Discriminant Partial Least Square on Compositional Data: a Comparison with the Log-Contrast Principal Component Analysis[1]

Gallo M.,
Department of Social Science, University of Naples – L'Orientale, Italy
mgallo@unior.it

Mahdi S.,
Department of computer Science, mathematics and Physics, University of the West Indies – Barbados
smahdi@uwichill.edu.bb

Abstract: *Discriminant Partial Least Squares for Compositional data (DPLS-CO) was recently proposed by Gallo (2008). The aim of this paper is to show that DPLS-CO is a better dimensionality reduction technique than the LogContrats Principal Component Analysis (LCPCA) for dimensional reduction aimed at discrimination when a compositional training dataset is available.*

Keywords: Compositional observation, Dimension reduction, Linear discrimination**.**

1. Introduction

When discrimination is the goal and dimension reduction is needed Principal Component Analysis (PCA) is the most commonly used technique. Since PCA is only capable of identifying gross variability and is not capable of distinguishing between-groups and within-groups variability, as is the explicit goal of the simple Linear Discriminant Analysis (LDA). LDA has more advantages from a theoretical point of view and it is a better technique to use for discrimination. In Barker and Rayner (2003) it is shown, with a formal statistical explanation, the direct connection between Discriminant Partial Least Squares (DPLS) and LDA. This connection suggests that DPLS to be better dimensionality reduction techniques than PCA for dimensional reduction aimed at discrimination when a training set is available.

In many disciplines, for instance, geology, biochemistry, economics, business administration and sociology, when the discrimination between more groups is the goal, often the observational data are compositional. In such a case, the components of the observational vectors of the data set represent ratios of a fixed total, and the data provide us only with relative and not absolute information. So only ratios between components are meaningful, and those ratios are independent from the arbitrary total. In mathematical terms, a vector $\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_J)$ is compositional if the elements $\tilde{w}_j \geq 0$ (for $j = 1, \ldots, J$) and $\tilde{w}_1 + \ldots + \tilde{w}_J = 1$.

The adequate analysis and relative interpretation of results from this kind of data should be properly incorporated into statistical modeling for discrimination which takes into account the constant-sum constraint that each compositional vector presents. For this purpose, Linear Discriminant Analysis (LDA) for discriminating between two compositions was proposed by Aitchison (1986); more recently, Gallo (2008) proposed the Discriminant Partial Least Squares technique for Compositional data (DPLS-CO), where the connections between DPLS-CO and logcontrast LDA were fully showed.

[1] The present paper is financially supported by the University of Naples – L'Orientale (Department of Social Science 60% fund 2008).

The aim of this paper is to show than DPLS-CO is to be preferred to logcontrast PCA when discrimination between two compositions is the goal and dimension reduction is needed.

2. Compositional data

Let us denote $\mathbf{W}\,(N, J)$ a compositional data matrix. In **W,** all elements are positive and each row is constrained to the unit-sum $\mathbf{W}\hat{\mathbf{1}}_J = \hat{\mathbf{1}}_N$ where $\hat{\mathbf{1}}_J$ and $\hat{\mathbf{1}}_N$ are unit vectors dimension J and *N*, respectively. Let $\mathbf{Q} = \left[N\mathbf{I}_N - \hat{\mathbf{1}}_N\hat{\mathbf{1}}_N^t\right]$ be the product between $1/N$ and the usual centering projector. then $\mathbf{W}^t\mathbf{QW}$ is the covariance matrix of $\mathbf{W}$ refereed to as crude *covariance matrix*.

The unit-sum constraint for each row of $\mathbf{W}$ implies four difficulties: 1) Negative bias, 2) Subcomposition, 3) Basis and 4) Null correlation. Each row and column of $\mathbf{W}^t\mathbf{QW}$ has zero-sum, that is, $\mathbf{W}^t\mathbf{QW}\hat{\mathbf{1}}_J = \hat{\mathbf{0}}_J$ where $\hat{\mathbf{0}}_J$ is a *J* dimensional zero vector. Therefore each variable has a covariance sum equivalent to a negative variance (the first difficulty). No-relationship exists between the crude covariance matrix and the crude subcomposition covariance one. Therefore the variation of subcomposition can substantially influence the covariance (the second difficulty). Likewise in the subcomposition, it is not easy to select a basis for the composition (which is the third difficulty).

