Sparse PCA for compositional data

Nickolay T. Trendafilov and Michele Gallo

Abstract A great number of procedures for sparse principal component analysis (PCA) were proposed in the last decade. However, they cannot be applied directly for PCA of compositional data (CoDa). We introduce a new procedure for sparse PCA which takes into account the additional constraints specific for CoDa. The proposed method is very effective to find logcontrasts in data, which is illustrated on a real example.

Key words: compositions, log-ratios, steepest descent, Stiefel manifold

1 Introduction

Compositional data (CoDa) consist of vectors of positive elements which sum to one, or in general, to some fixed constant $\kappa > 0$, which is usually 1 or 100. The CoDa are commonly present in a number of experimental fields, and often their high dimensionality requires principal component analysis (PCA) in attempt to find adequate low-dimensional description of the compositional variability (1).

The sample space for a p-dimensional compositional vector v is the simplex defined as $S^p = \{v > 0; v^\top 1_p = \kappa\}$. For this sample space, the internal simplicial operation of perturbation, the external operation of powering, and the simplicial metric define a Hilbert space (4). Thus, compositional problems can be investigated within this space with its specific algebraic-geometric structure.

Nickolay T. Trendafilov (Supported by a grant RPG-2013-211 from The Leverhulme Trust, UK)
Department of Mathematics and Statistics, Open University, United Kingdom, e-mail: Nickolay.Trendafilov@open.ac.uk

Michele Gallo
Department of Human and Social Sciences, University of Naples "L'Orientale", Italy, e-mail: mgallo@unior.it

Unfortunately, working in a simplex turns out to be unpopular. The widely accepted approach is to move from a constrained sample space (simplex) to an unconstrained space, as the Euclidean real space, by log-ratio transformations, opening up all available standard multivariate techniques. Several log-ratio transformations are proposed in the literature among which the centered log-ratio (`clr`) and the isometric log-ratio (`ilr`) transformations seem to be the most widely accepted by statisticians. Applying a log-ratio transformation to the input data for compositional analysis, solves the paradox of subcompositional coherence (2). The property of subcompositional coherence is indeed one of the cornerstones of compositional data analysis: results should be the same for components in a full composition as in any subcomposition, where the subcomposition has been closed again to give κ-sum. An example that is often given of subcompositional incoherence is that the correlation coefficient between two components in a subcomposition is not the same as that for the same two components in the full composition, see (10) and (6, p.141).

In this work, we follow the approach to move from simplex to real space by `clr` transformation and develop sparse PCA applied for compositions. A PCA of `clr` transformed data gives one zero eigenvalue with corresponding eigenvector with all equal elements. "The remaining eigenvalues are positive and, because the corresponding eigenvectors are orthogonal to the final eigenvector, they define contrasts (that is, linear functions whose coefficients sum to zero)" (8, p.347).

During the last decade it became clear that the analysis of high-dimensional data requires methods that produce results involving only a small portion of the original (some times huge) number of variables. The reason for this is twofold: such results are easier to interpret, and most of the involved variables are not really important for the main reasoning behind the data. A great number of procedures for sparse PCA (12) were proposed . However, they cannot be applied directly for PCA of CoDa, because such component loadings have additional constraint to add up to zero.

We introduce a new procedure for sparse PCA which takes into account the additional constraints specific for CoDa. Let R denote a sample covariance/correlation matrix and D – an $r \times r$ diagonal matrix containing its r largest eigenvalues. Consider the following sparse PCA formulation proposed recently by (12):

$$\min_{A \in \mathscr{O}(p,r)} \|A\|_{\ell_1} + \mu \|\mathrm{diag}(A^\top R A) - \mathrm{diag}(D)\|_2^2 \,, \tag{1}$$

where $\mathscr{O}(p,r)$ is the Stiefel manifold of all $p \times r$ orthonormal matrices A, i.e. $A^\top A = I_r$, and diag() is a matrix-to-vector operator taking the main diagonal of the matrix argument. The matrix norm $\|A\|_{\ell_1}$ is defined as usual $\|A\|_{\ell_1} = \mathrm{trace}[A^\top \mathrm{sign}(A)]$. The tuning parameter μ controls the importance of the two terms of the objective function (1). Clearly, the smaller μ the sparser component loadings A.

