# Reduction of Dimensionality for Classification

Carlos Cuevas-Covarrubias and Eva Riccomagno

**Abstract** We present an algorithm for the reduction of dimensionality useful in statistical classification problems where observations from two multivariate normal distributions are discriminated. It is based on Principal Components Analysis and consists of a simultaneous diagonalization of two covariance matrices. The criterion for reduction of dimensionality is given by the contribution of each principal component to the area under the ROC curve of a discriminant function. Linear and quadratic scores are considered, the focus being on the quadratic case.

## 1 Introduction

Principal components analysis (PCA) and linear discriminant analysis (LDA) are very important methods of multivariate statistics. Given a $p$-dimensional random variable $\mathbf{X} : \Omega \to \mathbb{R}^p$, PCA defines its optimal representation in a lower-dimensional subspace; this representation is usually assessed in terms of

Carlos Cuevas-Covarrubias
CIEMA, Facultad de Ciencias Actuariales, Universidad Anáhuac, México,
✉ ccuevas@anahuac.mx

Eva Riccomagno
Dipartimento di Matematica, Università degli Studi di Genova, Italy,
✉ riccomagno@dima.unige.it

ARCHIVES OF DATA SCIENCE, SERIES A
(ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 2, No. 1, 2017

DOI 10.5445/KSP/1000058749/25
ISSN 2363-9881

a percentage of total variation expressed as a function of the eigenvalues of the covariance matrix (Mardia et al, 1979). LDA assumes that $\Omega$, the sample space of $\mathbf{X}$, is partitioned into two different categories: $\Omega_0$ and $\Omega_1$. Given $\mathbf{x}$, a particular realization of $\mathbf{X}$, LDA is used to infer whether $\mathbf{x}$ corresponds to an observation coming from $\Omega_0$ or $\Omega_1$ (Mardia et al, 1979). The ROC[1] curve is one of many criteria to assess the quality of this classification procedure (Bamber, 1975). Although PCA and LDA appear in many standard textbooks of multivariate statistics, these methods are usually discussed independently. In this article, we assume that both classes of conditional distributions of $\mathbf{X}$ are multivariate normal. We explore an original combination of PCA and LDA where the area under the ROC curve appears as the link between both methods. The objective is to reduce dimensionality while preserving as much separability as possible. Our proposal is interesting for several reasons: It gives PCA a wider context of application, and it also helps to make LDA a more explanatory technique. Some of the ideas introduced in this article were previously discussed in Cuevas-Covarrubias (2003) and in Cuevas-Covarrubias (2013). However, these previous works mainly concentrates in the geometrical aspects of this method. Now, our discussion is different, it is presented from a pattern recognition perspective. The geometrical aspects are not analyzed and heteroscedasticity as an important source of information is considered. The main contribution of this article is given in Sects. 5 and 6 where a new approach for PCA is presented. Our proposal is somehow similar to the one in Chang (1983); both methods use a principal components transformation to discriminate two multivariate normal densities. However, contrasting with Chang's method, our proposal does not assume a common covariance matrix in both distributions, and it concentrates on the case where both covariance matrices are different. Our method diagonalizes both matrices simultaneously, and it always results in a set of mutually independent components.

## 2 Classification rules and ROC curves

Consider a random variable $\mathbf{X} : \Omega \to \mathbb{R}^p$ defined on a sample space $\Omega$ which is partitioned into two categories: $\Omega_0$ and $\Omega_1$. Given $x = X(\omega)$, the objective is to infer whether $\omega \in \Omega_0$ or $\omega \in \Omega_0$. Discriminant Analysis is aimed to summarize our vector of covariates $\mathbf{X}$ into a univariate risk score $S : \mathbb{R}^p \to \mathbb{R}$, useful to

[1] Receiver Operating Characteristic

discriminate between $\Omega_0$ and $\Omega_1$. Discrimination takes place according to the following decision rule:

$$\text{Classify in } \begin{cases} \Omega_1 \; if \;\; S > t \\ \Omega_0 \; if \; S \leq t, \end{cases} \tag{1}$$

where $t$ is an arbitrary decision threshold.

