Computational Statistics and Data Analysis 54 (2010) 3446–3457

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Stepwise estimation of common principal components

Nickolay T. Trendafilov

Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK

ARTICLE INFO

Article history:
Received 24 October 2009
Received in revised form 8 February 2010
Accepted 11 March 2010
Available online 20 March 2010

Keywords:
Simultaneous diagonalization
Dimensionality reduction
Power iterations for k symmetric positive definite matrices

ABSTRACT

The standard common principal components (CPCs) may not always be useful for simultaneous dimensionality reduction in k groups. Moreover, the original *FG* algorithm finds the CPCs in arbitrary order, which does not reflect their importance with respect to the explained variance. A possible alternative is to find an approximate common subspace for all k groups. A new stepwise estimation procedure for obtaining CPCs is proposed, which imitates standard PCA. The stepwise CPCs facilitate simultaneous dimensionality reduction, as their variances are decreasing at least approximately in all k groups. Thus, they can be a better alternative for dimensionality reduction than the standard CPCs. The stepwise CPCs are found sequentially by a very simple algorithm, based on the well-known power method for a single covariance/correlation matrix. Numerical illustrations on well-known data are considered.

1. Introduction

The common principal components (CPC) model was introduced and studied by Flury (1988). It is one of many possible generalizations of standard principal component analysis (PCA) of several covariance matrices (Jolliffe, 2002). The initial motivation for introducing CPC was to study discrimination problems where the group covariance matrices are not equal as required by linear discriminant analysis, but more generally share common principal axes (Krzanowski, 1984; Flury, 1988).

There are k normal populations with the same mean vector and it is assumed that their $p \times p$ covariance matrices $\Sigma_i, i = 1, 2, \ldots, k$ are all positive definite (p.d.) and different. The hypothesis of the CPC model is that the covariance matrices Σ_i are simultaneously diagonalizable, i.e.:

$$H_{\text{CPC}} : Q^\top \Sigma_i Q = D_i^2, \quad i = 1, 2, \ldots, k$$

where Q is a *common* $p \times p$ orthogonal matrix for all populations and D_i^2 are positive diagonal matrices *specific* to each population. The CPC estimation problem is, for given sample covariance matrices, S_i, to find their common eigenvectors and corresponding (different) eigenvalues. Flury (1984) proposed the maximum likelihood procedure for their estimation, which is implemented in the highly efficient *FG* algorithm based on Jacobi rotations. He also established a number of useful properties of these estimators: e.g. the statistic for testing the CPC null hypothesis H_{CPC} has asymptotically a χ^2 distribution.

The CPC model is mainly criticized because it is essentially a method for simultaneous diagonalization of several positive definite matrices, rather than a method for dimensionality reduction which is usually the main goal in data analysis. Indeed, the CPC model produces specific eigenvalues which may not be simultaneously ordered in decreasing order in all groups. In other words, the "common subspace need not be associated with the largest latent roots in each group, so it is not appropriate for dimensionality reduction" (Schott, 1988).

To remedy this problem with the CPC model, Krzanowski (1984) proposed a simple intuitive procedure to estimate approximately the CPCs. It is based on the PCA of the pooled sample covariance matrix and the total sample covariance

E-mail address: N.Trendafilov@open.ac.uk.

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.03.010

matrix, followed by comparison of their eigenvectors. He also proposed a descriptive procedure for testing the hypothesis that the subspaces spanned by the first r principal components of the k groups are the same (Krzanowski, 1979). A more precise approximate procedure for testing the same hypothesis was developed by Schott (1988). It is based on the fact that "…if the subspaces spanned by the first m principal components for the groups are the same, then they will be identical also to the subspace spanned by the first m principal components of the sum…" (Schott, 1988, p. 230). The test derivations suggested essentially the same method for approximate estimation of the CPCs: PCA of the pooled sample covariance matrix. This approximate CPC estimation is briefly outlined and illustrated on a small data set in Schott (1988). In the sequel, this type of approximate CPC estimation is referred to as common subspace analysis.

Clearly, a more precise method for CPC estimation is required which can replace Flury's diagonalization procedure and affords simultaneous dimensionality reduction in k groups. This paper proposes a new procedure for *stepwise* CPC estimation, which mimics the standard PCA performance.

If one is interested in considering CPCs one after another, then $m = 1$ in Schott's assumptions. Suppose that two covariance matrices S_1 and S_2 have common eigenvector (CPC) q. Then it is an eigenvector of $S_1 + S_2$ too. Indeed, if $S_1 q = \lambda q$ and $S_2 q = \mu q$, then $(S_1 + S_2)q = (\lambda + \mu)q$. However, note also that the common eigenvector q of the two covariance matrices S_1 and S_2 is an eigenvector of $\frac{1}{2}\left(\frac{S_1}{\lambda} + \frac{S_2}{\mu}\right)$ too. It turns out that this is a more appropriate choice than $S_1 + S_2$ for recovering the CPCs of S_1 and S_2.

