# Some relations between the comparison of covariance matrices and principal component analysis

Bernhard FLURY
*Department of Statistics, Purdue University, West Lafayette, IN 47907, USA*

Received November 1982
Revised February 1983

*Abstract*: While principal component analysis is a widely used technique in applied multivariate analysis, little attention is normally given to the comparison of covariance matrices. Based on Roy's largest and smallest roots' criterion, we expose some known properties of the eigenvectors of the matrix $\Sigma_1^{-1}\Sigma_2$. The linear combinations defined by these eigenvectors are discussed as a generalisation of principal component analysis to two groups, which can be useful in the case $\Sigma_1 \neq \Sigma_2$. The technique is illustrated by an example. A similar approach to the comparison of covariance matrices, based on the notion of Mahalanobis distance, is sketched. Finally, three equivalent conditions are given for the condition that two covariance matrices have identical principal axes. This leads to the definition of four degrees of similarity of two covariance matrices.

## 1. Principal Component Analysis

Principal Component Analysis (PCA) is a popular technique in multivariate statistics. Its main purpose is to transform $p$ variables $X' = (X_1, \ldots, X_p)$ by an orthogonal matrix $B$ into $p$ new, uncorrelated variables $U = B'X$. We will assume in the sequel that $\mathrm{E}(X) = 0$ and $\mathrm{E}(XX') = \Sigma$, where $\Sigma$ is positive definite and symmetric. Let $\Sigma\beta_i = \lambda_i\beta_i$ $(i = 1, \ldots, p)$, where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_p > 0$ are the eigenvalues of $\Sigma$, and $\beta_i$ are the associated eigenvectors normalized by $\beta_i'\beta_i = 1$, and let $U_i = \beta_i'X$ denote the $i$-th principal component (PC). The most important properties of PC's are as follows:

(i) $U_1$ has the largest variance of all normalized linear combinations of $X$. $U_i$ $(2 \leqslant i \leqslant p)$ has the largest variance of all linear combinations of $X$ uncorrelated with $U_1, \ldots, U_{i-1}$.

(ii) Let $B = (\beta_1, \ldots, \beta_p)$ and $U = (U_1, \ldots, U_p) = B'X$, then $\mathrm{E}(UU') = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$.

(iii) $B'B = I_p$, that is, $B$ is orthogonal. Therefore the PC-transformation can be considered as an orthogonal rotation of the coordinate system.

A detailed mathematical treatment of PCA can be found e.g. in Anderson [1] or

0167-9473/83/$3.00 © 1983, Elsevier Sience Publishers B.V. (North-Holland)

Mardia et al. [8]. For an application-oriented introduction, see Flury & Riedwyl [3]. Note that for normal populations, the maximum likelihood estimates of $\boldsymbol{B}$ and $\Lambda$ are found simply by replacing $\Sigma$ by the sample covariance matrix $\boldsymbol{S}$.

## 2. Comparison of covariance matrices

Methods for comparing two or more covariance matrices have been given little attention in applied statistical analysis. Usually, tests to compare covariance matrices are used only to check assumptions for other multivariate procedures such as MANOVA or linear discriminant analysis. For some test criteria, see Anderson [1, p. 247ff], Roy [13, p. 28], Morrison [9, p. 247ff], Mardia et al. [8, p.123ff]. However, it is not generally known that the union–intersection test introduced by Roy [14] leads in a simple way to a generalization of PCA, which we will discuss in Section 3.

Roy's test is based on the following considerations: Let $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ denote independent $p$-dimensional random vectors with covariance matrices $\Sigma_1$ and $\Sigma_2$, respectively. Obviously,

$$\Sigma_1 = \Sigma_2 \Leftrightarrow \operatorname{var}(\boldsymbol{a}'\boldsymbol{X}^{(1)}) = \operatorname{var}(\boldsymbol{a}'\boldsymbol{X}^{(2)}) \forall \boldsymbol{a} \in \mathbb{R}^p.$$

Since $\operatorname{var}(\boldsymbol{a}'\boldsymbol{X}^{(i)}) = \boldsymbol{a}'\Sigma_i\boldsymbol{a}$ $(i = 1,2)$, the covariance matrices are identical iff $\boldsymbol{a}'\Sigma_2\boldsymbol{a}/\boldsymbol{a}'\Sigma_1\boldsymbol{a} = 1 \ \forall \boldsymbol{a} \in \mathbb{R}^p$, or

$$\min_{a \in \mathbb{R}^p} \frac{\boldsymbol{a}'\Sigma_2\boldsymbol{a}}{\boldsymbol{a}'\Sigma_1\boldsymbol{a}} = \max_{a \in \mathbb{R}^p} \frac{\boldsymbol{a}'\Sigma_2\boldsymbol{a}}{\boldsymbol{a}'\Sigma_1\boldsymbol{a}} = 1. \tag{2.1}$$

Let $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ denote sample covariance matrices, then Roy's test is based on the maximization and minimization of

$$F(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{S}_2\boldsymbol{a}/\boldsymbol{a}'\boldsymbol{S}_1\boldsymbol{a} \tag{2.2}$$

which leads to the problem of finding the eigenvalues of the matrix $\boldsymbol{S}_1^{-1}\boldsymbol{S}_2$ (as will be shown in Section 3). As a test statistic, Roy proposed to use the so called 'largest and smallest root criterion' $(l_1, l_p)$, where $l_1$ and $l_p$ are the largest and smallest characteristic roots of $\boldsymbol{S}_1^{-1}\boldsymbol{S}_2$ respectively.

Though Roy's test seems to be well known, little attention has been given by applied statisticians to the case $\Sigma_1 \neq \Sigma_2$, which is considered as an unpractical violation of assumptions, not worth analyzing in detail. In the following sections, we try to demonstrate that the comparison of two covariance matrices, using as a generalization of PCA, can be a useful tool of multivariate data analysis.

