Regularized-Generalized PLS-DA

Pietro Amenta
Department of Analysis of Economic and Social Systems,
University of Sannio, Via delle Puglie 3, 82100, Benevento, Italy
amenta@unisannio.it

Abstract: *Linear Discriminant Analysis leads to unstable models and poor predictions in the presence of quasi collinearity among variables or in situations where the number of variables is large with respect to the samples. Partial Least Squares Discriminant Analysis (PLS-DA) was than proposed to overcome the multicollinearity problem and defined as a straightforward extension of the PLS regression. Generalized PLS-DA (GPLS-DA) and "Between" PLS-DA (B-PLS-DA) are two suitable extension of PLS-DA. A simple regularization procedure is proposed to cope with the problems of quasi collinearity or multicollinearity. It is shown that the GPLS-DA and Between PLS-DA are the two end points of a continuum approach.*

Keywords: Discriminant analysis, Multicollinearity, Partial Least Squares, Regularization.

1. Introduction

In the chemometrics, pattern recognition, ecology and in other fields, we very often have to deal with the study of groups and with the research of their separation aiming to fix a decision role. The most known statistical methods of discrimination are Linear Discriminant Analysis (LDA) developed by Fisher (1936) and Quadratic Discriminant Analysis, closely related to LDA, which allows the intraclass covariance matrices to differ between classes. Both these methods are likely to lead to unstable models and poor predictions in the presence of quasi collinearity among variables or in situations where the number of variables is large with respect to the samples (Naes and Indhal, 1998). A number of theoretical approaches have been suggested to deal with these problems. Early proposals on this topic are the contributions of Smidt and McDonald (1976), Wold (1976) and Friedman (1989).

PLS is a routinely used methodology for both classification and discrimination problems with multicollinear data, even if it was not born with these purposes. The usual PLS Discriminant Analysis (PLS-DA) (Sjöström *et al.*, 1986) was proposed to overcome the multicollinearity problem of LDA and it can be defined as a straightforward extension of the PLS regression. Unfortunately, in some situations, a misuse of PLS-DA can lead to a biased solution which does not answer the given problem of discrimination. As pointed out by Barker and Rayens (2003) and Sabatier *et al.* (2003), it is absolutely impossible to interpret PLS-DA with respect to the between groups sum of squares and cross product variance matrix like LDA because PLS-DA corresponds to extracting the unit eigenvector associated to the dominant eigenvalue of an alterated version of this matrix. Barker and Raynes (2003), Nocairi *et al.* (2005) and Sabatier *et al.* (2003) highlighted that a more suitable version of PLS-DA should be based on the ordinary "between" groups sum of squares and cross product variance (B-PLS-DA) according to the maximation of a covariance criterium. Sabatier *et al.* (2003) provided a simple extension of PLS-DA called Generalized PLS-DA which is close to the generalization of PLS proposed by Cazes (1997), but for the context of discrimination; moreover it is based on the eigenanalysis of a matrix equivalent to that of LDA. This approach provides equal results to LDA if the variance and covariance matrix of the predictor variables is not ill-conditioned.

As B-PLS-DA corresponds to a shrinkage of inverse of the the variance and covariance matrix of the predictor variables towards the identity matrix, GPLS-DA (LDA) and B-PLS-DA can be then considered two end points of a continuum approach by incoporating a ridge type of regularization into GPLS-DA. This new regularized formulation of GPLS-DA (LDA) and B-PLS-DA is conceptually easy to understand and implement by handling the multicollinearity problem. We refer to it as “Regularized-Generalized PLS-DA”. With this continuum approach, we will have LDA and B-PLS-DA as two end points, if the variance and covariance matrix of the predictor variables is not ill-conditioned, while, in presence of multicollinearity, the two end points will be GPLS-DA and B-PLS-DA. In the latter case, GPLS-DA can be viewed as ridge version of Discriminant Analysis (Smidt and McDonal, 1976; Rencher, 1998).