**Definition 1.** Given a risk score $S$ with class conditional distribution functions

$$F_0(t) = \Pr[S \leq t|\Omega_0] \text{ and } F_1(t) = \Pr[S \leq t|\Omega_1],$$

its *ROC* curve is the following set:

$$ROC = \{(x,y)|x = 1 - F_0(t), y = 1 - F_1(t), -\infty < t < \infty\}. \tag{2}$$

The ROC curve as in Definition 1 is a graphical summary of the global performance of the score. The area covered by the ROC-curve (denoted as $A$) is a measure of the quality of $S$ (Bamber, 1975). $A = \Pr[S_0 \leq S_1]$; where $S_0 \sim F_0$ and $S_1 \sim F_1$ are two independent random variables. Thus, if we randomly select one $\omega_0 \in \Omega_0$, one $\omega_1 \in \Omega_1$, and we evaluate $S$ on each of these, $A$ is the probability that both observations are ordered according to Equation 1. The closer $A$ is to 1, the better the the performance of $S$ (Bamber, 1975). ROC curves are invariant to monotonous transformations of the score, this property is the basis of the following definition:

**Definition 2.** Let $S$ be a single variable used to discriminate between $\Omega_0$ and $\Omega_1$. If there is a monotonous transformation $T(S)$ of $S$ such that its class conditional distributions are both normal, the *ROC* curve of $S$ (and of $T(S)$) is given by

$$ROC = \{(u,v)|u = 1 - \Phi(t), v = 1 - \Phi\Big(\frac{t-d}{r}\Big), -\infty < t < \infty\}, \tag{3}$$

where $\Phi$ denotes a standard normal distribution function, $d = \frac{E(T(S)|\Omega_1) - E(T(S)|\Omega_0)}{\sqrt{Var(T(S)|\Omega_0)}}$ and $r^2 = \frac{Var(T(S)|\Omega_1)}{Var(T(S)|\Omega_0)}$. Any variable $S$ with this property is said to be a binormal score with parameters $(d, r^2)$. The area under the ROC curve of such a score is:

$$A = \Phi\Big(\frac{d}{\sqrt{1+r^2}}\Big) \tag{4}$$

Table 1 shows the numerical value of $A$ for a binormal model with different values of $d$ and $r^2$ (see Equation 4).

**Table 1** Area under the ROC curve of a binormal score (Equation 4)

| $d$ | $r^2 = 1.01$ | $r^2 = 1.25$ | $r^2 = 1.50$ | $r^2 = 1.75$ | $r^2 = 2.00$ | $r^2 = 4.00$ |
|---|---|---|---|---|---|---|
| 0.00 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| 0.50 | 0.63 | 0.63 | 0.62 | 0.61 | 0.61 | 0.59 |
| 1.00 | 0.75 | 0.74 | 0.73 | 0.72 | 0.71 | 0.67 |
| 1.50 | 0.85 | 0.84 | 0.82 | 0.81 | 0.80 | 0.75 |
| 2.00 | 0.92 | 0.90 | 0.89 | 0.88 | 0.87 | 0.81 |

**Table 2** Numerical Integration of the ROC curve of $Q$

| $d$ | $r^2 = 1.01$ | $r^2 = 1.25$ | $r^2 = 1.50$ | $r^2 = 1.75$ | $r^2 = 2.00$ | $r^2 = 4.00$ |
|---|---|---|---|---|---|---|
| 0.00 | 0.50 | **0.53** | **0.56** | **0.58** | **0.60** | 0.70 |
| 0.50 | 0.63 | 0.63 | 0.62 | **0.63** | **0.64** | 0.71 |
| 1.00 | 0.75 | 0.74 | 0.73 | 0.72 | **0.72** | 0.74 |
| 1.50 | 0.85 | 0.84 | 0.82 | 0.81 | 0.80 | 0.78 |
| 2.00 | 0.92 | 0.90 | 0.89 | 0.88 | 0.87 | 0.83 |

## 3 Discrimination under normality

### *3.1 The univariate case*

Let $X : \Omega \to \mathbb{R}$ be a univariate random variable such that its class conditional distributions are: $N(0,1)$ when the observations come from $\Omega_0$; and $N(d, r^2)$ when they come from $\Omega_1$. Clearly, $X$ is a binormal score. As a direct consequence of the Neyman-Pearson Theorem, the best way to exploit $X$ is through its likelihood ratio. Thus, the quadratic function in Equation 5 is an optimal discriminant score (Mardia et al, 1979).

$$Q(X) = log\frac{f_1(X)}{f_0(X)} + \frac{d^2}{r^2} = (1 - \frac{1}{r^2})X^2 + 2\frac{d}{r^2}X \tag{5}$$

In Equation 5 $f_0$ and $f_1$ represent the class conditional density functions of $X$. Large values of $Q$ indicate that $\Omega_1$ is more likely than $\Omega_0$. We do not have a closed form for the area under the ROC curve of this likelihood ratio; nevertheless, very accurate approximations can be obtained using numerical methods. Table 2 shows the area under the ROC curve of $Q$ for different values of $d$ and $r^2$. The ROC curve for $Q$ was calculated according to Equation 2, and the area under the curve was approximated by the trapezoid rule.