## 3. The comparison of covariance matrices as a generalization of principal component analysis to two groups

Suppose the random vectors $\boldsymbol{X}^{(i)}$ of $p$ components have positive definite symmetric covariance matrices $\Sigma_i$ $(i = 1,2)$. Since we shall be interested only in variances and covariances in the section, we shall assume that the mean vectors are zero. Moreover, in developing the ideas and algebra here, the actual distribution of $\boldsymbol{X}^{(i)}$ is irrelevant except for the covariance matrices. Let $\boldsymbol{b}$ denote a $p$-component column vector. The variance of $\boldsymbol{b}'\boldsymbol{X}^{(i)}$ is

$$\mathrm{E}(\boldsymbol{b}'\boldsymbol{X}^{(i)})^2 = \boldsymbol{b}'\Sigma_i\boldsymbol{b} \quad (i = 1,2). \tag{3.1}$$

We wish to find the linear combination with largest ratio of variances, that is, we wish to maximize the function $h(\boldsymbol{b}) = \boldsymbol{b}'\Sigma_2\boldsymbol{b}/\boldsymbol{b}'\Sigma_1\boldsymbol{b}$ over $\boldsymbol{b} \in \mathbb{R}^p$. Since $h(\boldsymbol{b}) = h(c\boldsymbol{b})$ $\forall c \in \mathbb{R}$ $(c \neq 0)$, maximization can be performed without loss of generality under the restriction $\boldsymbol{b}'\Sigma_1\boldsymbol{b} = 1$ (which means that the linear combination has variance 1 in group 1). Thus we wish to maximize

$$H_1(\boldsymbol{b}) = \boldsymbol{b}'\Sigma_2\boldsymbol{b} - \lambda(\boldsymbol{b}'\Sigma_1\boldsymbol{b} - 1), \tag{3.2}$$

where $\lambda$ is a Lagrange multiplier. The vector of partial derivatives is

$$\frac{\partial H_1}{\partial \boldsymbol{b}} = 2\Sigma_2\boldsymbol{b} - 2\lambda\Sigma_1\boldsymbol{b}. \tag{3.3}$$

A vector $\boldsymbol{b}$ maximizing the ratio of variances must satisfy the expression (3.3) set equal to zero, that is

$$(\Sigma_2 - \lambda\Sigma_1)\boldsymbol{b} = \boldsymbol{0}, \tag{3.4}$$

or by multiplying from the left by $\Sigma_1^{-1}$

$$\left(\Sigma_1^{-1}\Sigma_2 - \lambda\boldsymbol{I}_p\right)\boldsymbol{b} = \boldsymbol{0}. \tag{3.5}$$

It can be shown, that $\Sigma_1^{-1}\Sigma_2$ has $p$ positive characteristic roots; let these be denoted by $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$. Multiplying (3.4) from the left by $\boldsymbol{b}'$ yields

$$\boldsymbol{b}'\Sigma_2\boldsymbol{b} = \lambda\boldsymbol{b}'\Sigma_1\boldsymbol{b} = \lambda. \tag{3.6}$$

This shows that if $\boldsymbol{b}$ satisfies (3.4), then the ratio of variances of the linear combinations $\boldsymbol{b}'\boldsymbol{X}^{(1)}$ and $\boldsymbol{b}'\boldsymbol{X}^{(2)}$ is $\lambda$. Thus for the largest ratio of variances we should use the largest root $\lambda_1$ and the associated eigenvector $\boldsymbol{\beta}_1$ normalized such that $\boldsymbol{\beta}_1'\Sigma_1\boldsymbol{\beta}_1 = 1$. The linear combinations

$$V_1^{(1)} = \boldsymbol{\beta}_1'\boldsymbol{X}^{(1)} \quad \text{and} \quad V_1^{(2)} = \boldsymbol{\beta}_1'\boldsymbol{X}^{(2)} \tag{3.7}$$

have the largest ratio of variances. We will call $(V_2^{(1)},\ V_1^{(2)})$ the first generalized PC's.

In the sequel we will assume that all eigenvalues of $\Sigma_1^{-1}\Sigma_2$ are distinct. For a derivation in the case of multiple characteristic roots, one can proceed along the treatment of PC's given by Anderson [1, p. 273ff]. Suppose now that the first $q$ $(1 \leqslant q < p)$ pairs of generalized PC's have been defined by

$$V_i^{(1)} = \boldsymbol{\beta}_i'\boldsymbol{X}^{(1)}; \qquad V_i^{(2)} = \boldsymbol{\beta}_i'\boldsymbol{X}^{(2)}, \tag{3.8}$$

where $\boldsymbol{\beta}_i$ is the $i$-th eigenvector of $\Sigma_1^{-1}\Sigma_2$, and that the following properties hold:

$$\begin{aligned} &(1) \quad \operatorname{var}\left(V_i^{(1)}\right) = 1; \quad \operatorname{var}\left(V_i^{(2)}\right) = \lambda_i, \quad 1 \leqslant i \leqslant q; \\ &(2) \quad \operatorname{cov}\left(V_i^{(1)},\ V_j^{(1)}\right) = \operatorname{cov}\left(V_i^{(2)},\ V_j^{(2)}\right) = 0 \quad \text{for } 1 \leqslant i < j \leqslant q. \end{aligned} \tag{3.9}$$

We wish now to find the $(q+1)$-th generalized PC's, by maximizing the ratio of variances among all linear combinations $\boldsymbol{b}'\boldsymbol{X}^{(1)}$ and $\boldsymbol{b}'\boldsymbol{X}^{(2)}$ that are uncorrelated with $V_i^{(1)}$ and $V_i^{(2)}$, $1 \leqslant i \leqslant q$, respectively, subject to $\boldsymbol{b}'\Sigma_1\boldsymbol{b} = 1$. Thus we wish to minimize the function

$$h(\boldsymbol{b}) = \boldsymbol{b}'\Sigma_2\boldsymbol{b}/\boldsymbol{b}'\Sigma_1\boldsymbol{b}, \tag{3.10}$$

under the restrictions

$$\begin{array}{llll} \text{(i)} & \boldsymbol{b}'\Sigma_1\boldsymbol{b} = 1, & & \\ \text{(ii)} & \boldsymbol{b}'\Sigma_1\boldsymbol{\beta}_i = 0 & (1 \leqslant i \leqslant q) & \text{(lack of correlation in group 1)}, \\ \text{(iii)} & \boldsymbol{b}'\Sigma_2\boldsymbol{\beta}_i = 0 & (1 \leqslant i \leqslant q) & \text{(lack of correlation in group 2)}. \end{array} \tag{3.11}$$

