LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK
SONDERFORSCHUNGSBEREICH 386

Boulesteix:

A note on between-group PCA

Sonderforschungsbereich 386, Paper 397 (2004)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

A note on between-group PCA

Anne-Laure Boulesteix

September 24, 2004

Department of Statistics, University of Munich

Akademiestr.1, D-80799 Munich (Germany)

email: boulesteix@stat.uni-muenchen.de

Keywords: Classification, dimension reduction, feature extraction, linear discriminant analysis, partial least squares, principal component analysis.

Abstract

In the context of binary classification with continuous predictors, we proove two properties concerning the connections between Partial Least Squares (PLS) dimension reduction and between-group PCA, and between linear discriminant analysis and between-group PCA. Such methods are of great interest for the analysis of high-dimensional data with continuous predictors, such as microarray gene expression data.

1 Introduction

Classification, i.e. prediction of a categorical variable using predictor variables is an important field of applied statistics. Suppose we have a $p \times 1$ random vector $\mathbf{x} = (X_1, \ldots, X_p)^T$, where $X_1, \ldots, X_p$ are continuous predictor variables. Y is a categorical response variable and can take values $1, \ldots, K$ ($K \geq 2$). It can also be denoted as group membership. Many dimension reduction and classification methods are based on linear transformations of the random vector $\mathbf{x}$ of the type

$$Z = \mathbf{a}^T \mathbf{x}, \tag{1}$$

where $\mathbf{a}$ is a $p \times 1$ vector. In linear dimension reduction, the focus is on the linear transformations themselves, whereas linear classification methods aim to predict the response variable Y via linear transformations of $\mathbf{x}$. However, both approaches are strongly connected, since the linear transformations which are output by dimension reduction methods can sometimes be used as new predictor variables for classification. In this short note, we study the connection between some well-known dimension reduction and classification methods.

Principal component analysis (PCA) consists to find uncorrelated linear transformations of the random vector $\mathbf{x}$ which have high variance. The same analysis can be performed on the variable $E(\mathbf{x}|Y)$ instead of $\mathbf{x}$. In this paper, this approach is denoted as between-group PCA and examined in section 2. An alternative approach for linear dimension reduction is Partial Least Squares (PLS), which aims to find linear transformations which have high covariance with the response Y. In section 3, the PLS approach is briefly presented and a connection between between-group PCA and the first PLS component is shown for the case $K = 2$.

If one assumes that $\mathbf{x}$ has a multivariate normal distribution within each group and that

the within-group covariance matrix is the same for all the groups, decision theory tells us that the optimal decision function is a linear transformation of $\mathbf{x}$. This approach is called linear discriminant analysis. An overview of discriminant analysis can be found in Hastie et al. (2001). For $K = 2$, we show in section 4 that under a stronger assumption, the linear transformation of $\mathbf{x}$ obtained in linear discriminant analysis is the same as in between-group PCA.

In the whole paper, $\boldsymbol{\mu}$ denotes the mean of the random vector $\mathbf{x}$ and $\boldsymbol{\Sigma}$ its covariance. For $k = 1, \ldots, K$, $\boldsymbol{\mu}_k$ denotes the within-group mean vector of $\mathbf{x}$ and $\boldsymbol{\Sigma}_k$ the within-group covariance matrix for group k. In addition, we assume $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$, $\forall i \neq j$. $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ for $i = 1, \ldots, n$ denote independent identically distributed realizations of the random vector $\mathbf{x}$ and Y_i denotes the group membership of the ith realization. $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_K$ are the sample within-group mean vectors calculated from the data set $(\mathbf{x}_i, Y_i)_{i=1,\ldots,n}$.

2 Between-group PCA

2.1 Definition

Linear dimension reduction consists to define new random variables $Z_1, \ldots, Z_m$ as linear combinations of $X_1, \ldots, X_p$, where m is the number of new variables. For $j = 1, \ldots, m$, Z_j has the form

$$Z_j = \mathbf{a}_j^T \mathbf{x},$$

where $\mathbf{a}_j$ is a $p \times 1$ vector. In Principal Component Analysis (PCA), $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^p$ are defined successively as follows.

Definition 1 . Principal Components.

$\mathbf{a}_1$ *is the* $p \times 1$ *vector maximizing* $VAR(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\mathbf{\Sigma}\mathbf{a}$ *under the constraint* $\mathbf{a}_1^T\mathbf{a}_1 = 1$. *For* $j = 2, \ldots, m$, $\mathbf{a}_j$ *is the* $p \times 1$ *vector maximizing* $VAR(\mathbf{a}^T\mathbf{x})$ *under the constraints* $\mathbf{a}_j^T\mathbf{a}_j = 1$ *and* $\mathbf{a}_j^T\mathbf{a}_i = 0$ *for* $i = 1, \ldots, j-1$.

The vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ defined in definition 1 are the (normalized) eigenvectors of the matrix $\mathbf{\Sigma}$. The number of eigenvectors with strictly positive eigenvalues equals $rank(\mathbf{\Sigma})$, which is $p-1$ if $X_1, \ldots, X_p$ are linearly independent. $\mathbf{a}_1$ is the eigenvector of $\mathbf{\Sigma}$ with the greatest eigenvalue, $\mathbf{a}_2$ is the eigenvector of $\mathbf{\Sigma}$ with the second greatest eigenvalue, and so on. For an extensive overview of PCA, see e.g. Jolliffe (1986).

In PCA, the new variables $Z_1, \ldots, Z_m$ are built indepently of Y and the number of new variables m is at most $p-1$. If one wants to build new variables which contain information on the categorical response variable Y, an alternative to PCA is to look for linear combinations of $\mathbf{x}$ which maximize $VAR(E(\mathbf{a}^T\mathbf{x}|Y))$ instead of $VAR(\mathbf{a}^T\mathbf{x})$. In the following, this approach is denoted as between-group PCA. $\mathbf{\Sigma}_B$ denotes the between-group covariance matrix:

$$\mathbf{\Sigma}_B = COV(E(\mathbf{x}|Y)). \tag{2}$$

In between-group PCA, $\mathbf{a}_1, \ldots, \mathbf{a}_m$ are defined as follows.

Definition 2 . **Between-group Principal Components**.

$\mathbf{a}_1$ *is the* $p \times 1$ *vector maximizing* $VAR(E(\mathbf{a}^T\mathbf{x}|Y)) = \mathbf{a}^T\mathbf{\Sigma}_B\mathbf{a}$ *under the constraint* $\mathbf{a}_1^T\mathbf{a}_1 = 1$. *For* $j = 2, \ldots, m$, $\mathbf{a}_j$ *is the* $p \times 1$ *vector maximizing* $VAR(\mathbf{a}^T\mathbf{x}|Y)$ *under the constraints* $\mathbf{a}_j^T\mathbf{a}_j = 1$ *and* $\mathbf{a}_j^T\mathbf{a}_i = 0$ *for* $i = 1, \ldots, j-1$.

The vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ defined in definition 2 are the eigenvectors of the matrix $\mathbf{\Sigma}_B$. Since $\mathbf{\Sigma}_B$ is of rank at most $K-1$, there are at most $K-1$ eigenvectors with strictly positive eigenval-

ues. Since $E(\mathbf{a}^T\mathbf{x}|Y) = \mathbf{a}^T E(\mathbf{x}|Y)$, between-group PCA can be seen as PCA performed on the random vector $E(\mathbf{x}|Y)$ instead of $\mathbf{x}$. In the next section, the special case $K = 2$ is examined.

2.2 A special case: $K = 2$

If $K = 2$, $\mathbf{\Sigma}_B$ has only one eigenvector with strictly positive eigenvalue. This eigenvector is denoted as $\mathbf{a}_B$. $\mathbf{a}_B$ can be derived from simple computations on $\mathbf{\Sigma}_B$.

$$\begin{aligned}
\mathbf{\Sigma}_B &= p_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu})(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^T + p_2(\boldsymbol{\mu}_2 - \boldsymbol{\mu})(\boldsymbol{\mu}_2 - \boldsymbol{\mu})^T \\
&= p_1(\boldsymbol{\mu}_1 - p_1\boldsymbol{\mu}_1 - p_2\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - p_1\boldsymbol{\mu}_1 - p_2\boldsymbol{\mu}_2)^T \\
&\quad + p_2(\boldsymbol{\mu}_2 - p_1\boldsymbol{\mu}_1 - p_2\boldsymbol{\mu}_2)(\boldsymbol{\mu}_2 - p_1\boldsymbol{\mu}_1 - p_2\boldsymbol{\mu}_2)^T \\
&= p_1 p_2^2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T + p_2 p_1^2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \\
&= p_1 p_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \\
\mathbf{\Sigma}_B(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) &= p_1 p_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).
\end{aligned}$$

Since

$$p_1 p_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0, \tag{3}$$

$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is an eigenvector of $\mathbf{\Sigma}_B$ with strictly positive eigenvalue. Since $\mathbf{a}_B$ has to satisfy $\mathbf{a}_B^T\mathbf{a}_B = 1$, we obtain

$$\mathbf{a}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||. \tag{4}$$

In practice, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are often unknown and must be estimated from the available data set $(\mathbf{x}_i, Y_i)_{i=1,\ldots,n}$. $\mathbf{a}_B$ may be estimated by replacing $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ by $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ in equation (4):

$$\hat{\mathbf{a}}_B = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)/||\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2||. \tag{5}$$

Between-group PCA is applied by Culhane et al. (2002) in the context of high-dimensional microarray data. However, Culhane et al. (2002) formulate the method as a data matrix decom-

position (singular value decomposition) and do not define the between-group principal components theoretically. In the following section, we examine the connection between-group PCA and Partial Least Squares.