Similarly to the *crude covariance matrix*, each row and column of the *crude correlation matrix* of $\mathbf{W}$ has also a zero-sum. Therefore the correlation between two variables is not free to range over the usual full interval $[-1,1]$. The negative bias causes a radical difference from the standard interpretation of correlation between variables. Zero correlation between two ratios does not mean that there is no association (the latter difficulty). Moreover the no interpretability of the crude covariance structure is not the only problem of compositional data. Unfortunately, compositional data often exhibit curvature when standard multivariate methods are employed.

Aitchison (1986) richly described the properties of compositional data and proposed an alternative form of logratio, where the more useful one is based upon a geometric mean. It consists in replacing the natural non-negative condition by the following stronger assumption on strict positive quantities. He proposed to transform each element of $\mathbf{W}$ (w_{ij}) into the logratio $log\left[w_{ij}/g(\mathbf{w})\right]$. This is because the relative matrix of centered logratio $\mathbf{Z}$, with generic element $z_{ij} = log\left[w_{ij}/g(\mathbf{w})\right]$ where $g(\mathbf{w})$ is a geometric mean of a *J* dimensional composition, is adequate for a low-dimensional description of compositional variability. Moreover, a generalization of the logratios – called *logcontrasts* – have particular and researched proprieties in compositional data analysis. Logcontrast of $\mathbf{w}$ is any loglinear combination $\mathbf{c}^t\, log\, \mathbf{w} = c_1\, log\, w_1 + \ldots + c_J\, log\, w_J$ with $c_1 + \ldots + c_J = 0$, where logcontrast with the geometric mean $g(\mathbf{w})$ presents the following property: $\mathbf{c}^t\, log\, \mathbf{w} = \mathbf{c}^t\, log(\mathbf{w}/g(\mathbf{w}))$.

3. DPLS-CO

Modeling the relation between two blocks of variables is a common practice in many fields especially in chemistry and chemical engineering. The objective is usually to extract the information in the descriptor matrix $\mathbf{X}\,(N, J)$ and in the response matrix $\mathbf{Y}\,(N, G)$, where the members of the **Y** can be expressed as a function of the members of the **X**. Traditionally, the modeling of **Y** by means of **X** is done through multiple linear regression, which works well as long as the ***X***-variables are fairly few and weakly uncorrelated. When handling numerous and collinear ***X***-variables as well as in applications involving many variables with few observations, PLS allows us to investigate more complex problems than before, and to also analyze the available data in a more realistic way.

Let $\mathbf{Y}\ (N,G)$ be a "dummy matrix" that expresses the G classes of the training data set with n_g $(g=1,\ldots,G)$ observations on each class. When the goal of analysis is to find a model that separates classes of observations on the basis of the values of their ***X***-variables, it is well known that PLS (maximizing covariance) is related to CCA (maximizing correlation)

$$\mathrm{cov}^2(\mathbf{Xa},\mathbf{Yb}) = \mathrm{var}(\mathbf{Xa})\mathrm{cor}^2(\mathbf{Xa},\mathbf{Yb})\mathrm{var}(\mathbf{Yb}) \tag{3.1}$$

where a principal component analysis in the ***X***-space and a principal component analysis in ***Y***-space provide the penalties. The equation (3.1) can be rewritten as

$$\mathrm{var}(\mathbf{Xa})\mathrm{cor}^2(\mathbf{Xa},\mathbf{Yb}) = \frac{\mathrm{cov}^2(\mathbf{Xa},\mathbf{Yb})}{\mathrm{var}(\mathbf{Yb})}, \tag{3.2}$$

provided that, var(**Yb**) is not zero.