2. Notation

The main goal of collecting the values taken by n statistical units on p variables in a matrix $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}^T$ of order $n \times p$ is, generally, the comparison either of the statistical units or the variables. For the former comparison, if we choose to compute a distance between the statistical units, then we have to define a $p \times p$ symmetric positive definite matrix $\mathbf{Q_X}$ defining the scale of the several variables. If we perform the comparison between the variables by a linear correlation coefficient, then we have to define a positive diagonal matrix $\mathbf{D}$ collecting the weights of the statistical units. Then we can consider the notation of the statistical study (triplet) $(\mathbf{X}, \mathbf{Q_X}, \mathbf{D})$ (Escoufier, 1987) to describe the data and their use. The triplet allows to present the factorial methods in a single theoretical framework by using suitable choices for the metrics $\mathbf{Q_X}$ and $\mathbf{D}$ where $\mathbf{D} = diag(d_1, ..., d_n)$ specifies the weights metric in the vectorial space $\Re^n$ of variables with $\sum_{i=1}^{n} d_i = 1$ (for instance $d_i = 1/n$) and $\mathbf{Q_X}$ defines the metric measuring the distance between the data vectors $\mathbf{x}_j$, $\mathbf{x}_k$ of two statistical units *j*, *k* in $\Re^p$ given by $(\mathbf{x}_j - \mathbf{x}_k)^T \mathbf{Q_X} (\mathbf{x}_j - \mathbf{x}_k)$. We assume that $\mathbf{X}$ is mean centred with respect to $\mathbf{D}$ ($\mathbf{1}_n^T \mathbf{DX} = \mathbf{0}$ with $\mathbf{1}_n$ unitary column vector). We highlight that the ordinary Principal Component Analysis on covariance matrix is the analysis of the statistical study $(\mathbf{X}_c, \mathbf{I}_p, \frac{1}{n}\mathbf{I}_n)$ where $\mathbf{X}_c$ is the centred data matrix.

Let $(\mathbf{Y}, \mathbf{Q_Y}, \mathbf{D})$ be the statistical study associated with the matrix Y of order $(n \times q)$, collecting an additional set of q criterion variables observed on the same n statistical units. $\mathbf{Q_Y}$ is the $(q \times q)$ metric of the statistical units in $\Re^q$. If $\mathbf{Y}$ is a qualitative variable with q categories then $\mathbf{U}$ is the $\mathbf{D}$-uncentered binary indicator matrix related to the complete disjunctive coding of this variable. In this case, $\mathbf{D_Y} = \mathbf{U}^T \mathbf{DU}$ is the $(q \times q)$ diagonal matrix of the each category relative frequencies, $\mathbf{G} = \mathbf{D_Y}^{-1} \mathbf{U}^T \mathbf{DX}$ is the $(q \times p)$ matrix collecting the explanatory variables averages for each categories and $\mathbf{B} = \mathbf{G}^T \mathbf{D_Y} \mathbf{G} = \mathbf{X}^T \mathbf{DUD_Y}^{-1} \mathbf{U}^T \mathbf{DX}$ is the between groups variance matrix.

3. A brief summary of Generalized PLS-DA

Generalized PLS Discriminant Analysis is defined (Sabatier *et al.*, 2003) as the PLS analysis of the triplets $(\mathbf{U}, \mathbf{Q_Y}, (1/n) \times \mathbf{I}_n)$ and $(\mathbf{X}, \mathbf{Q_X}, (1/n) \times \mathbf{I}_n)$ with $\mathbf{Q_Y} = \mathbf{D_Y}^{-1}$ and $\mathbf{Q_X} = (\mathbf{X}^T \mathbf{X})^{-1}$. The first GPLS-DA axis w_1 is given by the diagonalization of the matrix $\mathbf{B}(\mathbf{X}^T \mathbf{X})^{-1}$ and it is equivalent to LDA if $\mathbf{X}^T \mathbf{X}$ is not singular. The first PLS-DA axis is instead given by the eigenanalysis of the matrix $\mathbf{X}^T \mathbf{DUU}^T \mathbf{DX}$ obtained by the PLS analysis of the triplets $(\mathbf{U}, \mathbf{I}_q, (1/n) \times \mathbf{I}_n)$ and $(\mathbf{X}, \mathbf{I}_p, (1/n) \times \mathbf{I}_n)$. It is evident that the matrix diagonalized by PLS-DA is an alterated version of matrix $\mathbf{B}$ and so we do not have an optimality condition for PLS-DA when discrimination is the goal. According the Barker and Raynes (2003) and Nocairi *et al*. (2005) suggestions, B-PLS-DA