The stepwise CPCs minimize the same objective function as the standard CPCs, but in a sequential manner. The benefit of using stepwise CPCs is that one can find only a prescribed number $r(<p)$ of them, and be sure that every such CPC explains more variance than the following one. Moreover, when Flury's specific eigenvalues are (or can be made) simultaneously decreasing in all groups, then the stepwise CPCs coincide with them. However, if this is not the case, then the stepwise CPCs manage to keep such an ordering at least approximately. Thus, the stepwise CPCs can be a better alternative for dimensionality reduction than the standard CPCs.

The paper is organized as follows. The CPC estimation problem is formulated in Section 2. Stepwise CPCs are introduced and studied in Section 3. In Section 4, the classical power method (Golub and Van Loan, 1996) for diagonalization of a single symmetric p.d. matrix is extended to simultaneously diagonalize several matrices and produce stepwise CPCs. Finally, the stepwise CPCs are illustrated with several standard data sets.

2. CPC estimation problem

The CPC model was introduced and studied by Flury (1984, 1988). There are considered to be k normal populations and it is assumed that their $p \times p$ covariance matrices Σ_i, $i = 1, 2, \ldots, k$ are p.d. The hypothesis of the CPC model is that they can be decomposed simultaneously as:

$$\Sigma_i = QD_i^2Q^\top, \tag{1}$$

where Q is a common $p \times p$ orthogonal matrix for all Σ_i, $i = 1, 2, \ldots, k$, and D_i^2 are positive $p \times p$ diagonal matrices.

The CPC estimation problem is then for given $p \times p$ sample covariance matrices S_i, $i = 1, 2, \ldots, k$ with $n_i(>p)$ degrees of freedom, to find their common eigenvectors and corresponding (different) eigenvalues, such that:

$$S_i \approx QD_i^2Q^\top \tag{2}$$

as closely as possible in some sense. Maximum likelihood estimation of the parameters Q and $D_i^2, i = 1, 2, \ldots, k$ is formulated in Flury (1988) as the following optimization problem:

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{k} n_i \left[\log(\det(QD_i^2Q^\top)) + \operatorname{trace}((QD_i^2Q^\top)^{-1}S_i)\right] \\ & = \sum_{i=1}^{k} n_i \left[\log(\det(D_i^2)) + \operatorname{trace}(D_i^{-2} \odot (Q^\top S_i Q))\right] \end{aligned} \tag{3}$$

$$\text{Subject to} \quad (Q, D_1, D_2, \ldots, D_k) \in \mathcal{O}(p) \times \mathcal{D}(p)^k, \tag{4}$$

where $\odot$ denotes the Hadamard matrix product. The Lie group of all $p \times p$ orthogonal matrices is denoted by $\mathcal{O}(p)$ and $\mathcal{D}(p)^k = \underbrace{\mathcal{D}(p) \times \cdots \times \mathcal{D}(p)}_{k}$, where $\mathcal{D}(p)$ is the linear subspace of all $p \times p$ diagonal matrices. The first order optimality conditions for a stationary point of the CPC objective function (3) are given in the following:

Theorem 2.1. *A necessary condition for* $(Q, D_1, \ldots, D_k) \in \mathcal{O}(p) \times \mathcal{D}(q)^k$ *to be a stationary point of the CPC objective function (3) is that the following* $k + 1$ *conditions must hold simultaneously:*

(a) $\sum_{i=1}^{k} n_i Q^\top S_i Q D_i^{-2}$ *is symmetric;*
(b) $\operatorname{diag}(Q^\top S_i Q) = D_i^2$.

After substitution of D_i^2 from (b) into (3), CPC estimation can be redefined as the following profile likelihood problem:

$$\text{Minimize} \quad \sum_{i=1}^{k} n_i \log(\det(\operatorname{diag}(Q^\top S_i Q))) \tag{5}$$

$$\text{Subject to} \quad Q \in \mathcal{O}(p), \tag{6}$$

which is, in fact, what the *FG* diagonalization algorithm solves (Flury, 1988), rather than the original problem (3)–(4).

3. Stepwise CPCs

The solution Q of the CPC problem (5)–(6) is an orthogonal matrix containing all the CPCs, i.e. all CPCs are found simultaneously. Instead, one can try to imitate standard PCA by finding the CPCs one after another. Such an alternative stepwise solution can be useful for simultaneous dimensionality reduction in all k groups. It eliminates the need for finding all CPCs first, and then choosing, say, the two CPCs explaining the largest total variances in all groups.

3.1. Simultaneously decreasing eigenvalues in all groups

For this purpose, one can rewrite the CPC problem (5)–(6) in vectorized form:

$$\text{Minimize} \quad \sum_{i=1}^{k} n_i \sum_{j=1}^{p} \log(q_j^\top S_i q_j) \tag{7}$$

$$\text{Subject to} \quad [q_1, q_2, \ldots, q_p] \in \mathcal{O}(p). \tag{8}$$

If the (*FG*) solution of the CPC problem (5)–(6), gives eigenvalues q_i simultaneously decreasing in all k groups, i.e.

$$q_1^\top S_i q_1 \geq q_2^\top S_i q_2 \geq \cdots \geq q_p^\top S_i q_p \tag{9}$$

for every $i = 1, 2, \ldots, k$, then obviously

$$\sum_{i=1}^{k} n_i \log(q_1^\top S_i q_1) \geq \sum_{i=1}^{k} n_i \log(q_2^\top S_i q_2) \geq \cdots \geq \sum_{i=1}^{k} n_i \log(q_p^\top S_i q_p).$$