Since $\Sigma_2\boldsymbol{\beta}_i = \lambda_i\Sigma_1\boldsymbol{\beta}_i$, (ii) implies (iii), and restriction (iii) can therefore be omitted. Thus we wish to maximize the function

$$H_{q+1}(\boldsymbol{b}) = \boldsymbol{b}'\Sigma_2\boldsymbol{b} - \lambda(\boldsymbol{b}'\Sigma_1\boldsymbol{b} - 1) - \sum_{i=1}^{q} \mu_i\boldsymbol{b}'\Sigma_1\boldsymbol{\beta}_i, \tag{3.12}$$

where $\lambda$ and $\mu_i$ $(1 \leqslant i \leqslant q)$ are Lagrange multipliers. The vector of partial derivatives is

$$\frac{\partial H_{q+1}}{\partial \boldsymbol{b}} = 2\Sigma_2\boldsymbol{b} - 2\lambda\Sigma_1\boldsymbol{b} - \sum_{i=1}^{q} \mu_i\Sigma_1\boldsymbol{\beta}_i. \tag{3.13}$$

Setting (3.13) equal to zero and multiplying from the left by $\boldsymbol{\beta}_j'$ $(1 \leqslant j \leqslant q)$ yields

$$2\boldsymbol{\beta}_j'\Sigma_2\boldsymbol{b} - 2\lambda\boldsymbol{\beta}_j'\Sigma_1\boldsymbol{b} - \mu_j\boldsymbol{\beta}_j'\Sigma_1\boldsymbol{\beta}_j - \sum_{\substack{i=1\\ i\neq j}}^{q} \mu_i\boldsymbol{\beta}_j'\Sigma_1\boldsymbol{\beta}_i = 0 \quad (1 \leqslant j \leqslant q). \tag{3.14}$$

The first two terms vanish by (ii) and (iii). By assumption (2) the last sum also vanishes. Since $\boldsymbol{\beta}_j'\Sigma_1\boldsymbol{\beta}_j = 1$ by assumption (1), it follows that $\mu_j = 0$. Therefore, (3.13) set equal to zero gives

$$\Sigma_2\boldsymbol{b} = \lambda\Sigma_1\boldsymbol{b}, \tag{3.15}$$

which shows that $\boldsymbol{b}$ and $\lambda$ must satisfy the same equation as the previous solutions $\boldsymbol{\beta}_i$ and $\lambda_i$ $(1 \leqslant i \leqslant q)$. Since the $q$ largest eigenvalues and associated eigenvectors do not satisfy (ii) and (iii) by assumption (1), the next candidates are obviously $\boldsymbol{\beta}_{q+1}$ and $\lambda_{q+1}$. Multiplying $\Sigma_2\boldsymbol{\beta}_{q+1} = \lambda_{q+1}\Sigma_1\boldsymbol{\beta}_{q+1}$ from the left by $\boldsymbol{\beta}_i'$ and $\Sigma_2\boldsymbol{\beta}_i = \lambda_i\Sigma_1\boldsymbol{\beta}_i$ from the left by $\boldsymbol{\beta}_{q+1}'$ $(1 \leqslant i \leqslant q)$ yields

$$\lambda_{q+1}\boldsymbol{\beta}_i'\Sigma_1\boldsymbol{\beta}_{q+1} = \lambda_i\boldsymbol{\beta}_{q+1}'\Sigma_1\boldsymbol{\beta}_i \quad (1 \leqslant i \leqslant q). \tag{3.16}$$

Since these quantities are scalars, $\boldsymbol{\beta}_i'\Sigma_1\boldsymbol{\beta}_{q+1} \neq 0$ would imply $\lambda_i = \lambda_{q+1}$, which contradicts the assumption of simplicity of all eigenvalues. Therefore, conditions (ii) and (iii) hold, and the linear combinations $V_{q+1}^{(1)} = \boldsymbol{\beta}_{q+1}'\boldsymbol{X}^{(1)}$ and $V_{q+1}^{(2)} = \boldsymbol{\beta}_{q+1}'\boldsymbol{X}^{(2)}$ are the $(q+1)$-th generalized PC's.

Note that, since $\Sigma_1$ and $\Sigma_2$ are assumed to be positive definite, this procedure can be continued. In the last step, the smallest eigenvalue $\lambda_p$ and the associated eigenvector $\boldsymbol{\beta}_p$ are used to define the $p$-th generalized PC's.

Let

$$V^{(1)} = \left(V_1^{(1)},\dots,V_p^{(1)}\right)', \qquad V^{(2)} = \left(V_1^{(2)},\dots,V_p^{(2)}\right)',$$
$$B = \left(\beta_1,\dots,\beta_p\right), \qquad \Lambda = \operatorname{diag}(\lambda_1,\dots,\lambda_p).$$

The above results are summarized in the following theorem.

**Theorem 1.** *Let* $(V^{(1)},V^{(2)}) = (B'X^{(1)}, B'X^{(2)})$ *denote the generalized PC's. Then*

(i) $\mathrm{E}(V^{(1)}V^{(1)\prime}) = I_p$; $\mathrm{E}(V^{(2)}V^{(2)\prime}) = \Lambda$; *that is, the generalized PC's are uncorrelated in both groups.*

(ii) *The i-th pair of generalized PC's* $(V_i^{(1)}, V_i^{(2)}$ *has the largest ratio of variances of all linear combinations uncorrelated with the previous generalized PC's.*

Thus it seems to be justified to name $(V^{(1)}, V^{(2)})$ generalized PC's and the transformation $V = B'X$ (where the subscript is omitted for simplicity) the generalized PC-transformation.

**Remarks.** (1) The generalized PC's can be found more easily by noting the following facts: Since $\Sigma_1$ is regular, there is a regular $p \times p$-matrix $C$ with $C'\Sigma_1 C = I_p$. It can be shown, that the eigenvalues of $\psi = C'\Sigma_2 C$ are identical to the eigenvalues of $\Sigma_1^{-1}\Sigma_2$, and that each eigenvector $\beta_i$ of $\Sigma_1^{-1}\Sigma_2$ corresponds to an eigenvector $\gamma_i = C^{-1}\beta_i$ of $\psi$ (see, e.g., Anderson [1, p. 308]). Since $\psi$ is a positive definite symmetric matrix, the theory of PC's can be used directly to find the generalized PC's. The above derivation is given to stress the similarity between the results presented here and principal components as discussed by Anderson [1, p. 273ff] or Morrison [9, p. 267ff].