The sparse PCA defined in (1) can be easily adapted to analyze CoDa. We show that one simply needs to solve (1) on the Stiefel sub-manifold $\mathscr{O}_0(p,r)$ of all $p \times r$ orthonormal matrices A with zero column sums, i.e. $A^\top A = I_r$ and $A^\top 1_{p\times 1} = 0_{r\times 1}$:

$$\min_{A \in \mathscr{O}_0(p,r)} \|A\|_{\ell_1} + \mu \|\mathrm{diag}(A^\top R A) - \mathrm{diag}(D)\|_2^2 \,. \tag{2}$$

2 Steepest Descent Matrix Algorithm

We solve (2) by making use of the projected gradient approach. The standard gradient method is designed for optimization problems without constraints, while the projected gradient method extends the same idea for constrained optimization problems. The key feature of the projected gradient method is that it simultaneously keeps the steepest descent direction and moves on $\mathscr{O}_0(p,r)$, the constrained set of (2). This is achieved by projecting the steepest descent at each step of the algorithm on the constrained manifold $\mathscr{O}_0(p,r)$. For this reason, one needs to have a projection formula (operator) of an arbitrary $p \times r$ matrix Z onto $\mathscr{O}_0(p,r)$ at A, which is to say on its tangent space $\mathscr{T}_A\mathscr{O}_0(p,r)$ at A. The strong point of the projected gradient approach is that it leads to globally convergent algorithms, i.e. converging from any initial $A_0 \in \mathscr{O}_0(p,r)$.

Detailed consideration of the projected gradient approach and the Stiefel submanifold $\mathscr{O}_0(p,r)$ of all $p \times r$ orthonormal matrices with zero column sums is available in (11). Here we briefly revisit some useful results. It can be shown that the tangent space at any $A \in \mathscr{O}_0(p,r)$ is given by

$$\begin{aligned}\mathscr{T}_A\mathscr{O}_0(p,r) &= \{H \in R^{p\times r} |\, H^\top J_p A + A^\top J_p H = 0_r\} \\ &= \{H \in R^{p\times r} |\, A^\top J_p H \text{ is skew-symmetric}\}\ ,\end{aligned} \tag{3}$$

where $J_p = I_p - \frac{1}{p}1_{p\times 1}1_{p\times 1}^\top$ is the $p \times p$ centering matrix. Then, it follows from (3) that each element H of the tangent space $\mathscr{T}_Q\mathscr{O}_0(p,r)$ has the form $H = QK + (J_p - QQ^\top)W = QK + (I_p - QQ^\top)J_pW$ for some skew-symmetric K and arbitrary W. Therefore, it follows from (11, p.448) that the projection of a general $Z \in R^{p\times r}$ onto the tangent space $\mathscr{T}_A\mathscr{O}_0(p,r)$ is defined by

$$\pi_{\mathscr{T}_A\mathscr{O}_0(p,r)}(Z) = A\frac{A^\top Z - Z^\top A}{2} + (J_p - AA^\top)Z\ . \tag{4}$$

After having the projection formula (4) one needs the gradient of the objective function (2) with respect to the Frobenius matrix norm $\|Z\|_F^2 = \text{trace}(Z^\top Z)$. We employ the following smooth approximation $\|A\|_{\ell_1} \approx \text{trace}[A^\top \tanh(\gamma A)]$, for some large $\gamma > 0$. Other smooth approximations for $\|A\|_{\ell_1}$ are also available. Then, the gradient of (2) is given by:

$$\begin{aligned}\nabla_A = {} & \tanh(\gamma A) + (\gamma A) \odot [1_{p\times 1}1_{r\times 1}^\top - \tanh(\gamma A) \odot \tanh(\gamma A)] + \\ & \mu RA\text{Diag}(\text{diag}(A^\top RA) - \text{diag}(D))\ ,\end{aligned} \tag{5}$$

where $\odot$ denotes the Hadamard (element-wise) matrix product and Diag() is a vector-to-matrix operator producing a diagonal matrix from its vector argument.

Following the projected gradient formalism and applying the projection formula (4) we arrive at a matrix ordinary differential equation (ODE) of the form:

$$\frac{dA}{dt} = A\frac{A^\top \nabla_A - \nabla_A^\top A}{2} + (J_p - AA^\top)\nabla_A \ . \tag{6}$$

The ODE (6) defines steepest descent matrix flow on $\mathscr{O}_0(p,r)$ for the objective function (2). Starting with an initial point, we can use this ODE to find the solution of (2). Alternatively, one can use (4) and (6) and employ iterative algorithms for optimization on matrix manifolds which are freely available (9).