The best discriminant scores are those based on the likelihood ratio; however, the advantages of using a quadratic score are not always evident. As a

matter of fact, in terms of $A$, there is no evident improvement unless $(r^2-1)$ is comparatively larger than $d$. Consider the following transformation of $Q$ (where $r^2>1$):

$$\frac{r^2}{r^2-1}Q = X^2 + 2\frac{d}{r^2-1}X. \tag{6}$$

Clearly, the relative importance of the quadratic term in Equation 6 increases as $r^2$ diverges from 1. By comparing Tables 1 and 2 for any fixed $d$, it is possible to see how, when $r^2 \leq d+1$, $X$ and $Q$ are equivalent in terms of the area under their ROC curves. However, once $r^2 \geq d+1$ the area corresponding to $Q$ increases with $r^2$ while the one corresponding to $X$ continues decreasing to $\frac{1}{2}$. In this way, Equation 6 and Table 2 suggest a practical rule of thumb: *If $d \leq (r^2-1)$, then use a quadratic discriminant score.*

### *3.2 The multivariate case*

Let $\mathbf{X}$ be a random vector such that its class conditional distributions are both multivariate normal (Mardia et al, 1979). Then, its conditional density functions given $\Omega_i$ are:

$$f_i(\mathbf{x}) = \frac{1}{\|2\pi\Sigma_i\|^{1/2}} \exp -\frac{1}{2}(\mathbf{x}-\mu_i)^t \Sigma_i^{-1}(\mathbf{x}-\mu_i) \tag{7}$$

with $i = 0,1$. The problem of discriminating into two multivariate normal distributions is well known and extensively discussed in the literature (Mardia et al, 1979). When a vector $\mathbf{X}$ is normally distributed in both classes we refer to it as a *Multivariate Normal Score* (MNS). Here, we concentrate on the canonical form of a MNS (Kullback, 1968, p. 194-197). This is:

$$(X|\Omega_0) \sim N(\mathbf{0},\mathbf{I}) \text{ and } (X|\Omega_1) \sim N(\delta,\Lambda) \tag{8}$$

where $\mathbf{I}$ is the identity matrix and $\Lambda$ is a diagonal matrix.

When $\Lambda = \mathbf{I}$, the likelihood ratio is transformed into the following linear function:

$$S_\ell = \delta^t \mathbf{X}. \tag{9}$$

Because $\mathbf{X}$ is a MNS, $S_\ell$ is a univariate binormal score with $d = \sqrt{\delta^t\delta}$ and $r^2 = 1$. Thus, its ROC curve is as given as in Definition 2. If $\Lambda \neq \mathbf{I}$, the likelihood ratio is transformed into the following quadratic discriminant function.

$$S_q = \sum_{i=1}^{p}(1-\frac{1}{\lambda_i})X_i^2 + 2.\sum_{i=1}^{p}\frac{\delta_i}{\lambda_i}X_i \tag{10}$$

Please compare $S_q$ in Equation 10 with $Q$ in Equation 5.

## 4 Linear Discriminant Analysis

Let $\mathbf{X}$ be a $p$-dimensional vector of covariates such that its class conditional density functions are both multivariate normal; *i.e.* as given in Equation 7. The parameters $\mu_i$ and $\Sigma_i$ represent the class conditional expectation and class conditional covariance matrix of $\mathbf{X}$ given $\Omega_i$. Let $\theta$ be a constant $p$-dimensional vector. In principle, any linear combination $S = \theta^t(\mathbf{X} - \mu_0)$ is a unidimensional summary of $\mathbf{X}$ that could be used in a classification rule like the one in Equation 1. Thus, it is important to find that linear score $S$ with the best global performance. Given the normality assumption on $\mathbf{X}$, any linear combination of its components is a binormal score and the area under its ROC curve is

$$A_S(\theta) = \Phi\Big[\frac{\theta^t(\mu_1-\mu_0)}{\sqrt{\theta^t(\Sigma_0+\Sigma_1)\theta}}\Big]. \tag{11}$$

We are interested on finding $\theta_*$ such that $A_S(\theta_*) > A_S(\theta)$ for any $\theta \neq \theta_*$