(2) Though the above results are not essentially new, they seem to be rather unknown to applied statisticians. We feel that in the case of two groups the generalized PC's could be as useful as PC's in the one sample case. In the next section, we illustrate this by an example.

(3) Of course, generalized PC's are of no use if $\Sigma_1 = \Sigma_2$, in which case $\Sigma_1^{-1}\Sigma_2 = I_p$. The same is true for proportionality, i.e. $\Sigma_2 = c\Sigma_1$ $(c > 0)$, in which case all eigenvalues of $\Sigma_1^{-1}\Sigma_2$ are equal to $c$. The method makes sense, if $\lambda_1 > \lambda_p$ (actually, this is again an analogy to PCA). In practical applications, the equality of $\Sigma_1$ and $\Sigma_2$ can most easily be checked by the largest and smallest roots of $S_1^{-1}S_2$. For $S_1$ and $S_2$ being independently distributed as Wishart with the same parameter matrix $\Sigma_1 = \Sigma_2$ and degrees of freedom $n_1$ and $n_2$ respectively, the tables of Pillai [11] or the charts of Heck [5] (see also Kres, [6]) can be used to get quantiles of the distributions of the extreme roots for some values of $n_1$ and $n_2$. For testing proportionality, the distribution of the ratio of the extreme eigenvalues of $S_1^{-1}S_2$ has been obtained by Pillai et al. [12].

(4) If $\Sigma_1 = I_p$, it is easy to see that the above derivation will yield the PC's of $X^{(2)}$, since $\Sigma_1^{-1}\Sigma_2 = \Sigma_2$ in this case. The same holds if $\Sigma_1 = cI_p$ $(c > 0)$. Thus, once more, the notion of generalized PC's seems to be justified. However, a price has to be paid for this generalization: While, in PCA, the transformation $U = B'X$ can

be considered as an orthogonal rotation of the coordinate system ($B'B = I_p$), the same is in general not true in generalized PCA. That is, the eigenvectors of $\Sigma_1^{-1}\Sigma_2$ are in general not orthogonal, and $B'B$ is in general not diagonal. However, we can ask for conditions, under which the eigenvectors of $\Sigma_1^{-1}\Sigma_2$ are orthogonal. An answer to this question will be given in Section 5.

(5) It is not important, which group is considered as the first and which as the second, for the eigenvalues of $\Sigma_2^{-1}\Sigma_1$ are inverses of the eigenvalues of $\Sigma_1^{-1}\Sigma_2$; the eigenvectors are identical, but ordered in reverse.

(6) It can be shown that, provided that all eigenvalues of $\Sigma_1^{-1}\Sigma_2$ are simple, the generalized PC-transformation is (up to scale factors) the only linear transformation yielding uncorrelated variables in both groups simultaneously.

## 4. An example

Flury and Riedwyl [3] measured the following 6 variables on 100 real and 100 forged swiss bank-notes:

$X_1$: length of the bank-note,
$X_2$: width of the bank-note, measured on the left side,
$X_3$: width of the bank-note, measured on the right side,
$X_4$: width of the lower margin,
$X_5$: width of the upper margin,
$X_6$: length of the print diagonal from the lower left to the upper right corner.

The sample covariance matrices, based on 99 degrees of freedom each, were:

*real notes*:

$$S_1 = \begin{pmatrix} 0.1502 & 0.0580 & 0.0573 & 0.0571 & 0.0145 & 0.0055 \\ 0.0580 & 0.1326 & 0.0859 & 0.0567 & 0.0491 & -0.0431 \\ 0.0573 & 0.0859 & 0.1236 & 0.0582 & 0.0306 & -0.0238 \\ 0.0571 & 0.0567 & 0.0582 & 0.4132 & -0.2635 & -0.0002 \\ 0.0145 & 0.0491 & 0.0306 & -0.2635 & 0.4212 & -0.0753 \\ 0.0055 & -0.0431 & -0.0238 & -0.0002 & -0.0753 & 0.1998 \end{pmatrix},$$

*forged notes*:

$$S_2 = \begin{pmatrix} 0.1240 & 0.0315 & 0.0240 & -0.1006 & 0.0194 & 0.0116 \\ 0.0315 & 0.0650 & 0.0468 & -0.0240 & -0.0119 & -0.0050 \\ 0.0240 & 0.0468 & 0.0889 & -0.0186 & 0.0001 & 0.0342 \\ -0.1006 & -0.0240 & -0.0186 & 1.2813 & -0.4902 & 0.2385 \\ 0.0194 & -0.0119 & 0.0001 & -0.4902 & 0.4045 & -0.0221 \\ 0.0116 & -0.0050 & 0.0342 & 0.2358 & -0.0221 & 0.3112 \end{pmatrix}$$

We can now define the sample generalized PC's as the linear combinations given by the characteristic vectors of $S_1^{-1}S_2$, replacing $\Sigma_1$ and $\Sigma_2$ in Section 3 by their sample counterparts. The eigenvalues and eigenvectors of $S_1^{-1}S_2$ are as follows: eigenvalues:

6.2225 1.6745 1.0516 0.9003 0.5455 0.2839 eigenvectors:

| | | | | | |
|---|---|---|---|---|---|
| 0.9751 | −0.0718 | −1.4129 | 1.9840 | −1.3421 | −0.3961 |
| 0.7054 | 0.0426 | 1.0120 | 1.3528 | 3.3632 | −1.1742 |
| 0.4192 | 1.4190 | 1.9213 | −1.6155 | −2.5544 | −0.3740 |
| −2.2562 | −0.4762 | −0.3505 | −0.0446 | −0.2471 | −0.5121 |
| −1.5528 | 0.4905 | −1.3088 | −0.7537 | 0.0319 | −0.8418 |
| −1.0667 | 1.9275 | 0.1204 | 0.5800 | 0.6345 | 0.5866 |