3 A connection between PLS dimension reduction and between-group PCA

3.1 Introduction to PLS dimension reduction

Partial Least Squares (PLS) dimension reduction is another linear dimension reduction method. It is especially appropriate to construct new components which are linked to the response variable Y. Studies of the PLS approach from the point of view of statisticians can be found in e.g. Stone & Brooks (1990); Frank & Friedman (1993); Garthwaite (1994). In the PLS framework, $Z_1, \ldots, Z_m$ are not random variables which are theoretically defined and then estimated from a data set: their definition is based on a specific data set. Here, we focus on the binary case ($Y = 1, 2$), although the PLS approach can be generalized to multicategorical response variables (de Jong, 1993). For the data set $(\mathbf{x}_i, Y_i)_{i=1,\ldots,n}$, the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ are defined as follows (Stone & Brooks, 1990).

Definition 3 . PLS components

Let $\hat{COV}$ denote the sample covariance computed from $(\mathbf{x}_i, Y_i)_{i=1,\ldots,n}$. $\mathbf{a}_1$ is the $p \times 1$ vector maximizing $\hat{COV}(\mathbf{a}_1^T\mathbf{x}, Y)$ under the constraint $\mathbf{a}_1^T\mathbf{a}_1 = 1$. For $j = 2, \ldots, m$, $\mathbf{a}_j$ is the $p \times 1$ vector maximizing $\hat{COV}(\mathbf{a}^T\mathbf{x}, Y)$ under the constraints $\mathbf{a}_j^T\mathbf{a}_j = 1$ and $\hat{COV}(\mathbf{a}_j^T\mathbf{x}, \mathbf{a}_i^T\mathbf{x}) = 0$ for $i = 1, \ldots, j-1$.

In the following, the vector $\mathbf{a}_1$ defined in definition 3 is denoted as $\mathbf{a}_{PLS}$. An exact algorithm to compute the PLS components can be found in Martens & Naes (1989). Here, we study the connection between the first PLS component and the first between-group principal component.

Proposition 1 .

For a given data set $(\mathbf{x}_i, Y_i)_{i=1,\ldots,n}$, the first PLS component equals the first between-group principal component:

$$\mathbf{a}_{PLS} = \hat{\mathbf{a}}_B.$$

Proof. For all $\mathbf{a} \in \mathbb{R}^p$,

$$\begin{aligned}
\hat{COV}(\mathbf{a}^T\mathbf{x}, Y) &= \mathbf{a}^T\hat{COV}(\mathbf{x}, Y) \\
\hat{COV}(\mathbf{x}, Y) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbf{x}_i Y_i - \tfrac{1}{n^2}(\sum_{i=1}^n \mathbf{x}_i)(\sum_{i=1}^n Y_i) \\
&= \tfrac{1}{n}(n_1\hat{\boldsymbol{\mu}}_1 + 2n_2\hat{\boldsymbol{\mu}}_2) - \tfrac{1}{n^2}(n_1\hat{\boldsymbol{\mu}}_1 + n_2\hat{\boldsymbol{\mu}}_2)(n_1 + 2n_2) \\
&= \tfrac{1}{n^2}(nn_1\hat{\boldsymbol{\mu}}_1 + 2nn_2\hat{\boldsymbol{\mu}}_2 - n_1^2\hat{\boldsymbol{\mu}}_1 - 2n_1n_2\hat{\boldsymbol{\mu}}_1 - n_1n_2\hat{\boldsymbol{\mu}}_2 - 2n_2^2\hat{\boldsymbol{\mu}}_2) \\
&= n_1n_2(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)/n^2
\end{aligned}$$

The only unit vector maximizing $n_1n_2\mathbf{a}^T(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)/n^2$ is

$$\begin{aligned}
\mathbf{a}_{PLS} &= (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)/||\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1|| \\
&= \hat{\mathbf{a}}_B
\end{aligned}$$

□

Thus, the first component obtained by PLS dimension reduction is the same as the first component obtained by between-group PCA. This is an argument to support the (controversal) use of PLS dimension reduction in the context of binary classification. The connection between between-group PCA and linear discriminant analysis is examined in the next section.

4 A connection between LDA and between-group PCA

4.1 Linear discriminant analysis

In this section, linear discriminant analysis is briefly introduced. The connection to between-group PCA is examined in section 4.2.