This means that the vectorized CPC problem (7)–(8) can be solved sequentially by solving p identical problems, of the form:

$$\text{Minimize} \quad \sum_{i=1}^{k} n_i \log(q^\top S_i q) \tag{10}$$

$$\text{Subject to} \quad q^\top q = 1. \tag{11}$$

The first CPC to be found is q_p, which gives the minimum of (10) on the unit sphere in $\mathbb{R}^p$. The next CPC to be found is q_{p-1}, which gives the minimum of (10) on the unit sphere in $\mathbb{R}^p$ being orthogonal to q_p. Each minimum found this way is greater than the previous one, being found in an orthogonal subspace of the previous minimization domain. The objective function in (10) is bounded on the unit sphere in $\mathbb{R}^p$ as the quantities $q^\top S_i q$ are bounded below and above by the smallest and the largest original eigenvalues of S_i.

For dimensionality reduction purposes, it is more useful to obtain the CPCs in a reverse manner starting with q_1. This can be achieved by following the variational eigenvalue definition (Horn and Johnson, 1986). Indeed, denote by $Q_p = [q_1, q_2, \ldots, q_p]$ the $p \times p$ orthogonal matrix collecting the CPCs found by solving (10)–(11). Then the jth CPC q_j is a solution of the following minimization problem:

$$\text{Minimize} \quad \sum_{i=1}^{k} n_i \log(q^\top S_i q) \tag{12}$$

$$\text{Subject to} \quad q^\top q = 1 \quad \text{and} \quad q^\top Q_{j+1}^{\perp} = 0_{p-j}^\top, \tag{13}$$

where $Q_{j+1}^{\perp} = [q_{j+1}, \ldots, q_{p-1}, q_p]$ and 0_{p-j} is a zero vector of size $p - j$. Alternatively, the jth CPC q_j can be expressed as a solution of the following maximization problem:

$$\text{Maximize} \quad \sum_{i=1}^{k} n_i \log(q^\top S_i q) \tag{14}$$

$$\text{Subject to} \quad q^\top q = 1 \quad \text{and} \quad q^\top Q_{j-1} = 0_{j-1}^\top, \tag{15}$$

where $Q_0 := 0$ and $Q_{j-1} = [q_1, q_2, \ldots, q_{j-1}]$. To see this, consider a basis in $\mathbb{R}^p$ formed by the column vectors of Q_p, i.e. any $q \in \mathbb{R}^p$ can be written as a linear combination of the columns of Q_p as $q = \xi_1 q_1 + \cdots + \xi_p q_p$. Let first $q \perp Q_{j+1}^{\perp}$ and $q^\top q = 1$. Then q can be written as $q = \xi_1 q_1 + \cdots + \xi_j q_j$ with $\xi_1^2 + \cdots + \xi_j^2 = 1$. Taking into account the inequalities (9) it follows that:

$$\sum_{i=1}^{k} n_i \log(q^\top S_i q) = \sum_{i=1}^{k} n_i \log(\xi_1^2 q_1^\top S_i q_1 + \cdots + \xi_j^2 q_j^\top S_i q_j) \geq \sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j)$$

with equality when $q = q_j$, which implies:

$$\min_{\substack{q^\top q = 1 \\ q \perp Q_{j+1}^{\perp}}} \sum_{i=1}^{k} n_i \log(q^\top S_i q) = \sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j). \tag{16}$$

Now, let $q \perp Q_{j-1}$ and $q^\top q = 1$. Then q can be written as $q = \xi_j q_j + \cdots + \xi_p q_p$ with $\xi_j^2 + \cdots + \xi_p^2 = 1$, and from (9) it follows that:

$$\sum_{i=1}^{k} n_i \log(q^\top S_i q) = \sum_{i=1}^{k} n_i \log(\xi_j^2 q_j^\top S_i q_j + \cdots + \xi_p^2 q_p^\top S_i q_p) \leq \sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j)$$

with equality when $q = q_j$, which implies:

$$\max_{\substack{q^\top q = 1 \\ q \perp Q_{j-1}}} \sum_{i=1}^{k} n_i \log(q^\top S_i q) = \sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j). \tag{17}$$

Thus, the jth CPC q_j can be found as a solution of j problems identical to (14)–(15). The final solution $Q_p = [q_1, q_2, \ldots, q_p]$ coincides with the matrix solution of the CPC problem (5)–(6), but the process can be aborted at any $1 \leq j \leq p$ for dimensionality reduction. If the next q_{j+1} is required, its calculation needs the solution of a single maximization problem making use of the already known Q_j. There is no need to restart the whole cycle of maximization problems.

Theorem 3.1. *The first order optimality conditions for the problem (14)–(15) are, for $j = 1, 2, \ldots, p$:*

$$\pi_j \left(\sum_{i=1}^{k} \frac{n_i S_i}{q_j^\top S_i q_j} \right) q_j = 0_{p \times 1}, \tag{18}$$

where π_j denotes the projector $I_p - Q_j Q_j^\top$ and $g(q_j) = \sum_{i=1}^{k} \frac{n_i S_i q_j}{p t q_j^\top S_i q_j}$ is the gradient of the CPC objective function (14).