Under the null hypothesis $\Sigma_1 = \Sigma_2$ and under the assumption of $p$-variate normality in both groups,

$$\mathrm{P}(0.43 \leqslant \text{smallest eigenvalue} \leqslant \text{largest eigenvalue} \leqslant 2.31) = 0.95 \qquad (4.1)$$

approximately [1]. Thus it seems reasonable to perform the generalized PC-transformation to obtain a new set of variables which are uncorrelated in both groups. Looking at the eigenvalues, it seems especially important to analyze the first and the last generalized PC's, since their ratios of variances are quite different from 1. Figure 1 shows a scatterplot of

$$V_1 = 0.975\, X_1 + 0.705\, X_2 + 0.419\, X_3 - 2.256\, X_4 - 1.553\, X_5 - 1.067\, X_6$$

versus

$$V_6 = -0.396\, X_1 - 1.174\, X_2 - 0.374\, X_3 - 0.512\, X_4 - 0.842\, X_5 + 0.587\, X_6$$

in both groups. The group of real notes has a circular shape (due to the constraints $\boldsymbol{\beta}_i' \boldsymbol{S}_1 \boldsymbol{\beta}_i = 1$), while the forged notes have a larger variability in $V_1$ and a smaller variability in $V_6$. As a first result, it is quite astonishing that there is a linear combination for which the forged notes vary much less than the real notes. This demonstrates the quality of the forgerer's production, which seems, apart from some differences in the mean vectors, to be quite precise. On the other hand, the forged notes vary too much in $V_1$, but this variability is obviously caused by an inhomogenity in the group of forged notes. The same two subgroups were found by Flury and Riedwyl [3] using a clustering method. Though surprising at a first glance, this result can be explained: We would expect a similar covariance structure in both groups. However, if one group actually consists of two groups differing in mean vectors, it is obvious that the pooled subgroups would vary too much in the direction of the difference vector of the two subgroup means.

Since the eigenvalues of the intermediate generalized PC's $V_2$ to $V_5$ are rather close to 1, we give no interpretation to these. There seem to be no criteria available to test the hypothesis $\lambda_q = \cdots = \lambda_{q'}$ $(1 \leqslant q \leqslant q' \leqslant p)$. This example suggests that generalized PC's can be used in a way very similar to PC's (see, e.g., applications of PCA in Gnanadesikan [4]). They could also be helpful in quality control, when the production of two different machines or two different lots are compared.

Users of PCA often attempt to interpret the coefficients of some or all of the eigenvectors, and to simplify the structure of the PC-transformation, ignoring

[1] Quantities taken from an unpublished simulation study.

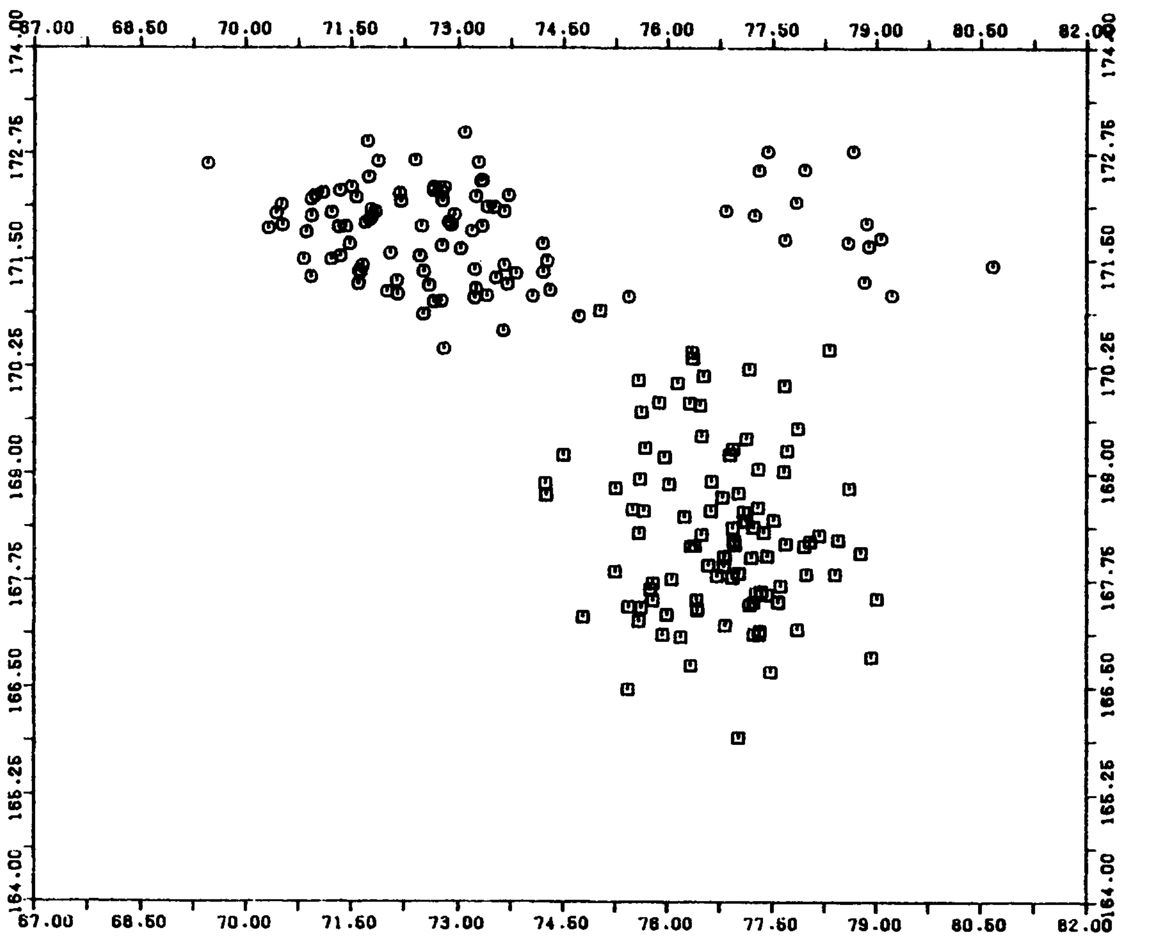

Fig. 1. First and last generalized principal components in the bank note example. Squares indicate real notes, circles forged notes.

coefficients close to zero and equating coefficients of similar magnitude. Using this method, biological applications sometimes yield interpretations of PC's as the well-known 'size'- and 'shape'-variables (see, e.g., Morrison [9], Pimentel [13]). In generalized PCA, the interpretation of the linear transformations is more complicated, since the relation of coefficients to two sets of variables must be considered. As the example shows, the interpretation of eigenvectors may be particularly important when the corresponding ratio of variances deviates strongly from 1. Flury [2] describes a method of stepwise simplification of the eigenvectors associated with the extreme characteristic roots.