Suppose $\mathbf{A}=[\mathbf{a}_1,\ldots,\mathbf{a}_k]$ is the first k coefficient vector and $\mathbf{Q}=\left[N\mathbf{I}_N-\hat{\mathbf{1}}_N\hat{\mathbf{1}}_N^t\right]$ is the product between $1/N$ and the usual centering projector, then $\mathbf{X}^t\mathbf{QX}$ is the covariance matrix of $\mathbf{X}$. Barker and Rayens (2003) propose to solve

$$\underset{\substack{\mathbf{a}\in\Re^J;\ \mathbf{b}\in\Re^G\\ \mathbf{a}^t\mathbf{X}^t\mathbf{QXA}=\hat{\mathbf{0}}}}{arg\,max}\left\{\frac{\mathrm{cov}^2(\mathbf{Xa},\mathbf{Yb})}{(\mathbf{a}^t\mathbf{a})\mathrm{var}(\mathbf{Yb})}\right\}=\{\mathbf{a}_{k+1},\mathbf{b}_{k+1}\} \tag{3.3}$$

where $\mathbf{a}_{k+1}$ is the eigenvector corresponding to the largest eigenvalue λ_{k+1} of the matrix $(\mathbf{I}_J-\mathbf{P}_k^X)\mathbf{X}^t\mathbf{QY}(\mathbf{Y}^t\mathbf{QY})^{-1}\mathbf{Y}^t\mathbf{QX}$ with $\mathbf{P}_k^X=\left(\mathbf{X}^t\mathbf{QXA}\left(\mathbf{A}^t\mathbf{X}^t\mathbf{QXX}^t\mathbf{QXA}\right)^{-1}\mathbf{A}^t\mathbf{X}^t\mathbf{QX}\right)$, and $\mathbf{b}_{k+1}=(\mathbf{Y}^t\mathbf{QY})^{-1}\mathbf{Y}^t\mathbf{QXa}_{k+1}$. Of course, when classification is the goal, only $\mathbf{a}_{k+1}$ is of interest.

Similarly, the aim of PLS for discrimination on compositional data (DPLS-CO) is to maximize $\mathrm{cov}^2(\mathbf{Za},\mathbf{Yb})/\mathrm{var}(\mathbf{Yb})$ subject to the usual constraints of PLS for discrimination plus the additional constraint that each coefficient vector has zero-sum: $\mathbf{a}^t\hat{\mathbf{1}}_J=0$. Since $\mathbf{a}^t\hat{\mathbf{1}}_J=0$ can be viewed as an orthogonality constraint with respect to a redefined inner product, Gallo (2008) showed that PLS for discrimination solutions always satisfy the additional constraint $\mathbf{a}^t\hat{\mathbf{1}}_J=0$ of compositional data:

$$\underset{\substack{\mathbf{a}\in\Re^J;\ \mathbf{b}\in\Re^G\\ \mathbf{a}^t\mathbf{Z}^t\mathbf{QZA}=\hat{\mathbf{0}}\\ \mathbf{a}^t\hat{\mathbf{1}}_J=0}}{arg\,max}\left\{\frac{\mathrm{cov}^2(\mathbf{Za},\mathbf{Yb})}{(\mathbf{a}^t\mathbf{a})\mathrm{var}(\mathbf{Yb})}\right\}=\underset{\substack{\mathbf{a}\in\Re^J;\ \mathbf{b}\in\Re^G\\ \mathbf{a}^t\mathbf{Z}^t\mathbf{QZA}=\hat{\mathbf{0}}}}{arg\,max}\left\{\frac{\mathrm{cov}^2(\mathbf{Za},\mathbf{Yb})}{(\mathbf{a}^t\mathbf{a})\mathrm{var}(\mathbf{Yb})}\right\}=\{\mathbf{a}_{k+1},\mathbf{b}_{k+1}\} \tag{3.4}$$

Hence, to implement DPLS-CO, one just needs to replace the original compositional data by their corresponding logcontrast ones and employ the PLS discrimination algorithms.