(Sabatier *et al.*, 2003) is instead given by the PLS analysis of the triplets $(\mathbf{U},\mathbf{D}_{\mathbf{Y}}^{-1},(1/n)\times\mathbf{I}_n)$ and $(\mathbf{X},\mathbf{I}_p,(1/n)\times\mathbf{I}_n)$ where the first axis is obtained by the diagonalization of the the ordinary between groups sum of squares and cross product variance $\mathbf{B}=\mathbf{X}^T\mathbf{DUD}_{\mathbf{Y}}^{-1}\mathbf{U}^T\mathbf{DX}$. We highlight that B-PLS-DA is also equivalent to a Between Principal Component Analysis (B-PCA) that is the PCA of the cloud of centers of gravity (Cailliez and Pages, 1976).

The GPLS-DA components $\mathbf{t}_1=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w}_1$ and $\mathbf{u}_1=\mathbf{YD}_{\mathbf{Y}}^{-1}\mathbf{c}_1$ associated with the first axes are the eigenvectors of $\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{U}}$ and $\mathbf{P}_{\mathbf{U}}\mathbf{P}_{\mathbf{X}}$, respectively, where $\mathbf{P}_{\mathbf{U}}$ and $\mathbf{P}_{\mathbf{X}}$ are the $\mathbf{D}$-orthogonal projection operators onto the vectorial subspaces $Im(\mathbf{U})$ and $Im(\mathbf{X})$. The other GPLS-DA components are based on the residual matrices $\mathbf{X}^{(s)}$ and $\mathbf{Y}^{(s)}$ ($s=2,\ldots,\min(p,q-1)$) by the $\mathbf{D}$-projections of $\mathbf{X}$ and $\mathbf{Y}$ onto the subspaces $\mathbf{T}^{(s)}$ spanned by $\mathbf{T}^{(s)}=[\mathbf{t}_1,\ldots,\mathbf{t}_s]$, respectively, such to maximize at each step the following squared covariance criteria

$$\begin{cases}\max\limits_{\mathbf{w}_s,\mathbf{c}_s}\operatorname{cov}^2\left(\mathbf{X}^{(s)}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w}_s,\mathbf{Y}^{(s)}\mathbf{D}_{\mathbf{Y}}^{-1}\mathbf{c}_s\right)\\ \left\|\mathbf{w}_s\right\|^2_{(\mathbf{X}^T\mathbf{X})^{-1}}=1\\ \left\|\mathbf{c}_s\right\|^2_{\mathbf{D}_{\mathbf{Y}}^{-1}}=1\end{cases}\quad\text{or}\quad\begin{cases}\max\limits_{\mathbf{w}_s}\sum_{j=1}^{q}\operatorname{cov}^2(\tilde{\mathbf{y}}_j,\mathbf{t}_s)\\ \left\|\mathbf{w}_s\right\|^2_{(\mathbf{X}^T\mathbf{X})^{-1}}=1\end{cases}$$