**Theorem 1.** *Let* $\mathbf{X}$ *be a MNS and let* $\mathbf{A} = (\mu_1-\mu_0)(\mu_1-\mu_0)^T$ *and* $\mathbf{B} = (\Sigma_0 + \Sigma_1)$. *Then, no linear combination of the elements of* $\mathbf{X}$ *has an area under its ROC larger than*

$$A_S(\theta_*) = \Phi(\sqrt{\varphi}) \tag{12}$$

*where* $\varphi$ *is the only positive eigenvalue of* $\mathbf{B}^{-1}\mathbf{A}$, $\theta_*$ *its corresponding normalized eigenvector and* $\Phi$ *the cumulative distribution function of a Standard Normal.*

It is possible show that that

$$\varphi = (\mu_1-\mu_0)^t B^{-1}(\mu_1-\mu_0) \tag{13}$$

and

$$\theta_* = \frac{1}{\sqrt{\varphi}}B^{-1}(\mu_1-\mu_0). \tag{14}$$

Theorem 1 is demostrated in Su and Liu (1993).

## 5 Principal Components Analysis for Discrimination

The canonical form of a MNS $\mathbf{X}$, as given in Sect. 3, implies that its components are simultaneously independent in $\Omega_0$ and $\Omega_1$. This may look too restrictive; however, every MNS can be transformed to its canonical form with just a linear transformation.

**Definition 3.** Let $\mathbf{X}$ be a multivariate score such that its class conditional covariance matrices $(\Sigma_0, \Sigma_1)$ are both positive definite and with all their eigenvalues with multiplicity one; also let $(\mu_0, \mu_1)$ be its class conditional expectations. The Principal Components Vector *(PCV)* of $\mathbf{X}$ is:

$$\mathbf{Z} = \Gamma^T \Sigma_0^{-1/2}(\mathbf{X} - \mu_0), \tag{15}$$

where $\Gamma\Lambda\Gamma^T$ is the spectral decomposition of $\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}$, and $\Sigma_0^{1/2}$ is a square root of $\Sigma_0$.

$E(\mathbf{Z}|\Omega_0) = \mathbf{0}$, $Var(\mathbf{Z}|\Omega_0) = \mathbf{I}$, $E(\mathbf{Z}|\Omega_1) = \delta$ and $Var(\mathbf{Z}|\Omega_1) = \Lambda$, where $\delta = \Gamma^T \Sigma_0^{-1/2}(\mu_1 - \mu_0)$ and $\Lambda = diag\{\lambda_i\}_{i=1}^p$ is the matrix of eigenvalues of $\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}$. We will refer to $\mathbf{T} = \Sigma_0^{-1/2}\Gamma$ as the *principal components transformation matrix*.

*Property 1.* Given the conditions of Definition 3, the principal components transformation matrix always exists. If the elements of the diagonal of $\Lambda$ are all different, then the transformation matrix is unique.

*Proof.* If $\Sigma_0$ and $\Sigma_1$ are positive definite, then it is always possible to construct

$$\mathbf{T} = \Sigma_0^{-\frac{1}{2}}\Gamma \tag{16}$$

where $\Sigma_0^{\frac{1}{2}}$ is any square root of $\Sigma_0$ and $\Gamma^T \Sigma_0^{-\frac{1}{2}}\Sigma_1\Sigma_0^{-\frac{1}{2}}\Gamma = \Lambda$. Clearly $\mathbf{T}^T\Sigma_1\mathbf{T} = \Lambda$ and because $\Gamma$ is orthonormal $\mathbf{T}'\Sigma_0\mathbf{T} = \mathbf{I}$. We conclude that under these assumptions a matrix $\mathbf{T}$ always exists. We now show that if $\mathbf{T}$ exists, then it is unique.

Let $\mathbf{T}$ be a square and nonsingular matrix such that $\mathbf{T}^T\Sigma_0\mathbf{T} = \mathbf{I}$ and $\mathbf{T}^T\Sigma_1\mathbf{T} = \mathbf{D}$ for a diagonal matrix $\mathbf{D}$. This is true if and only if $(\mathbf{T}^T)^{-1} = \Sigma_0\mathbf{T}$ and $(\mathbf{T}^T)^{-1}\mathbf{D} = \Sigma_1\mathbf{T}$. Therefore,

$$(\Sigma_0^{-1}\Sigma_1)\mathbf{T} = \mathbf{T}\mathbf{D} \tag{17}$$

In other words, the columns of **T** form a set of linearly independent eigenvectors of $\mathbf{R} = \Sigma_0^{-1}\Sigma_1$ and the elements of the diagonal of **D** are their respective eigenvalues (Harville, 1997, p. 562). If the eigenvalues in **D** are all different, the dimension of $\{\upsilon|(\mathbf{R} - d_i\mathbf{I})\upsilon = 0\}$ is 1, making **T** unique. □