## 5. Relations between generalized principal components, principal components and Mahalanobis-distance

Another approach to the comparison of covariance matrices, which turns out to be similar to Roy's approach, is as follows: Suppose the $p$-dimensional random vectors $X^{(1)}$ and $X^{(2)}$ are independent, each with mean vector $\mathbf{0}$ and covariance matrix $\Sigma_i$ $(i = 1,2)$. Consider the Mahalanobis-distance $d_1^2(a) = a'\Sigma_1^{-1}a$ and $d_2^2(a) = a'\Sigma_2^{-1}a$ for points $a \in \mathbb{R}^p$. We ask for points in which the two measures of

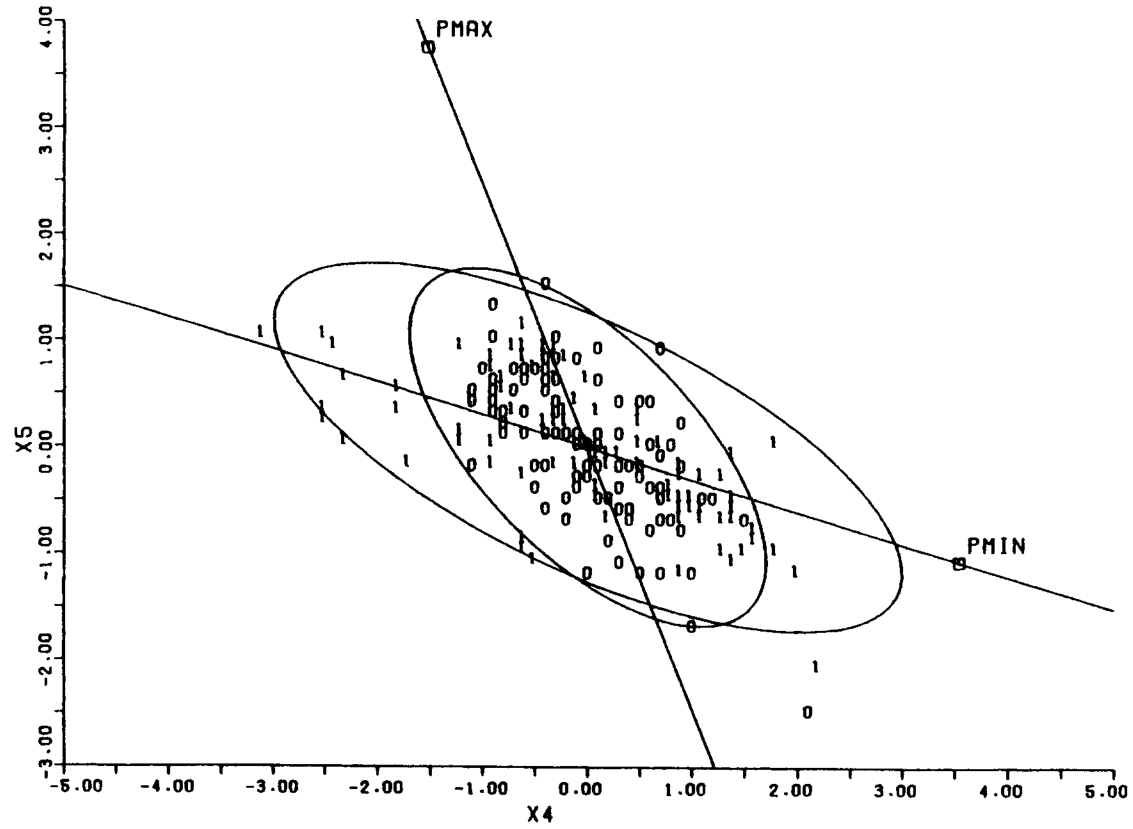

Fig. 2. Points in $\mathbb{R}^2$ with extreme ratios of Mahalanobis-distances. The two ellipses correspond to the sample covariance matrices of $X_4$ and $X_5$ in the bank note example. 0 stands for a real note, 1 for a forged note. Data of both groups were shifted to a common origin.

distance are as different as possible, i.e. for which the ratio $q(a) = d_1^2(a)/d_2^2(a)$ is as large or as small as possible.

Figure 2 illustrates this for the case $p = 2$. The covariance matrices used are taken from the example of Section 4, using varables $X_4$ and $X_5$. Two ellipses indicate the sets of all points in $\mathbb{R}^2$ having Mahalanobis-distance 5 from the origin, where the larger ellips stands for the forged notes, the smaller for the real notes. The ratio $q$ is maximum and minimum at all points on a straight line through the origin and PMAX or PMIN, respectively. It is easy to show that maximization and minimization of $q(a)$ leads to the equation $\Sigma_1^{-1}a = k\Sigma_2^{-1}a$. Multiplying this equation from the left by $\Sigma_1^{-1}\Sigma_2$ yields $\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}a = k\Sigma_1^{-1}a$, which is, with $b = \Sigma_1^{-1}a$, exactly the equation treated in Section 3. Thus the resulting eigenvalues $k_1 \geqslant k_2 \geqslant \cdots \geqslant k_p > 0$ are identical with the eigenvalues of the generalized PC's. However, in general the eigenvectors $a_i$ of this approach are not identical with the eigenvectors of $\Sigma_1^{-1}\Sigma_2$.

The following theorem gives a relation between PCA and generalized PCA.

**Theorem 2.** *Let $\Sigma_1$ and $\Sigma_2$ denote two positive definite, symmetric $p \times p$-matrices. Let $\beta_1, \ldots, \beta_p$ denote the eigenvectors of $\Sigma_1^{-1}\Sigma_2$, normalized according to $\beta_i'\Sigma_1\beta_i = 1$ $(i = 1, \ldots, p)$, and let $B = (\beta_1, \ldots, \beta_p)$. Then the following conditions are equivalent:*

(i) *the principal axes of* $\Sigma_1$ *and* $\Sigma_2$ *are identical,*
(ii) *the columns* of $\boldsymbol{B}$ *are orthogonal* ($\boldsymbol{B}'\boldsymbol{B}$ *is diagonal*),
(iii) $\Sigma_1^{-1}\Sigma_2$ *is symmetric,*
(iv) $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$.