with $\mathbf{t}_s=\mathbf{X}^{(s-1)}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w}_s$, $\mathbf{X}^{(0)}=\mathbf{X}$, $\mathbf{Y}^{(0)}=\mathbf{Y}$ and $\tilde{\mathbf{Y}}=\mathbf{YD}_{\mathbf{Y}}^{-1}$. The GPLS-DA solutions of order s are then given by the eigenvector $\mathbf{w}_s=(\mathbf{X}^T\mathbf{X})^{1/2}\tilde{\mathbf{w}}_s$ where $\tilde{\mathbf{w}}_s$ is linked to the higher eigenvalue λ of the eigen-system $(\mathbf{X}^T\mathbf{X})^{-1/2}\mathbf{X}^{(s-1)T}\mathbf{DYD}_{\mathbf{Y}}^{-1}\mathbf{Y}^T\mathbf{DX}^{(s-1)}(\mathbf{X}^T\mathbf{X})^{-1/2}\tilde{\mathbf{w}}_s=\lambda\tilde{\mathbf{w}}_s$. If $\mathbf{X}^T\mathbf{X}$ is not ill-conditioned then the GPLS-DA and the LDA solutions of order s are equal. Finally, we note that the maximum number of GPLS-DA axes is less or equal to $\min(p,q-1)$ analogously to LDA.

4. Regularized - Generalized PLS-DA

It is evident that the B-PLS-DA matrix can be obtained from that of GPLS-DA by shrinkage of matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ towards the identity matrix. We can consider then a bridge between B-PLS-DA and GPLS-DA adopting a gradual shrinkage of the inverse of the the variance and covariance matrix of the predictor variables towards the identity matrix. GPLS-DA (LDA) and B-PLS-DA can be then considered two end points of a continuum approach by incoporating a ridge type of regularization into GPLS-DA and considering a convex combination of the two matrices.

Regularized-Generalized PLS Discriminant Analysis (R-GPLS-DA) is then defined as the PLS analysis of the triplets $\{\mathbf{U},\mathbf{D}_{\mathbf{Y}}^{-1},(1/n)\times\mathbf{I}_n\}$ and $\{\mathbf{X},\left[(1-\alpha)\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I}_p\right]^{-1},(1/n)\times\mathbf{I}_n\}$ with $\alpha\in[0,1[$. It is possible to show that the R-GPLS-DA axis $\mathbf{w}_s$ of order s maximizes the criteria

$$\begin{cases}\max\limits_{\mathbf{w}_s,\mathbf{c}_s}\operatorname{cov}^2\left(\mathbf{X}^{(s)}\left[(1-\alpha)\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I}_p\right]^{-1}\mathbf{w}_s,\mathbf{Y}^{(s)}\mathbf{D}_{\mathbf{Y}}^{-1}\mathbf{c}_s\right)\\ \left\|\mathbf{w}_s\right\|^2_{\left[(1-\alpha)\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I}_p\right]^{-1}}=1\\ \left\|\mathbf{c}_s\right\|^2_{\mathbf{D}_{\mathbf{Y}}^{-1}}=1\end{cases}\quad\text{or}\quad\begin{cases}\max\limits_{\mathbf{w}_s}\sum_{j=1}^{q}\operatorname{cov}^2(\tilde{\mathbf{y}}_j,\mathbf{t}_s)\\ \left\|\mathbf{w}_s\right\|^2_{\left[(1-\alpha)\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I}_p\right]^{-1}}=1\end{cases}$$

with $\mathbf{t}_s = \mathbf{X}^{(s)}\left[(1-\alpha)\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I}_p\right]^{-1}\mathbf{w}_s$, and where $\mathbf{w}_s$ is given by the diagonalization of the matrix $\mathbf{B}\left[(1-\alpha)\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I}_p\right]^{-1}$. The R-GPLS-DA solution of order s is then given by the eigenvector $\mathbf{w}_s=\hat{\mathbf{Q}}_{\mathbf{X}}^{1/2}\tilde{\mathbf{w}}_s$ where $\tilde{\mathbf{w}}_s$ is linked to the higher eigenvalue λ of the eigen-system $\hat{\mathbf{Q}}_{\mathbf{X}}^{-1/2}\mathbf{X}^{(s-1)T}\mathbf{DYD}_{\mathbf{Y}}^{-1}\mathbf{Y}^T\mathbf{DX}^{(s-1)}\hat{\mathbf{Q}}_{\mathbf{X}}^{-1/2}\tilde{\mathbf{w}}_s=\lambda\tilde{\mathbf{w}}_s$, with $\hat{\mathbf{Q}}_{\mathbf{X}}^{-1/2}=\left[(1-\alpha)\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I}_p\right]^{-1/2}$.