When **X** is a MNS, **Z** is just its canonical representation. Therefore, according to equations (13) and (14) , the linear combination of the elements of **Z** with maximum area under its ROC curve has the following vector of coefficients

$$\xi_* = \left[\frac{\delta_1}{1+\lambda_1}, \frac{\delta_2}{1+\lambda_2}, ..., \frac{\delta_p}{1+\lambda_p}\right]^t, \tag{18}$$

and

$$A_Z(\xi_*) = \Phi\left[\sqrt{\sum_{i=1}^{p} \frac{\delta_i^2}{1+\lambda_i}}\right]. \tag{19}$$

Equation 18 shows how each component $Z_i$ is weighted by the ratio of its difference in means ($\delta_i$) and the sum of its class conditional variances ($1+\lambda_i$). A very important property of the principal components transformation is that $\theta_* = \Sigma_0^{-\frac{1}{2}}\Gamma\xi_*$. This implies that $S_*$ is invariant to the principal components transformation. Therefore, we can say that in terms of $S_*$, there is no loss of information when **X** is transformed into **Z**.

## 6 Reduction of Dimensionality

### *6.1 When $\delta$ is the main source of information*

Each principal component $Z_i$ is a linear score itself and the area under its ROC curve is

$$A_i = \Phi\left[\sqrt{\frac{\delta_i^2}{1+\lambda_i}}\right]. \tag{20}$$

The main objective of PCA is to represent random vectors in a linear space of lower dimension. $A_Z(\xi_*)$ can be used as a criterion to asses and control this reduction of dimensionality (see Equation 19 and (22)). Once **X** is transformed into its principal components vector **Z**, its $p$ components are ordered as $Z_{(1)}, Z_{(2)}, ..., Z_{(p)}$ where

$$\frac{\delta_{(i)}^2}{1+\lambda_{(i)}} \geq \frac{\delta_{(i+1)}^2}{1+\lambda_{(i+1)}} \tag{21}$$

with $i = 1,2,...p-1$. After ordering the components of $\mathbf{Z}$, the sequence $\{\mathscr{R}_k\}_{i=1}^p$ of log odds ratios $\mathscr{R}_k = \frac{\log \frac{A_{Z|k}}{1-A_{Z|k}}}{\log \frac{A_Z(\xi_*)}{1-A_Z(\xi_*)}}$ is computed. In this ratio

$$A_Z|k = \Phi\left[\sqrt{\sum_{i=1}^{k} \frac{\delta_i^2}{1+\lambda_i}}\right] \tag{22}$$

is the maximum area under the ROC curve that can be obtained with a linear combination of the first $k$ principal components $(Z_{(1)}, Z_{(2)}, ..., Z_{(k)})$. Any dimension reduction can imply a smaller area under the ROC curve of the final linear score. Therefore, if the minimum log odds ratio that can be afforded is a $100(1-p)\%$ of $\log \frac{A_Z(\xi_*)}{1-A_Z(\xi_*)}$, the new multivariate score in a lower dimension is obtained by selecting the first $k$ components, where $k$ is the minimum $k$ such that $\mathscr{R}_k \geq p$.

### *6.2 When $\Lambda$ is the only source of information ($\delta = \mathbf{0}$)*

Let $\mathbf{X}$ be a vector of covariates as in Definition 3, and let $\mathbf{Z}$ represent its vector of principal components. If $\delta = \mathbf{0}$, the covariance matrices are the only information we can use to discriminate $\Omega_0$ from $\Omega_1$; under this circumstances

$$S_q = \sum_{i=1}^{p} (1 - \frac{1}{\lambda_i}) Z_i^2. \tag{23}$$

Our objective is to identify those components with the most important contribution to $A_q$ (the area under the ROC curve of $S_q$) and use them to represent $\mathbf{X}$ in a space of lower dimension. For a moment, let us assume that a particular $Z_i$ with $\lambda_i > 1$ is our only discriminant score. Given our current assumption on $\delta$, the class conditional distributions of $Z_i^2$ are[2]: $\left(Z_i^2|\Omega_0\right) \sim \chi_{(1)}^2$ and $\left(\frac{Z_i^2}{\lambda_i}|\Omega_1\right) \sim \chi_{(1)}^2$, when we would classify in $\Omega_1$ when $Z_i^2 > t$ for a certain threshold $t$. Therefore,

[2] $\chi_{(k)}^2$ indicates a Chi-square distribution with $k$ degrees of freedom.