**Proof**: We show that (i) ⇒ (ii) ⇒ (iii) ⇒ (iv) ⇒ (i).

(*i*) ⇒ (*ii*): Let $\gamma_1, \ldots, \gamma_p$ denote a common set of eigenvectors of $\Sigma_1$ and $\Sigma_2$, normalized according to $\gamma_i'\gamma_i = 1$ $(i = 1, \ldots, p)$, and let $\Gamma = (\gamma_1, \ldots, \gamma_p)$. Then $\Gamma'\Sigma_1\Gamma = \Delta_1$ and $\Gamma'\Sigma_2\Gamma = \Delta_2$, where $\Delta_1$ and $\Delta_2$ are diagonal matrices. (Note that, contrary to the previous sections, no fixed order of the eigenvectors is assumed here. In general, the rank order of the diagonal elements is not identical in $\Delta_1$ and $\Delta_2$.) Thus $\Sigma_1 = \Gamma\Delta_1\Gamma'$ and $\Sigma_2 = \Gamma\Delta_2\Gamma'$. The equation $\Sigma_2\boldsymbol{\beta} = \lambda\Sigma_1\boldsymbol{\beta}$ can be written as $\Gamma\Delta_2\Gamma'\boldsymbol{\beta} = \lambda\Gamma\Delta_1\Gamma'\boldsymbol{\beta}$. It follows that $(\Delta_1^{-1}\Delta_2 - \lambda \boldsymbol{I}_p)\Gamma'\boldsymbol{\beta} = \boldsymbol{0}$. Let $\boldsymbol{\xi}_i$ denote the $i$-th eigenvector of $\Delta_1^{-1}\Delta_2$. Since $\Delta_1^{-1}\Delta_2$ is diagonal, it follows that $\boldsymbol{\xi}_i'\boldsymbol{\xi}_j = 0$ $(i \neq j)$. Also, $\boldsymbol{\beta}_i = \Gamma\boldsymbol{\xi}_i$ implies that $\boldsymbol{\beta}_i'\boldsymbol{\beta}_j = \boldsymbol{\xi}_i'\Gamma'\Gamma\boldsymbol{\xi}_j = \boldsymbol{\xi}_i'\boldsymbol{\xi}_j = 0$ $(i \neq j)$. Therefore, the eigenvectors $\boldsymbol{\beta}_j$ are orthogonal.

(*ii*) ⇒ (*iii*): Let $\boldsymbol{\Omega} = \Sigma_1^{-1}\Sigma_2$ and let $\boldsymbol{\beta}_i$ $(i = 1, \ldots, p)$ denote the eigenvectors of $\boldsymbol{\Omega}$, i.e. $\boldsymbol{\Omega}\boldsymbol{\beta}_i = \lambda_i\boldsymbol{\beta}_i$. By multiplication from the left by $\boldsymbol{\Omega}^{-1}$ we get $\boldsymbol{\beta}_i = \lambda_i\boldsymbol{\Omega}^{-1}\boldsymbol{\beta}_i$, and therefore $\boldsymbol{\beta}_i'\boldsymbol{\beta}_j = \lambda_i\boldsymbol{\beta}_i'\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\beta}_j/\lambda_j$. Since $\boldsymbol{\beta}_i'\boldsymbol{\beta}_j = 0$ $(i \neq j)$, $\boldsymbol{\beta}_i'\boldsymbol{\beta}_j = \boldsymbol{\beta}_i'\boldsymbol{\Omega}'^{-1}\boldsymbol{\Omega}\boldsymbol{\beta}_j$ holds for all $i, j$. In matrix form, this can be written as

$$\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{B}'\boldsymbol{\Omega}'^{-1}\boldsymbol{\Omega}\boldsymbol{B}. \tag{5.1}$$

Multiplication from the left by $\boldsymbol{B}'^{-1}$ and from the righty by $\boldsymbol{B}^{-1}$ yields $\boldsymbol{I}_p = \boldsymbol{\Omega}'^{-1}\boldsymbol{\Omega}$. Therefore $\boldsymbol{\Omega} = \boldsymbol{\Omega}'$.

(*iii*) ⇒ (*iv*): Since $\Sigma_1^{-1}\Sigma_2$ is symmetric, it follows that $\Sigma_1^{-1}\Sigma_2 = (\Sigma_1^{-1}\Sigma_2)' = \Sigma_2\Sigma_1^{-1}$. Multiplication on the left and right by $\Sigma_1$ gives $\Sigma_2\Sigma_1 = \Sigma_1\Sigma_2$.

(*iv*) ⇒ (*i*): See Nef [10, p. 229, problem 4].

The following two corollaries can easily be shown:

(1) If $\Sigma_1$ and $\Sigma_2$ have identical principal axes, the eigenvectors of $\Sigma_1^{-1}\Sigma_2$ are (up to differences in scale) identical with the eigenvectors of $\Sigma_1$ and $\Sigma_2$ (or can be chosen in this way, if $\Sigma_1^{-1}\Sigma_2$ has multiple roots). Every eigenvalue of $\Sigma_1^{-1}\Sigma_2$ corresponds to the ratio of an eigenvalue of $\Sigma_2$ and one of $\Sigma_1$.

(2) Under the same conditions, the eigenvectors of $\Sigma_1^{-1}\Sigma_2$ are identical with the eigenvectors of the Mahalanobis-distance approach explained at the beginning of

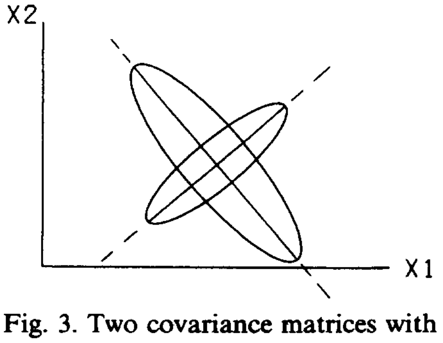

Fig. 3. Two covariance matrices with identical principal axes.

this section; that is, the points in $\mathbb{R}^p$ defined by this approach lie exactly on the common principal axes.