If $\mathbf{x}$ is assumed to have a multivariate normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ within class k,

$$\begin{aligned} P(Y=k|\mathbf{x}) &= p_k \cdot f(\mathbf{x}|Y=k)/f(\mathbf{x}) \\ \ln P(Y=k|\mathbf{x}) &= \ln p_k - \ln f(\mathbf{x}) - \ln(\sqrt{2\pi}|\boldsymbol{\Sigma}_k|^{1/2}) - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_k). \end{aligned}$$

The Bayes classification rule predicts the class of an observation $\mathbf{x}_0$ as

$$\begin{aligned} C(\mathbf{x}_0) &= \arg\max_k P(Y=k|\mathbf{x}) \\ &= \arg\max_k(\ln p_k - \ln(\sqrt{2\pi}|\boldsymbol{\Sigma}_k|^{1/2}) - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_k)). \end{aligned}$$

For $K=2$, the discriminant function d_{12} is

$$\begin{aligned} d_{12}(\mathbf{x}) &= \ln P(Y=1|\mathbf{x}) - \ln P(Y=2|\mathbf{x}) \\ &= -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_1) + (\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_2) \\ &\quad + \ln p_1 - \ln p_2 - \ln(\sqrt{2\pi}|\boldsymbol{\Sigma}_1|^{1/2}) + \ln(\sqrt{2\pi}|\boldsymbol{\Sigma}_2|^{1/2}) \end{aligned}$$

If one assumes $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, d_{12} is a linear function of $\mathbf{x}$ (hence the term linear discriminant analysis):

$$\begin{aligned} d_{12}(\mathbf{x}) &= (\mathbf{x} - \tfrac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})^T\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \ln p_1 - \ln p_2 \\ &= \mathbf{a}_{LDA}^T\mathbf{x} + b, \end{aligned}$$

where

$$\mathbf{a}_{LDA} = \mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{6}$$

and

$$b = -\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln p_1 - \ln p_2. \tag{7}$$

4.2 A property

Proposition 2 .

If $\mathbf{\Sigma}$ is assumed to be of the form $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$, where $\mathbf{I}_p$ is the identity matrix of dimensions $p \times p$ and σ is a scalar, $\mathbf{a}_{LDA}$ and $\mathbf{a}_B$ are collinear.

Proof. The proof follows from equations (4) and (6). □

Thus, we showed the strong connection between linear discriminant analysis and between-group PCA in the case $K = 2$. In practice, $\mathbf{a}_B$ is estimated by $\hat{\mathbf{a}}_B$ and $\mathbf{a}_{LDA}$ is estimated by $\hat{\mathbf{a}}_{LDA} = 2(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)/\hat{\sigma}$, where $\hat{\sigma}$ is an estimator of σ. Thus, $\hat{a}_B$ and $\hat{a}_{LDA}$ are also collinear.

The assumption about the structure of $\mathbf{\Sigma}$ is quite strong. However, such an assumption can be wise in practice when the available data set contains a large number of variables p and a small number of observations n. If $p > n$, which often occurs in practice (for instance in microarray data analysis), $\hat{\mathbf{\Sigma}}$ can not be inverted, since it has rank at most $n-1$ and dimensions $p \times p$. In this case, it is sensible to make strong assumptions on $\mathbf{\Sigma}$. Proposition 2 tells us that between-group PCA takes only between-group correlations into account, not within-group correlations.

5 Discussion

We showed the strong connection between PLS dimension reduction for classification, between-group PCA and linear discriminant analysis for the case $K = 2$. PCA and PLS are useful techniques in practice, especially when the number of observations n is smaller than the number of variables p, for instance in the context of microarray data analysis (Nguyen & Rocke, 2002). The connection between PLS and between-group PCA can also justify the use of PLS dimension reduction in the classification framework. In future work, one could examine the connection between the three approaches for multicategorical response variables.

References

CULHANE, A. C., PERRIERE, G., CONSIDINE, E., GOTTER, T. & HIGGINS, D. (2002). Between-group analysis of microarray data. *Bioinformatics* **18**, 1600–1608.

DE JONG, S. (1993). Simpls. an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18**, 251–253.

FRANK, I. E. & FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–135.

GARTHWAITE, P. H. (1994). An interpretation of partial least squares. *J.Amer.Stat.Assoc.* **89**, 122–127.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. H. (2001). *The elements of statistical learning*. New York: Springer-Verlag.

JOLLIFFE, I. T. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

MARTENS, H. & NAES, T. (1989). *Multivariate Calibration*. New York: Wiley.

NGUYEN, D. & ROCKE, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**, 39–50.

STONE, M. & BROOKS, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *J.R.Statist.Soc.B* **52**, 237–269.