4. DPLS-Co versus LCPCA in discrimination

Each composition $\mathbf{w}$ is a $J-1$ dimensional additive logistic normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ where $\mathbf{w}$ is the one-to-one transformation to $\mathbf{z}$ defined by a $J-1$ dimensional normal distribution with mean vector $\boldsymbol{\lambda}$ and covariance matrix $\mathbf{W}^t\mathbf{QW}$ and density function

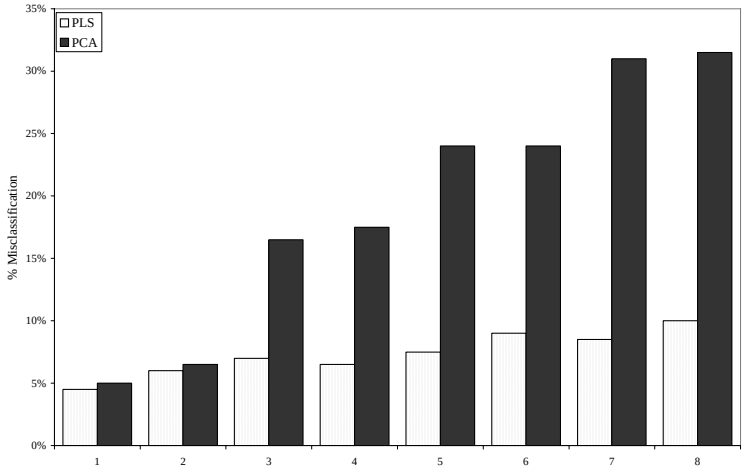
$$\phi(z)=(2\pi)^{-(J-1)/2}(J-1)^{-1/2}(w_1\ldots w_J)^{-1}\left|\mathbf{W}^t\mathbf{QW}-J^{-1}\hat{\mathbf{1}}_J\right|^{-1/2}$$
$$\exp\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\lambda})^t\left|\mathbf{W}^t\mathbf{QW}-J^{-1}\hat{\mathbf{1}}_J\right|^{-1}(\mathbf{z}-\boldsymbol{\lambda})\right] \tag{3.5}$$

with $\hat{\mathbf{1}}_J$ is a J x J matrix of units (Aitchison, 1982). Thus, to illustrate that DPLS-CO is to be preferred to LCPCA, we generated 100 observations on each of two groups, with means (10,10,10) and (15,10,10) for first and second group, respectively. The chosen dispersion matrix is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix},$$

where σ^2 was varied from one to eight. To compare the classification performance of DPLS-CO and LCPCA, the misclassification rate was considered.

For each σ^2 configuration, we have summarized. in Figure 1, the percentage of misclassification obtained with DPLS-CO and LCPCA.



It is has been observed that as σ^2 increases, the misclassification rate for DPLS-CO remains approximately constant while for LCPCA this rate increases noticeable.

5. Summary and conclusion

LCPCA is the best approach when the goal is to identify the maximum gross variability and the data are compositional, but it is not particularly reliable when it comes to distinguish among-group and within-group variability. In this case, LDA is to be preferred to any other methods when it is use is possible; however, often LDA cannot be performed when the number of variables is very large relative to the number of observations. Therefore, dimension reduction is necessary and LCPCA and DPLS-CO is the more appropriate method in such situations.

DPLS-CO will outperform LCPCA when discrimination is the goal and within-groups variability dominates the among-groups variability. This is because DPLS-CO is directly connected to LDA.

Bibliography

Aitchison J. (1982), The Statistical Analysis of Compositional Data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 44 (2), 139–177.

Aitchison J. (1986), *The Statistical Analysis of Compositional Data*. London, UK: Chapman and Hall.

Barker M., Rayens W. (2003), Partial least squares for discrimination. *Journal of Chemometrics*, 17, 166–173.

Barceló-Vidal C., Martín-Fernández J.A., Pawlowsky-Glahn V. (2001), Mathematical foundations of compositional data analysis. In Ross G. ed. *Proceedings of IAMG'01 – The sixth annual conference of the International Association for Mathematical Geology*. Electronic publication.

Gallo M. (2003), Partial Least Squares for Compositional Data: An Approach based on the splines. *Italian Journal of Applied Statistics*, 15, 349–358.

Gallo M (2008), Discriminant Partial Least Squares analysis on compositional data. *Statistical Modeling* (*In Press).*