3 Numerical example: `Ischia12`

`Ischia12` contains data for the category of hotels preferred by foreign tourists that visit the island of Ischia in 2012. The six parts of compositions (or parts) considered are: one star (1S), two stars (2S), three stars (3S), four stars (4S), five stars (5S) and others (Oth - bed and breakfast, private apartments etc.). We solve (2) by taking R the covariance matrix of the clr-transformed data:

$$R = \begin{pmatrix} 3.9730 & .2838 & -.9194 & -.9789 & -1.4162 & -.9423 \\ .2838 & 1.8167 & -.5479 & -.5164 & -.9983 & -.0379 \\ -.9194 & -.5479 & .5030 & .4172 & .4182 & .1290 \\ -.9789 & -.5164 & .4172 & .4720 & .4244 & .1817 \\ -1.4162 & -.9983 & .4182 & .4244 & 1.8201 & -.2482 \\ -.9423 & -.0379 & .1290 & .1817 & -.2482 & .9176 \end{pmatrix} ,$$

and considering $r = 2$ sparse components such that $D = \mathrm{Diag}(5.5329, 2.3748)$.

The variance $V_o = \mathrm{trace}(D)$ of the first two principal components of R is 7.9077, and the percentage of the explained by them variance is 83.22% ($= 100V_o/\mathrm{trace}(R)$). The sparse loadings A are orthonormal, but the components that they form are *not* uncorrelated as the original principal components. Thus, $V_s = \mathrm{trace}(A^\top RA)$ gives an over-estimation of the variance of the sparse components. To overcome this problem (13) introduced *adjusted* variance which takes into account the correlations among the sparse components. We adapt their definition to our case when the sparse loadings are obtained from covariance matrix. Consider the following Cholesky decomposition $A^\top RA = CC^\top$. Then, the adjusted variance of the sparse components is given by $V_a = \mathrm{trace}(C \odot C)$.

The choice of the tuning parameter μ is of great importance. In sparse PCA, tuning parameters as μ are usually found by cross-validation for large applications (13). Another popular option is employing information criteria. For small applications, as those considered in the paper, the optimal tuning parameter μ can be easily located by solving the problem for several values of μ and compromising between sparseness and fit. Another option is to solve (2) for a range of values of μ and choose the most appropriate of them based on some index of sparseness. Here we use the following one:

$$\mathrm{IS} = \frac{V_a V_s}{V_o^2} \times \frac{\#_0}{pr} \ , \tag{7}$$

which is introduced in (12). V_a, V_s and V_o were defined above, and $\#_0$ is the number of zeros among all pr loadings of A. IS increases with the goodness-of-fit (V_s/V_o), the higher adjusted variance (V_a/V_o) and the sparseness.

We solve (2) for several values of μ and the results are summarized in Table 1. The second and the third columns give the percentage of explained and explained adjusted variances, respectively TV $= 100V_s/\text{trace}(R)$ and TVA $= 100V_a/\text{trace}(R)$.

Table 1 Solutions of (2) for a range of μ values for comparing their merits

μ	TV	TVA	# 0s	IS
10.0	80.5870	79.4583	3	.2312
9.5	80.5615	79.4092	3	.2309
9.0	80.5126	79.3290	3	.2306
8.5	80.5112	79.3270	3	.2306
8.0	80.4960	79.3127	4	.3073
7.5	80.4916	79.3099	4	.3073
7.0	80.4882	79.3079	4	.3073
6.5	80.4798	79.3004	4	.3072
6.0	80.4615	79.2851	4	.3071
5.5	80.4486	79.2759	4	.3070
5.0	80.4196	79.2514	4	.3068
4.5	80.3991	79.2389	4	.3067
4.0	80.3569	79.2087	4	.3064
3.5	80.2549	79.1288	4	.3057
3.0	80.0848	79.0023	4	.3045
2.5	79.6932	78.7040	4	.3019
2.0	78.8883	78.0568	4	.2964
1.5	77.5811	76.9538	4	.2874
1.0	75.5226	75.1768	4	.2733
0.5	77.9744	74.2917	5	.3485
0.0	35.0617	35.0596	8	.1183