$$A_q = Pr[X \leq \lambda Y],$$

where $X$ and $Y$ represent two independent random variables identically distributed as $\chi^2_{(1)}$. Clearly, $A_q$ (the area unde the ROC curve of $S_q$) is an increasing function of $\lambda_i$. If $\lambda_i < 1$ the decision to classify in $\Omega_1$ would be taken when $Z_i^2 < t$; in this case

$$A_q = 1 - Pr[X \leq \lambda Y]$$

is a decreasing function of $\lambda_i$. Thus, if we consider the complete set of components with

$$0 < \lambda_1 \leq \lambda_2, \ldots \leq \lambda_p < \infty$$

we can conclude that:

- if $\lambda_1 > 1$, then $Z_p$ is the principal component with the largest contribution to $A_q$.
- if $\lambda_p < 1$, then $Z_1$ is the principal component with the largest contribution to $A_q$.
- when $\lambda_1 < 1$ and $\lambda_p > 1$, the maximum contribution to $A_q$ corresponds $Z_1$ if $\lambda_1^{-1} > \lambda_p$, and to $Z_p$ otherwise.

Once $Z_{(1)}$ (the first principal component) has been identified, the same criterion is applied to the rest of the components in order to find $Z_{(2)}$ (the second principal component). The process can be repeated until all the components of $\mathbf{Z}$ are ranked and ordered in terms of their contributions to $A_q$.

Let

$$S_{q|k} = \sum_{i=1}^{k} (1 - \frac{1}{\lambda_{(i)}}) Z^2_{(i)} \tag{24}$$

and let $A_{q|k}$ be the area under its ROC curve. Then, as in the previous subsection, we suggest to use the log odds ratio $\mathscr{R}_k = \dfrac{\log \frac{A_{q|k}}{1-A_{q|k}}}{\log \frac{A_q}{1-A_q}}$ in order to assess the reduction of dimensionality process. The areas $A_{q|k}$ and $A_q$ can be easily approximated using Monte Carlo methods.

### 6.2.1 When $\delta$ and $\Lambda$ are both relevant.

Finally, we consider those situations where $\delta$ and $\Lambda$ are equally important as sources of information to discriminate $\Omega_0$ from $\Omega_1$. These are cases where

despite $\|\delta\| > 0$, the cost of replacing $S_q$ with $S_*$ is too high (see the rule proposed in Sect. 3). The idea is to rank and select the components of $\mathbf{Z}$ in terms of their contribution to $A_q$. It is important to keep in mind that our objective is not only to know the actual value of $A_q$, but also to have a criterion to rank the components of $\mathbf{Z}$; thus, in a first stage we order the principal components with respect to the area under the ROC curve of their marginal quadratic discriminant function $Q$ (see eq. 5). Given the ordered set

$$\{Z_{(1)}, Z_{(2)}, ..., Z_{(k)}\}$$

of the first $k$ principal components we would use

$$S_{q|k} = \sum_{i=1}^{k} (1 - \frac{1}{\lambda_i}) Z_{(i)}^2 + 2 \sum_{i=1}^{k} \frac{\delta_i}{\lambda_i} Z_{(i)} \tag{25}$$

as a discriminant score. Again, if the minimum loss of information that we can afford is the $100(1-p)\%$ of $\log \frac{A_q}{1-A_q}$, we must take the minimum number of components such that $\mathscr{R}_k \geq p$. Given that we do not have a closed analytic form of $A_q$ nor of $A_{q|k}$, we suggest to use Monte Carlo approximations.

## 7 Practical Examples

*Example 1.* We analyze the skull morphometrics of grey kangaroos(*Macropus giganteous*) shown in (Andrews and Herzberg, 1985, p. 307-317). We focus on 50 skulls of the *Macropus giganteous* in order to discriminate male from female individuals. Each observation is a vector of 18 continuous variables. A previous analysis showed that 16 of these variables can be considered to be normally distributed. Two of them were transformed to normality with a Box-Cox transformation. We can assume that the linear score with maximum area under the ROC curve for this data is binormal with parameters $d = 3.33$ and $r^2 = 2.31$. Because $d > r^2 - 1$, we conclude that $S_*$ is enough to discriminate male from female skulls. The results of the PCA are shown in table 3.

The dimensionality of $\mathbf{X}$ can be significantly reduced: three independent principal components are enough to construct a linear score with an area of 0.91 under its ROC curve, and the first six components contain nearly the 90% of the total information. The graph shown in Fig. 1 confirms that a straight line can separate both groups despite its heteroscedasticity.