Note, that $g(q)$ is the gradient of the CPC objective function (14) embedded in $\mathbb{R}^p$, while the expression in (18) is the gradient of the CPC objective function on the unit sphere in $\mathbb{R}^p$ (Absil et al., 2008, Ch. 4.6.1).

In particular, for $j = 1$, the first order optimality conditions (18) are simplified to:

$$\left(n_1 \frac{S_1}{q_1^\top S_1 q_1} + \cdots + n_k \frac{S_k}{q_1^\top S_k q_1} - n I_p \right) q_1 = 0_{p \times 1}, \tag{19}$$

which is the first order optimality condition for the first common principal component q_1. If $q^\top S_1 q = q^\top S_2 q = \cdots = q^\top S_k q$, the stepwise CPC q is an eigenvector of the pooled covariance matrix $\bar{S} = \frac{n_1 S_1 + n_2 S_2 + \cdots + n_k S_k}{n}$, where $n = \sum_{i=1}^{k} n_i$, i.e. q coincides with the first eigenvector from the common space analysis of $\bar{S}$ (Krzanowski, 1984; Schott, 1988).

For equal numbers of observations in all groups (and $q := q_1$) Eq. (19) becomes:

$$\left(\frac{S_1}{q^\top S_1 q} + \cdots + \frac{S_k}{q^\top S_k q} - k I_p \right) q = 0_{p \times 1}, \tag{20}$$

which can be seen as a generalized definition of the symmetric eigenvalue problem for k matrices simultaneously.

For $j = 2$, the first order optimality conditions (18) become:

$$\left(n_1 \frac{S_1}{q_2^\top S_1 q_2} + \cdots + n_k \frac{S_k}{q_2^\top S_k q_2} - n I_p \right) q_2 - q_1 \left(n_1 \frac{q_1^\top S_1 q_2}{q_2^\top S_1 q_2} + \cdots + n_k \frac{q_1^\top S_k q_2}{q_2^\top S_k q_2} \right) = 0_{p \times 1} \tag{21}$$

where the second term in (21) vanishes if the matrices S_i for $i = 1, 2, \ldots, k$ are simultaneously (exactly) diagonalizable. In this case, the first order optimality conditions for the second common principal component q_2 would be:

$$\left(n_1 \frac{S_1}{q_2^\top S_1 q_2} + \cdots + n_k \frac{S_k}{q_2^\top S_k q_2} - n I_p \right) q_2 = 0_{p \times 1},$$

and so on for any q_j ($j = 3, \ldots, p$).

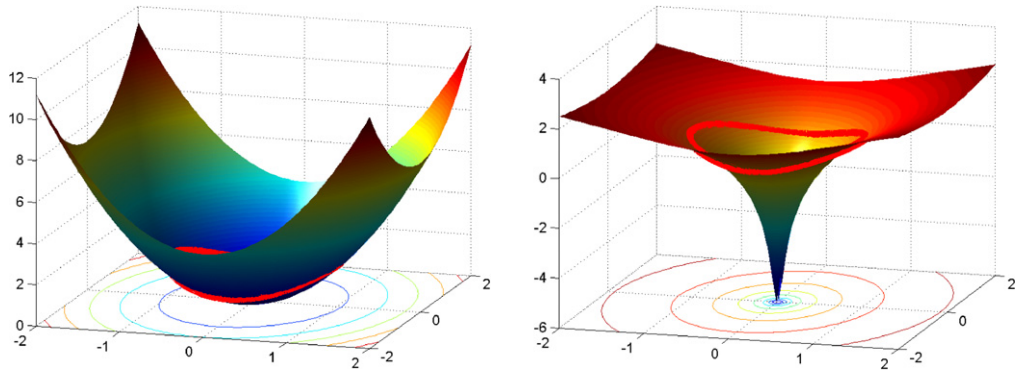

Fig. 1. Plots of the functions $q^\top Sq$ and $\log(q^\top Sq)$. The red contours on the surfaces denote the curves formed by the corresponding images of the unit circle $q^\top q = 1$, over which maxima and minima are sought. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In general, the matrices S_i for $i = 1, 2, \ldots, k$ are not exactly diagonalizable and the second term in (21) does not vanish. Then, the first order optimality conditions (21) for $j = 2$ can be rewritten in more convenient form as:

$$\left[(I_p - q_1 q_1^\top)\left(n_1 \frac{S_1}{q_2^\top S_1 q_2} + \cdots + n_k \frac{S_k}{q_2^\top S_k q_2}\right) - nI_p\right] q_2 = 0_{p\times 1}, \tag{22}$$

or in general for any q_j $(j = 2, \ldots, p)$ and $Q_{j-1} = [q_1, \ldots, q_{j-1}]$ as

$$\left[(I_p - Q_{j-1} Q_{j-1}^\top)\left(n_1 \frac{S_1}{q_j^\top S_1 q_j} + \cdots + n_k \frac{S_k}{q_j^\top S_k q_j}\right) - nI_p\right] q_j = 0_{p\times 1}. \tag{23}$$

In Section 4, power iterations based on Eqs. (19) and (23) are constructed for simultaneous (approximate) diagonalization of k symmetric matrices. This gives an efficient estimation procedure for finding the CPCs of S_i for $i = 1, 2, \ldots, k$.