The highest IS value is for $\mu = 0.5$. The corresponding sparse loadings are:

$$A_{\mu=0.5} = \begin{pmatrix} .855 & \\ & .707 \\ -.187 & \\ -.261 & \\ -.001 & -.707 \\ -.407 & \end{pmatrix}, A_{\mu=8.0} = \begin{pmatrix} -.883 & -.227 \\ & .797 \\ .190 & \\ .203 & \\ .353 & -.566 \\ .138 & \end{pmatrix}, A_{\mu=0.0} = \begin{pmatrix} -.707 & \\ .707 & \\ & \\ & .707 \\ & -.707 \\ & \end{pmatrix}.$$

For comparison, we reproduce the sparse loadings for $\mu = 8.0$. They are less sparse, but the explained and the adjusted explained variances are quite high and pretty close to the one of the first two principal components. Finally, the most extreme case is when $\mu = 0.0$ and the second term of (2) is switched off. The sparsest possible loadings $A_{\mu=0.0} \in \mathscr{O}_0(6,2)$ depend only on the initial value for solving (6).

To study compositional relationships, ratio and subcompositions one can use graphical tools as principal component plots as in Figure 1 or biplots (3; 7).

Here, we want to highlight the constant logcontrast in $A_{\mu=0.5}$ given as $4.736\log(Oth) + \log(1S) + 5.736\log(3S) = \text{constant}$. Thus, a proportionality relationship between the categories is expressed as:

$$log\left(\frac{1S}{3S}\right) \propto log\left(\frac{3S}{Oth}\right)^{4.736}. \tag{8}$$

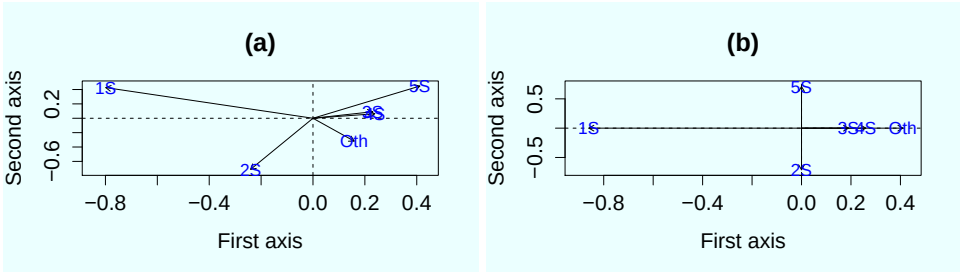

Fig. 1 Plots of principal component loadings (a) and $A_{\mu=0.5}(b)$
.

References

[1] Aitchison, J.: Principal component analysis of compositional data, Biometrika, **70**, 57–65 (1983)

[2] Aitchison, J.: The statistical analysis of compositional data, Chapman & Hall, London - New York (1986)

[3] Aitchison, J., Greenacre, M.: Biplots of compositional data, J Roy Stat Soc C-App, **51**, 375–392 (2002)

[4] Billheimer, D., Guttorp, P., Fagan, W.F.: Statistical interpretation of species composition, J Am Stat Assoc, **456**, 1205–1214 (2001)

[5] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C: Isometric logratio transformations for compositional data analysis, Math Geol, **35**(3), 279–300 (2003)

[6] Egozcue, J.J., Barcelo-Vidal, C., Martín-Fernández, J.A., Jarauta-Bragulat, E., Díaz-Barrero, J.L., Mateu-Figueras, G.: Elements of simplicial linear algebra and geometry, Compositional Data Analysis: Theory and Applications, John Wiley & Sons, Chichester (2011)

[7] Gallo, M., Buccianti, A.: Weighted principal component analysis for compositional data: application example for the water chemistry of the Arno river (Tuscany, central Italy), Environmetrics, **24**(6), 269–277 (2013)

[8] Jolliffe, I.T.: Principal Component Analysis. 2nd edn. Springer-verlag, New York (2002)

[9] Manopt 1.0.5. `http://www.manopt.org`. (2014)

[10] Pawlowsky-Glahn, V., Egozcue, J.J.: Geometric approach to statistical analysis on the simplex, Stoch Env Res Risk A, **15**, 384–398 (2001)

[11] Trendafilov, N.T.: DINDSCAL: direct INDSCAL. Stat. Comput. **22**, 445–454 (2012)

[12] Trendafilov, N.T.: From simple structure to sparse components: a review. Comp. Stat. to appear, (2014), DOI:10.1007/s00180-013-0434-5

[13] Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. J. Comput. Graph. Stat. **15**, 265–286 (2006)