**Table 3** Discriminating Male from Female skulls of the Macropus giganteous kind

| $Z_i$ | $A_i$ | Rank | $Z_{(q)}$ | $A_{S\|q}$ | $\log \frac{A_{S\|q}}{1-A_{S\|q}}$ | $100 * \mathscr{R}_i\%$ |
|---|---|---|---|---|---|---|
| 1 | 0.8240 | (1) | (1) | 0.8240 | 1.5436 | 45.90 |
| 2 | 0.5977 | (11) | (2) | 0.8850 | 2.0465 | 60.90 |
| 3 | 0.5573 | (13) | (3) | 0.9118 | 2.3360 | 69.54 |
| 4 | 0.7311 | (3) | (4) | 0.9284 | 2.5637 | 76.32 |
| 5 | 0.7772 | (2) | (5) | 0.9409 | 2.7679 | 82.40 |
| 6 | 0.7068 | (5) | (6) | 0.9502 | 2.9496 | 87.81 |
| 7 | 0.6869 | (7) | (7) | 0.9570 | 3.1046 | 92.42 |
| 8 | 0.6989 | (6) | (8) | 0.9596 | 3.1678 | 94.31 |
| 9 | 0.5051 | (18) | (9) | 0.9617 | 3.2242 | 95.98 |
| 10 | 0.6094 | (10) | (10) | 0.9635 | 3.2733 | 97.45 |
| 11 | 0.5455 | (15) | (11) | 0.9648 | 3.3121 | 98.60 |
| 12 | 0.6229 | (8) | (12) | 0.9653 | 3.3262 | 99.02 |
| 13 | 0.5593 | (12) | (13) | 0.9657 | 3.3393 | 99.41 |
| 14 | 0.5144 | (16) | (14) | 0.9660 | 3.3493 | 99.71 |
| 15 | 0.7133 | (4) | (15) | 0.9663 | 3.3575 | 99.95 |
| 16 | 0.5108 | (17) | (16) | 0.9663 | 3.3584 | 99.98 |
| 17 | 0.6167 | (9) | (17) | 0.9663 | 3.3588 | 99.99 |
| 18 | 0.5501 | (14) | (18) | 0.9663 | 3.3589 | 100.00 |

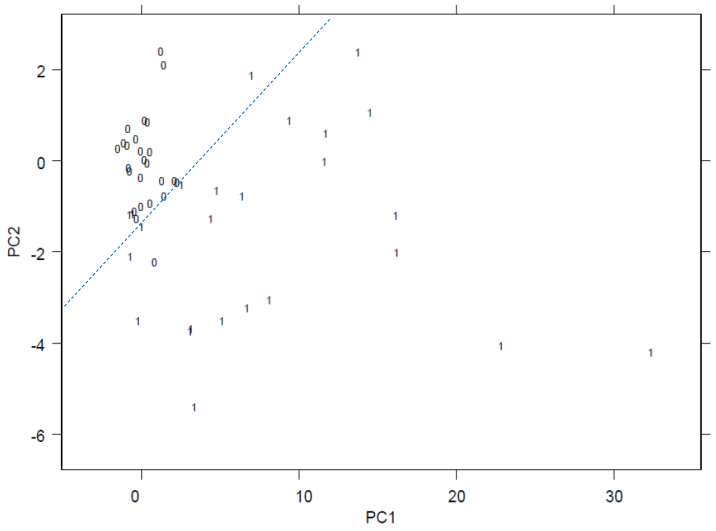


**Fig. 1** Macropus giganteous, second vs first principal component: male = 0, female = 1

*Example 2.* We analyze the physical characteristics of 77 urine samples; 33 with calcium oxalate crystals and 44 without crystals[3]. Six physical characteristics are measured: (1) specific gravity, (2) pH, (3) osmolarity (mOsm), (4) conductivity, (5) urea concentration and (6) calcium concentration. Five of these variables can be assumed as normaly distributed. A Box-Cox Transformation was applied to variable $X_6$ (calcium). The optimal linear score $S_*$ can be considered to be binormal with $d = 2.34$ and $r^2 = 8.81$: $A_S(\theta_*) = 0.77$ and $A_q = 0.95$. Therefore, we decided to base our PCA on the quadratic discriminant score $S_q$; the results are shown in table 4.