For $k = 1$, the optimality conditions (18) are reduced to the optimality conditions of standard PCA (Trendafilov, 2006; Trendafilov and Jolliffe, 2006):

$$\pi_j S_1 q_j = 0_{p\times 1}, \tag{24}$$

and for $j = 1$, one simply has:

$$(I_p - qq^\top)\frac{S_1 q}{q^\top S_1 q} = \left(\frac{S_1}{q^\top S_1 q} - I_p\right) q = 0_{p\times 1}, \tag{25}$$

which shows that the critical points of PCA and CPC coincide.

If the objective function $\log(q^\top Sq)$ is maximized at $q_\wedge$, then $q_\wedge$ also maximizes $q^\top Sq$. Indeed, $\log(q_\wedge^\top Sq_\wedge) \geq \log(q^\top Sq)$ implies $q_\wedge^\top Sq_\wedge \geq q^\top Sq$ for all q in a neighborhood of $q_\wedge$, because the logarithm is a monotonically increasing function. Thus, as can be expected, for $k = 1$ the CPC solution coincides with the PCA solution. Fig. 1illustrates this for $p = 2$ and covariance matrix $S = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$.

3.2. General case

Now consider the case when the standard CPCs obtained by solving the CPC problem (5)–(6) do not have eigenvalues simultaneously decreasing in all k groups. In fact, this case is more interesting because the dimensionality reduction based on the standard CPCs is ambiguous. Indeed, dimensionality reduction can be based on those CPCs with the largest total variances, but the ordering of the eigenvalues within the groups is unpredictable: small eigenvalues may be involved in the dimensionality reduction, while large eigenvalues are left out.

Let us consider again the stepwise CPCs obtained as solutions of (14)–(15). In this case, they are different from the standard CPCs of Flury, and do *not* minimize the CPC objective function (5). However, they are worth considering because: first, as before, they appear in order starting with the largest variance explained and followed by the next one in magnitude, and second, the eigenvalues produced tend to decrease simultaneously (or nearly) in all groups.

The stepwise CPCs q_j for $j = 1, 2, \ldots, p$ obtained as solutions of (14)–(15) by construction have the property:

$$\sum_{i=1}^{k} n_i \log(q_1^\top S_i q_1) \geq \sum_{i=1}^{k} n_i \log(q_2^\top S_i q_2) \geq \cdots \geq \sum_{i=1}^{k} n_i \log(q_p^\top S_i q_p). \tag{26}$$

For studying the features of the stepwise CPCs q_i, one relies on the fact that the logarithm is a concave function and satisfies Jensen's inequality:

$$\frac{\sum_{i=1}^{k} w_i \log(x_i)}{\sum_{i=1}^{k} w_i} \leq \log\left(\frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}\right), \tag{27}$$

for any $x_i > 0$ and $w_i \geq 0$. Then, after applying (27) one obtains that for each $j = 1, 2, \ldots, p$:

$$\sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j) \leq n \log\left(q_j^\top \left(\sum_{i=1}^{k} \frac{n_i}{n} S_i\right) q_j\right), \tag{28}$$

where the equality holds if $q_j^\top S_1 q_j = q_j^\top S_2 q_j = \cdots = q_j^\top S_k q_j$. In other words, the maximum of the objective function (14) will be achieved for such a q_j that makes all $q_j^\top S_i q_j$ as similar as possible.

Let $\bar{q}_1 \geq \bar{q}_2 \geq \cdots \geq \bar{q}_p$ denote the eigenvalues of the pooled covariance matrix $\bar{S} = \sum_{i=1}^{k} \frac{n_i}{n} S_i$ from the common space analysis (Krzanowski, 1984; Schott, 1988). Then, the inequality (28) can be rewritten as:

$$\sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j) \leq n \log\left(q_j^\top \bar{S} q_j\right) \leq n \log(\bar{q}_j^\top \bar{S} \bar{q}_j), \tag{29}$$

i.e. the jth objective function (14) is bounded above by n times the logarithm of the jth eigenvalue of $\bar{S}$. Note that, in general, the jth eigenvector $\bar{q}_j$ of $\bar{S}$ does *not* maximize the objective function (14), see also (19).

From inequality (29) and the variational properties of the eigenvalues of $\bar{S}$, one finds that for every $j = 1, 2, \ldots, p-1$:

$$\bar{q}_{j+1}^\top \bar{S} \bar{q}_{j+1} \leq q_j^\top \bar{S} q_j \leq \bar{q}_j^\top \bar{S} \bar{q}_j. \tag{30}$$

Indeed, for, say, q_1 ($q_1^\top q_1 = 1$) the variational properties of the eigenvalues of $\bar{S}$ give:

$$\bar{q}_2^\top \bar{S} \bar{q}_2 = \min_{\substack{q_1^\top q_1 = 1, \\ q_1^\top \bar{q}_1 = 0}} q_1^\top \bar{S} q_1 \leq q_1^\top \bar{S} q_1 \leq \max_{q_1^\top q_1 = 1} q_1^\top \bar{S} q_1 = \bar{q}_1^\top \bar{S} \bar{q}_1. \tag{31}$$