It is worthy noting that the above criteria are also equivalent to the criterion

$$\max\frac{\sum_{j=1}^{q}\operatorname{cov}^2(\tilde{\mathbf{y}}_j,\mathbf{t}_s)}{\|\mathbf{Xw}_s\|^2+k\|\mathbf{w}_s\|^2}\quad\text{with } k=\frac{\alpha}{1-\alpha}$$

which leads to a regressor proportional to ridge regression (Frank and Friedman, 1993; Bougeard *et al.*, 2008) in the univariate case $\mathbf{Y}=[\mathbf{y}]$ and with $\|\mathbf{w}_s\|^2=1$. It is then possible to express R-GPLS-DA as the PLS analysis of the triplets $\{\mathbf{U},\mathbf{D}_{\mathbf{Y}}^{-1},(1/n)\times\mathbf{I}_n\}$ and $\{\mathbf{X},\left[\mathbf{X}^T\mathbf{X}+k\mathbf{I}_p\right]^{-1},(1/n)\times\mathbf{I}_n\}$ with $k\in[0,\infty[$. It is evident that with both formulation of R-GPLS-DA, we obtain GPLS-DA (LDA) and B-PLS-DA as two end points of a continuum approach according to $\alpha\in[0,1[$ or $k\in[0,\infty[$. In presence of multicollinearity, R-GPLS-DA allows then to user to choose among a whole range of methods from which a suitable model can be selected by using a validation procedure.

Bibliography

Barker M., Rayens W. (2003), PLS for discrimination, Journal of Chemometrics, 17: 166-173.

Bougeard S., Hanafi M., Qannari E.M., (2008), Continuum redundancy–PLS regression: A simple continuum approach, Computational Statistics and Data Analysis Volume: 52, 7: 3686-3696.

Cailliez F., Pages J.P. (1975), Introduction à l'Analyse des Donneés, Smash Paris.

Cazes P. (1997), Adaptation de la régression PLS au cas de la régression aprés analyse des correspondances multiples, Revue de Statistique Appliquées, XLV(2): 89-99.

Escoufier Y. (1987), The duality diagram: a means of better practical applications, in: Legendre P, Legendre L. (Eds.) Development in numerical ecology. NATO advanced Institute, Springer Verlag, Berlin.

Fisher R.A. (1936), The use of multiple measurements in taxonomic problems, Annals of Eugenics 7: 179-188.

Frank I.E., Friedman, J.H. (1993), A statistical view of some chemometrics regression tools. Technometrics, 35: 109–135.

Friedman J.H. (1989), Regularized Discriminant Analysis, Journal of the American Statistical Association, 84: 165-175.

Naes T., Indahl U. (1998) A unified description of classical classification methods for multicollinear data, Journal of Chemometrics, 12: 205– 220.

Nocairi H., Qannari M. El, Vigneau E., Bertrand D. (2005), Discrimination on latent components with respect to patterns. Application to multicollinear data, Computational Statistics and Data Analysis, 48, 1: 139-147.

Sabatier R., Vivien M., Amenta P. (2003), Two approaches for Discriminant Partial Least Squares, in: Between Data Science and Everyday Web Practice, Shader, M., Gaul, W. & Vichi M. (Eds.), Springer-Verlag: 100-108.

Sjöström M., Wold S., Söderström B. (1986), PLS discriminant plots, in: Pattern Recognition in Practice II, Gelsema E.S., Kanal L.N. (Eds.), Elsevier, Amsterdam.

Smidt R.K., McDonald L.L. (1976), Ridge discriminant analysis. Research Paper No. 108, University of Wyoming, Department of Statistics.

Wold S. (1976), Pattern recognition by means of disjoint principal components models, Pattern Recognition, 8, 3: 127-139.