**Table 4** Principal Components Analysis; contribution to $A_{q|k}$

| Variable | $A_i$ | Rank | $Z_{(k)}$ | $A_{q\|k}$ | $\log \frac{A_{q\|k}}{1-A_{q\|k}}$ | Ratio |
|---|---|---|---|---|---|---|
| $Z_1$ | 0.8250 | (2) | (1) | 0.8464 | 1.7066 | 55.73 % |
| $Z_2$ | 0.8464 | (1) | (2) | 0.9332 | 2.6369 | 86.11 % |
| $Z_3$ | 0.6370 | (4) | (3) | 0.9373 | 2.7046 | 88.32 % |
| $Z_4$ | 0.5054 | (6) | (4) | 0.9454 | 2.8515 | 93.12 % |
| $Z_5$ | 0.6190 | (5) | (5) | 0.9549 | 3.0527 | 99.69 % |
| $Z_6$ | 0.6483 | (3) | (6) | 0.9553 | 3.0620 | 100.00 % |

The dimensionality of $\mathbf{X}$ can be reduced to the first three principal components keeping the 88% of the total information of $\mathbf{X}$. The graph in Fig. 2 shows the plot of $Z_{(2)}$ against $Z_{(1)}$. The urine samples without crystals are mainly concentrated within a circle of radius 2 centered at the origin, while those with crystals show a different location and are scattered all over the plane.

## 8 Conclusion

We have explored a new method for reduction of dimensionality based on the canonical form of the multivariate normal distribution. This method is flexible enough to be applied when a quadratic function is used instead of a linear one. To the best of our knowledge, this is the first time that the area under the ROC curve is applied as a criterion to assess reduction of dimensionality. Our proposal is based on a simultaneous diagonalization of two covariance matrices. This form of orthogonality suggests that our proposal may be useful as a

[3] The data set is in (Andrews and Herzberg, 1985, p. 249-252)

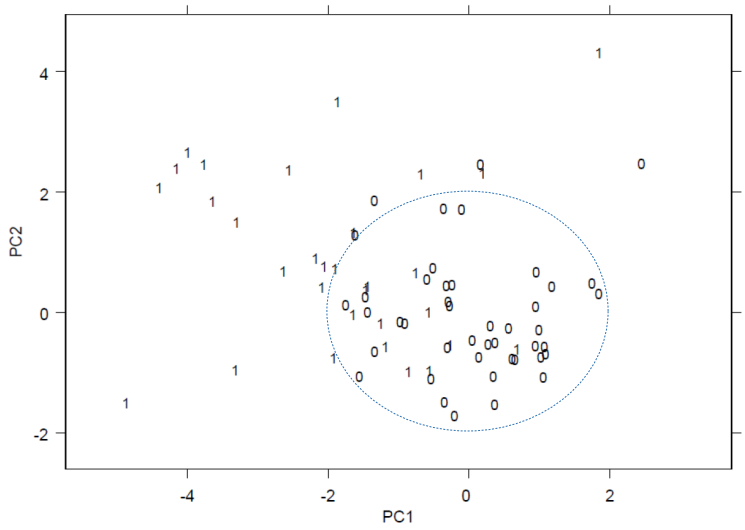


**Fig. 2** Urine chrystals, second vs first principal component: no crystals = 0, crystals = 1

way to eliminate colinearity in logistic regression. We can conclude that this proposal for PCA is parsimonious and interpretable. Our discussion has been focused on a two category discrimination context. Extensions to the three category problem are possible as long as the three covariance matrices involved are proportional. The multy class problem will be the objective of future research.

**Acknowledgements** The authors express their gratitude to the anonymous reviewers for their valuable comments, advice and support.

## References

Andrews DF, Herzberg AM (1985) Data, A collection of problems from Many Fields for the Student and Research Worker. Springer series in Statistics, Springer, New York, DOI 10.1007/978-1-4612-5098-2

Bamber D (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical

Psychology 12(4):387 – 415, DOI 10.1016/0022-2496(75)90001-2

Chang W (1983) Using principal components before separating a mixture of two multivariate normal distributions. Applied Statistics 32(3):420–431, DOI 10.2307/2347949

Cuevas-Covarrubias C (2003) Statistical Inference for ROC-Curves. PhD Dissertation, Statistics Department, University of Warwick, Coventry,UK

Cuevas-Covarrubias C (2013) Principal components analysis for a Gaussian mixture. In: Lausen B, van den Poel D, Ultsch A (eds) Algorithms from and for Nature and Life, Studies in Classification, Data Analysis and Knowledge Organization, Springer International Publishing, Cham, DOI 10.1007/978-3-319-00035-0_17

Harville DA (1997) Matrix Algebra from a Statistitian's Perspective. Springer, New York, DOI 10.1007/b98818

Kullback S (1968) Information Theory and Statistics. Dover Books on Mathematics, Dover Books, Dover

Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis. Academic Press, London

Su JQ, Liu JS (1993) Linear combinations of multiple diagnostic markers. Journal of the American Statistical Association 88(424):1350–1355, DOI 10.1080/01621459.1993.10476417