Then, the inequalities (30) imply

$$q_{j+1}^\top \bar{S} q_{j+1} \leq q_j^\top \bar{S} q_j. \tag{32}$$

In other words, the weighted sum $\sum_{i=1}^{k} \frac{n_i}{n} q_j^\top S_i q_j$ of the variances explained by each CPC q_j is not less than that explained by the next CPC q_{j+1}. For an equal number of observations in each group one simply has:

$$\sum_{i=1}^{k} q_1^\top S_i q_1 \geq \sum_{i=1}^{k} q_2^\top S_i q_2 \geq \cdots \geq \sum_{i=1}^{k} q_p^\top S_i q_p, \tag{33}$$

i.e. the total variances explained by the stepwise CPCs are ordered in decreasing order. Moreover, it follows from (28) that the individual eigenvalues $q_j^\top S_1 q_j, \ldots, q_j^\top S_k q_j$ for each $j = 1, 2, \ldots, p$ are made more similar. In other words, if a particular standard CPC produces a small eigenvalue in a certain group, then in the corresponding stepwise CPC this small value will be increased considerably more than the larger ones will be reduced. This usually results in simultaneously decreasing eigenvalues in all groups, or at least eigenvalues not increasing too much in the sense that the "misplaced" eigenvalues are quite close in magnitude. Thus, stepwise CPCs can be naturally used for dimensionality reduction and seem a better alternative than the standard CPCs.

Finally, one can see that in this case when the inequalities (9) do not hold, the problem (14)–(15) can still be obtained as:

$$\min_{\substack{q^\top q = 1 \\ q \perp Q_{j+1}^\perp}} \sum_{i=1}^{k} n_i \log(q^\top S_i q) = \sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j), \tag{34}$$

for $j = 1, 2, \ldots, p$. Thus, the nice variational property of the standard eigenvalues works in this approximate simultaneous diagonalization too. Let the columns of $Q_p = [q_1, \ldots, q_p]$ form a basis in $\mathbb{R}^p$ and assume that for some $q \in \mathbb{R}^p$, $q \perp Q_{j+1}^\perp$ and

$q^\top q = 1$. Then q can be written as $q = \xi_1 q_1 + \cdots + \xi_j q_j$ with $\xi_1^2 + \cdots + \xi_j^2 = 1$. Taking into account the inequalities (26) and (27) it follows that:

$$\begin{aligned}\sum_{i=1}^{k} n_i \log(q^\top S_i q) &= \sum_{i=1}^{k} n_i \log(\xi_1^2 q_1^\top S_i q_1 + \cdots + \xi_j^2 q_j^\top S_i q_j)\\ &\geq \sum_{i=1}^{k} n_i [\xi_1^2 \log(q_1^\top S_i q_1) + \cdots + \xi_j^2 \log(q_j^\top S_i q_j)]\\ &= \xi_1^2 \sum_{i=1}^{k} n_i \log(q_1^\top S_i q_1) + \cdots + \xi_j^2 \sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j)\\ &\geq (\xi_1^2 + \cdots + \xi_j^2) \sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j) = \sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j)\end{aligned}$$

with equality when $q = q_j$, which implies:

$$\min_{\substack{q^\top q = 1 \\ q \perp Q_{j+1}^\perp}} \sum_{i=1}^{k} n_i \log(q^\top S_i q) = \sum_{i=1}^{k} n_i \log(q_j^\top S_i q_j). \tag{35}$$

4. Power method for *k* symmetric matrices

In Section 3.1 it was shown that the first order optimality conditions (18) given in Theorem 3.1 can be rewritten as:

$$\left[\left(n_1 \frac{S_1}{q_1^\top S_1 q_1} + \cdots + n_k \frac{S_k}{q_1^\top S_k q_1}\right) - nI_p\right] q_1 = 0_{p\times 1}, \tag{36}$$

and for $j = 2, \ldots, p$

$$\left[(I_p - Q_{j-1}Q_{j-1}^\top)\left(n_1 \frac{S_1}{q_j^\top S_1 q_j} + \cdots + n_k \frac{S_k}{q_j^\top S_k q_j}\right) - nI_p\right] q_j = 0_{p\times 1}. \tag{37}$$

Eqs. (36)–(37) show that the CPCs can be found by solving sequentially p symmetric eigenvalue problems. The numerical solution of symmetric eigenvalue problems is a very active area of research. There exist a large number of approaches to attack the problem (Golub and Van Loan, 1996, pp. 391–469). In this section the standard power method (Golub and Van Loan, 1996, pp. 406–407) for finding the eigenvalues and eigenvectors of a single symmetric matrix is modified to find CPCs defined by Eqs. (36) and (37). It is well known from the 1950s (Faddeev and Faddeeva, 1963, pp. 430–443) that the power iteration method is a special case of the gradient ascent iterative method for quadratic forms over the unit sphere. For a modern treatment, see Absil et al. (2008, pp. 78–79).