As an illustration of the theorem, Figure 3 shows, for dimension $p = 2$, an example of two covariance matrices having identical principal axes. To examine orthogonality in practical applications, the matrix $\boldsymbol{B}'\boldsymbol{B}$ could be given. However, it seems to be more advantageous to compute the cosines of the angles between the eigenvectors of $\boldsymbol{S}_1^{-1}\boldsymbol{S}_2$. If $\phi_{ij}$ denotes the angle between eigenvectors $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, then

$$f_{ij} = \cos\phi_{ij} = \boldsymbol{\beta}_i'\boldsymbol{\beta}_j / \left(\boldsymbol{\beta}_i'\boldsymbol{\beta}_i\boldsymbol{\beta}_j'\boldsymbol{\beta}_j\right)^{1/2}. \tag{5.2}$$

The matrix $\boldsymbol{F} = (f_{ij})$ is comparable to a correlation matrix, since it contains ones in the main diagonal and zeros elsewhere if the eigenvectors are orthogonal. This might be a useful complement to a method proposed be Krzanowski [7], which compares the PC's of two groups by computing the angle between the hyperplanes spanned by the first $q$ PC's of each group.

In our example of section 4, the angles between the sample eigenvectors are

$$\boldsymbol{F} = \begin{pmatrix} 1.0000 & -0.1490 & 0.3026 & 0.2934 & -0.0123 & 0.0840 \\ -0.1490 & 1.0000 & 0.3606 & -0.2119 & -0.1817 & 0.0950 \\ 0.3026 & 0.3606 & 1.0000 & -0.3888 & 0.0391 & 0.0009 \\ 0.2934 & -0.2119 & -0.3888 & 1.0000 & 0.4662 & -0.1467 \\ -0.0123 & -0.1817 & 0.0391 & 0.4662 & 1.0000 & -0.2567 \\ 0.0840 & 0.0950 & 0.0009 & -0.1467 & -0.2567 & 1.0000 \end{pmatrix}$$

While the two extreme eigenvectors are nearly orthogonal ($\phi_{16} = 85.2°$), some intermediate eigenvectors deviate considerably from orthogonality. However, no criteria seem to be available to test the hypothesis of orthogonality. Estimation of the common principal axes of two groups is a problem, on which the author is presently working. Although a solution has been found for the case $p = 2$, a general solution seems to be rather difficult to obtain.

## 6. Final remarks

The theorem given in Section 5 suggests the following definition of *4 degrees of similarity of two covariance matrices* $\Sigma_1$ *and* $\Sigma_2$:

*degree* 1: $\Sigma_1 = \Sigma_2$ (identity),
*degree* 2: $\Sigma_1 = c\Sigma_2$, $c > 0$ (proportionality),
*degree* 3: $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$ (identical principal axes),
*degree* 4: $\Sigma_1$, $\Sigma_2$ arbitrary.

All pairs of covariance matrices of a degree are contained in the following degrees, that is, the ordering is hierarchical. For a given dimension $p$, the number of parameters in a pair of covariance matrices is $p(p+1)/2$, $p(p+1)/2+1$, $p(p+3)/2$ and $p(p+1)$ for degrees 1 to 4, respectively. Thus, if $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$ holds in the population, a lot of superfluous parameter estimation could be avoided.

As mentioned above, test criteria for degree 3 seem not to be known. However, we feel that estimation and testing of identical principal axes would not only reduce the number of estimated parameters, but could also lead to simpler interpretations of the PC's of two groups.

Moreover, under the assumptions of degree 3, computation of Mahalanobis-distances for classifying observations could be simplified as follows: Since the principal axes are identical, the same orthogonal transformation yields uncorrelated variables in both groups. As Mahalanobis-distance is unaffected by orthogonal transformations, it can be computed in the common space of principal components. The formula for computing the distance between an observation $x=(x_1, \ldots, x_p)'$ and the mean vector $\mu^{(j)}=(\mu_1^{(j)}, \ldots, \mu_p^{(j)})'$ $(j=1,2)$ becomes simply

$$D_j^2(x)=d_j'\Lambda_j^{-1}d_j=\sum_{i=1}^{p} d_{ji}^2/\lambda_{ji} \quad (j=1,2) \tag{6.1}$$

where $d_j=(d_{j1}, \ldots, d_{jp})=B'(x-\mu^{(j)})$, $\Lambda_j=\operatorname{diag}(\lambda_{j1}, \ldots, \lambda_{jp})=B'\Sigma_j B$, and $B$ is the matrix of (common) characteristic vectors of $\Sigma_1$ and $\Sigma_2$.

## Acknowledgement

Research for this paper was partly done while the author worked as an assistant in the Department of Statistics, University of Berne, Switzerland, and partly under contract No. 82.008.0.82 of the Swiss National Science Foundation at Purdue University. The author wishes to thank Henri Carnal, George McCabe, Walter Nef, K.C.S. Pillai, Hans Riedwyl and a referee for some helpful comments and remarks.

## References

[1] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis* (Wiley, New York, 1958).

[2] B. Flury, Analyse von Linearkombinationen mit extremen Varianzenquotienten, Dissertation, Universität Bern (1982).

[3] B. Flury and H. Riedwyl, *Angewandte Multivariate Statistik* (Gustav Fischer, Stuttgart & New York, 1983).

[4] R. Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations* (Wiley, New York, 1977).

[5] D.L. Heck, Charts of some upper percentage points of the distribution of the largest characteristic root, *Ann. Math. Stat.* **31** (1960) 625-642.

[6] H. Kres, *Statistische Tafeln zur Multivariaten Analysis* (Springer, Berlin, 1975)

[7] W.J. Krzanowski, Between-groups comparison of principal components, *J. Amer. Statist. Assoc.* **74** (1979) 703-707.

[8] K.V. Mardia, J.T. Kent and J.M. Bibby, *Multivariate Analysis* (Academic Press, London, 1979).

[9] D.F. Morrison, *Multivariate Statistical Methods* (McGraw-Hill, New York, 1976).

[10] W. Nef, *Lineare Algebra* (Birkhäuser, Basel, 1966).

[11] K.C.S. Pillai, Upper percentage points of the largest root of a matrix in multivariate analysis, *Biometrika* **54** (1967) 189-194.

[12] K.C.S. Pillai, S. Al-Ani and G.M. Jouris, On the distributions of the ratios of the roots of a covariance matrix and Wilk's criterion for tests of three hypotheses, *Ann. Math. Stat.* **40** (1969) 2033-2040.

[13] R.A. Pimentel, *Morphometrics* (Kendall/Hunt, Dubuque, IA, 1979).

[14] S.N. Roy, *Some Aspects of Multivariate Analysis* (Wiley, New York, 1957).