prespecifying the number of zeros for the first component it is equivalent to use a L0 norm constraint. Adachi & Trendafilov (2015) proposed a method called unpenalized sparse loading PCA (USLPCA) in which the total number of nonzero loadings (the cardinality of the losding matrix) for a set of k components is pre-specified, without using penalty functions. In their approach, note that they sparsify the loading matrix **A** instead of the weight matrix V. See later part 3.5.4

3. 2 Sparse MCA

In MCA since the columns of X represent categories of nominal variables, Bernard et al., (2012) proposed to globally select the variables, ie blocks of X and not separate columns or categories. Their penalty approach relies

upon a modification of equation (8) similar to the sparse-group lasso of Simon et al. (2013)

$$\hat{\beta}_{GL} = \arg\min_{\beta} \left[\left\| \mathrm{z} - \sum_{j=1}^{J} \mathbf{X}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j} \|\beta_j\|_1 \right]$$

The algorithm consists in alternating sparse-group lasso and SVD until convergence.

Mori et al. (2016) proposed a different method based upon the prespecifying sparsity approach of USLPCA.

It is important to note that one cannot derive a sparse CA from a sparse MCA: CA is equivalent to a MCA with only 2 blocks, and sparse MCA selects entire blocks of indicator variables. Selecting a block, ie discarding one of the 2 categorical variables, does not make sense here.

3.3 Sparse canonical correlation analysis

As noticed in section 2.4, a canonical correlation analysis between indicator matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ corresponding to the variables categories will give the same results as the CA of the contingency matrix N. However, in high dimensional data, when the variables outnumber the sample size or when the variables are highly correlated classical CCA is no longer appropriate. Like in high dimensional PCA, the main drawbacks are the unstability of the estimates and the lack of interpretability of the combinations based on a large number of original variables. In the CA context this may happen in case of large number of categories for one or both variables under study. For example in text mining one variable may represent categories of documents or authors and the other variable columns correspond to the terms of these documents. To overcome these traditional CCA limitations, sparse versions of CCA have been developped by several authors. They all improve the interpretability of canonical variables by restricting the linear combinations to small subsets of variables.

Wilms & Croux (2015) present a short review of the main existing methods before proposing their own sparse CCA method. As for sparse PCA, the sparse CCA methods can be grouped into two main families: one based on SVD and one in the regression framework. Parkhomenko et al. (2009) consider singular value decomposition to derive sparse singular vectors through an efficient iterative algorithm that alternately approximates the left and right singular vectors of the SVD using iterative soft-thresholding for feature selection. Their sparse

CCA (SCCA) method seeks sparsity in both sets of variables simultaneously. It incorporates variable selection and produces linear combinations of small subsets of variables from each group of measurements with maximal correlation.

A similar approach was taken by Witten et al. (2009) who apply a penalized matrix decomposition to the cross-product matrix $\mathbf{X}_1'\mathbf{X}_2$

In the general case, a limitation of these approaches is that they require the variables within each of the two datasets to be uncorrelated to guaranteed sparsity in the canonical vectors. But, in the particular case of correspondence analysis, where the variables are the categories indicators, the previous limitation is naturally satisfied since within each set, the indicator variables are orthogonal and consequently the matrices $\mathbf{X}_1'\mathbf{X}_1$ and $\mathrm{X}_2'\mathrm{X}_2$ are diagonal and regular.

Wilms & Croux (2015) consider the CCA problem from a predictive point of view and reformulate it into a regression framework. They induce sparsity in the canonical vectors by combining an alternating penalized regression approach with a lasso penalty.

3.4 Sparse SVD

The Penalized Matrix Decomposition (PMD) approach by Witten et al.(2009) solves the following optimization problem for the first pair of left and right singular vectors:

$$\text{m}\ \ \mathbf{u}'\mathbf{X}\mathbf{v} \text{ with } \|\mathbf{u}\|^2 = \|\mathbf{v}\|^2 = 1 \text{ and } P_1(\mathbf{u}) \leqslant c_1 P_2(\mathbf{v}) \leqslant c_2$$

where P_1 and P_2 are convex penalty functions such as, eg., the LASSO $P_1(\mathbf{u}) = \sum_{i=1}^{n} |u_i|$ and $P_2(\mathbf{v}) = \sum_{j=1}^{p} |v_j|$ or the fused LASSO constraints with c_1 and c_2 being positive constants. Note that $\max \mathbf{u}'\mathbf{X}\mathbf{v}$ is equivalent to $\min \|\mathbf{X} - \mathbf{u}\mathbf{v}'\|^2$ with the same constraints.

Following Witten et al. $\sum_{i=1}^{n} |u_i|$ and $\sum_{j=1}^{p} |v_j|$ will be denoted sumabsu and sumabsv respectively. sumabsu (resp sumabv) must be between 1 and the square root of the number of rows (resp. columns) of X. The smaller it is, the sparser $\mathbf{u}$ (resp) will be. The pseudo singular

value α is then equal to $\mathbf{u}'\mathbf{X}\mathbf{v}$ When sumabsu is large, ie when there is no penalty on **u**, PMD is equivalent to the sparse PCA of X with the SCoTLASS criterium but with a different algorithm. sPCA-rSVD which stands for sparse PCA via a onesided regularized SVD (Shen & Huang, 2008) is also a special case of PMD since its criterium is $\min_{\mathbf{u},\mathbf{v}}[\|\mathbf{X}-\mathbf{u}\mathbf{v}'\|^2 + P_\lambda(\mathbf{v})]$ with $\| \mathbf{u} \| = 1$ and P_λ a penalty function which is usually the lasso or L1 penalty.

If one looks for a similar degree of sparsity on **u** and **v**, Witten et al. suggest to use a unique parameter denoted sum abs such that: sum abs u $= \sqrt{n}$ sumabs and sumabsv $= \sqrt{p}$ sum abs.

Conclusion:

We have proposed a sparse version of correspondence analysis that highlights the most important categories that determine the underlying axes of dependence between two categorical variables. We believe that this method is particularly well suited to the case where at least one (if not both) variables has a high number of categories as we have illustrated on textual data. The method is flexible and allows different levels of sparsity for rows and columns. A simple and efficient deflation technique $pPMD$ has been proposed. As with other sparse methods, however, the search for solutions is carried out at the cost of losing properties characteristic of correspondence analysis: orthogonality, barycentric relations.

References

[1] Abdi H., B_era M. (2017) Correspondence Analysis. In: Alhajj R., Rokne J. (eds) Encyclopedia of Social Network Analysis and Mining. Springer, New York, NY

[2] Adachi, K., Trenda_lov, N.T. (2015) : Sparse principal component analysis subject to prespecied cardinality of loadings. Computational Statistics 31, 1{25}.

[3] D'Ambra, L., Lauro, N. C. (1992). Non symmetrical exploratory data analysis. Statistic a Applicata, 4(4), 511-529.

[4] B_ecue-Bertaut, M. (2019): Textual Data Science with R, Chapman and Hall/CRC

[5] Hall, P., Marron, J. S., Neeman, A. (2005) Geometric representation of high dimension, low sample size data. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 3, 427{444.

[6] Jolli_e, I.T., Trenda_lov, N.T. and Uddin, M. (2003) A modi_ed principal component technique based on the LASSO. Journal of Computational and Graphical Statistics, 12, 531{547

[7] Laclau C., Nadif, M. (2015) Diagonal Co-clustering Algorithm for Document Word Partitioning. In: Fromont E., De Bie T., van Leeuwen M. (eds)