Consider the following algorithm, where for clarity the indexes of the power iterations are given in parentheses:

set $\pi = I_p$
for $j = 1, 2, \ldots, p$
  set $x_{(0)}$ with $x_{(0)}^\top x_{(0)} = 1$
  $x_{(0)} \leftarrow \pi x_{(0)}$
  $\mu_{(0)}^i \leftarrow x_{(0)}^\top S_i x_{(0)}$ and $i = 1, 2, \ldots, k$
  for $l = 1, \ldots, l_{\max}$
    $S \leftarrow \frac{n_1 S_1}{\mu_{(l-1)}^1} + \cdots + \frac{n_k S_k}{\mu_{(l-1)}^k}$
    $y \leftarrow \pi S x_{(l-1)}$
    $x_{(l)} \leftarrow y/\sqrt{y^\top y}$
    $\mu_{(l)}^i \leftarrow x_{(l)}^\top S_i x_{(l)}$ and $i = 1, 2, \ldots, k$
  end
  $q_j \leftarrow x_{(l_{\max})}$ (and $\lambda_j^i \leftarrow \mu_{(l_{\max})}^i$)
  $\pi \leftarrow \pi - q_j q_j^\top$
end

The resulting vectors $q_j, j = 1, 2, \ldots, p$ are the required CPCs. If the eigenvalues of all S's are well separated, very few iterations are needed, say $l_{\max} = 3$. The starting unit vector $x_{(0)}$ is chosen randomly. For faster performance the following

rational start is suggested. Let $\bar{q}_j,\ j = 1, 2, \ldots, p$ be the eigenvectors of the pooled covariance matrix $\bar{S} = \frac{n_1S_1+n_2S_2+\cdots+n_kS_k}{n}$. Then, set $x_{(0)} \equiv \bar{q}_j$ for every consecutive run $j = 1, 2, \ldots, p$ in the above algorithm.

The difference from the classical power method is that the matrix S is updated at each iteration step. Nevertheless, this is a gradient ascent algorithm and the general theory of gradient methods applies (Absil et al., 2008, Ch. 3.6, 4). Particularly, it can be shown that the iterations produced by this modified power algorithm converge and that the CPC objective function (14) is not decreasing with such iterations.

Denote the CPC objective function by

$$f(q) = n_1 \log(q^\top S_1 q) + \cdots + n_k \log(q^\top S_k q) \tag{38}$$

and its gradient by

$$g(q) = \operatorname{grad} f(q) = \frac{n_1 S_1 q}{q^\top S_1 q} + \cdots + \frac{n_k S_k q}{q^\top S_k q}.$$

If x is not a stationary point of the CPCs problem (14)–(15), then x is not a CPC and the optimality condition (36) is not satisfied. Instead, the optimality condition (36) can be written as $g(x)/n = x + t\varepsilon$, for some unit vector ε and $t \geq 0$. As $x^\top g(x) = n$, it follows that ε is orthogonal to x.

At each particular orthogonal subspace, every iteration of the algorithm starts with some unit vector x, and finds the next iteration x_{next} as

$$x_{\text{next}} = \frac{g(x)}{\sqrt{g(x)^\top g(x)}} = \frac{x + t\varepsilon}{\sqrt{(x + t\varepsilon)^\top (x + t\varepsilon)}} = \frac{x + t\varepsilon}{\sqrt{1 + t^2}},$$

which is a well-known retraction on the unit sphere in $\mathbb{R}^p$ (Absil et al., 2008, Ch. 4.1.1).

Lemma 4.1. *The CPC objective function f is nondecreasing along the curve*

$$\gamma : t \to \frac{x + t\varepsilon}{\sqrt{1 + t^2}}$$

starting from $\gamma(0) = x$ at $t = 0$ and moving in direction ε which leads to the next iteration x_{next}.

Proof. The directional derivative of the CPC objective function f along the curve γ at $t = 0$ is:

$$\left.\frac{\mathrm{d}f}{\mathrm{d}t}\right|_{t=0} = \left(\left.\frac{\partial f}{\partial \gamma}\right|_{\gamma(0)=x}\right)^\top \left(\frac{x + t\varepsilon}{\sqrt{1 + t^2}}\right)' \Bigg|_{t=0} = n(x + t\varepsilon)^\top \varepsilon = nt \geq 0. \quad \square$$

The difference between the values of the CPC objective function at two such consecutive iterations can be estimated roughly as follows. For x_{next} close to x

$$\begin{aligned} f(x_{\text{next}}) - f(x) &= g(x)^\top (x_{\text{next}} - x) = n(x^\top + t\varepsilon^\top)\left(\frac{x + t\varepsilon}{\sqrt{1 + t^2}} - x\right) \\ &= n(\sqrt{1 + t^2} - 1) \geq 0, \end{aligned} \tag{39}$$

which indicates again that the CPC objective function is nondecreasing along the iterative steps. The sequence of function values at the iterative steps $f_1, f_2, \ldots, f_l, \ldots$ is nondecreasing and bounded above, and thus converging. Then, it is a Cauchy sequence, which implies that the difference between any two consecutive values f_l and f_{l+1} decreases, and thus, the right hand side of (39) should converge to 0, which can happen only if $t \to 0$.

Lemma 4.2. *Let $x_l, y_l \in \mathbb{R}^p$, such that $x_l^\top x_l \leq 1$ and $y_l^\top y_l \leq 1$. If $x_l^\top y_l \to 1$ when $l \to \infty$, then $(x_l - y_l)^\top (x_l - y_l) \to 0$.*

Proof. From the parallelogram identity

$$(x_l + y_l)^\top (x_l + y_l) + (x_l - y_l)^\top (x_l - y_l) = 2x_l^\top x_l + 2y_l^\top y_l$$

it follows that

$$(x_l - y_l)^\top (x_l - y_l) = x_l^\top x_l + y_l^\top y_l - 2x_l^\top y_l \leq 2 - 2x_l^\top y_l. \quad \square$$

Let x_l be the starting vector x for the lth power iteration and y_l be the corresponding output x_{next}. Then, from

$$x^\top x_{\text{next}} = x^\top \frac{x + t\varepsilon}{\sqrt{1 + t^2}} = \frac{1}{\sqrt{1 + t^2}}$$

and Lemma 4.2, it follows that $(x_l - y_l)^\top (x_l - y_l) = (y_{l-1} - y_l)^\top (y_{l-1} - y_l) \to 0$, as $t \to 0$ according to (39), i.e. the power iterations form a Cauchy sequence in $\mathbb{R}^p$, and thus converge.

Table 1
Eigenvalues for the Bone Data obtained by standard and stepwise CPCs.

Eigenvalue	Standard		Stepwise	
	Males	Females	Males	Females
1	4.54	3.76	4.54	3.76
2	1.08	1.60	1.10	1.57
3	0.12	0.14	0.66	0.41
4	0.68	0.37	0.12	0.13
Total	6.43	5.87	6.43	5.87

Table 2
Eigenvalues for the Real and Forged Bank Notes Data obtained by standard and stepwise CPCs.

Eigenvalue	Standard		Stepwise	
	Real	Forged	Real	Forged
1	0.29	0.12	0.68	1.02
2	0.68	1.02	0.28	0.13
3	0.09	0.12	0.09	0.12
4	0.04	0.03	0.04	0.03
Total	1.09	1.29	1.09	1.29

The magnitude of the error term is given by

$$t^2 = \left(\frac{g(x)}{n} - x\right)^\top \left(\frac{g(x)}{n} - x\right) \geq 0, \tag{40}$$

which indicates that t^2 is bounded (below and above). As $t \to 0$ according to (39), the right hand side of (40) reduces too. Thus, at every iteration step:

$$\left(\frac{g(x_{\text{next}})}{n} - x_{\text{next}}\right)^\top \left(\frac{g(x_{\text{next}})}{n} - x_{\text{next}}\right) \leq \left(\frac{g(x)}{n} - x\right)^\top \left(\frac{g(x)}{n} - x\right).$$

5. Numerical examples and illustrations

Situations where the inequalities (9) do and do not hold are illustrated in the following examples.

Example 1. First we consider the case where the *FG* algorithm produces CPCs with simultaneously decreasing eigenvalues in all groups. Such data sets are Jolicoeur's Bone Data and the Real and Forged Bank Notes Data considered in Flury (1984). The eigenvalues corresponding to the standard and stepwise CPCs are identical as shown in Tables 1 and 2. The standard CPCs can appear in any order, but the pairs of eigenvalues are the same. After they are all found one can reorder them properly. The stepwise CPCs automatically produce simultaneously decreasing eigenvalues in the groups.

The data on the head dimensions of 200 men and 59 women from (Flury, 1988) illustrate the other situation when the eigenvalues of the standard CPCs cannot be ordered simultaneously in decreasing order in all groups. In the first two columns of Table 3 are given the eigenvalues for the groups of men and women obtained by the *FG* algorithm. They differ slightly from the ones published in (Flury (1988) p. 98), but what matters is that they form the same pairs of eigenvalues. The problem is that there is no way to order the eigenvalues for men and women in decreasing order simultaneously. They can be ordered by ordering the sums of the pairs of eigenvalues (variances), which is possible after they are all found. In contrast to the *FG* solution the stepwise eigenvalues automatically appear ordered according to the magnitude of their sums and thus, one can easily reduce simultaneously the dimensionality in both groups. Note also that the variances (the eigenvalues' cumulative sums) explained by the stepwise CPCs are greater than the variances explained by the standard CPCs. Indeed, the first two stepwise CPCs explain 216.82, while the standard CPCs only 212.92.

The stepwise CPCs are different from the CPCs reported in (Flury (1988) p. 98):

$$Q = \begin{pmatrix} 0.48 & 0.35 & 0.55 & -0.52 & 0.27 & -0.02 \\ 0.41 & 0.38 & 0.22 & 0.78 & -0.17 & -0.05 \\ 0.40 & -0.78 & 0.31 & -0.00 & -0.35 & -0.07 \\ 0.16 & -0.35 & -0.07 & 0.29 & 0.88 & -0.01 \\ 0.36 & 0.01 & -0.31 & -0.06 & -0.05 & 0.88 \\ 0.53 & 0.08 & -0.67 & -0.18 & -0.07 & -0.47 \end{pmatrix}.$$

The minimum of the objective function (5) achieved by the *FG* algorithm for the standard CPCs is 4767.48, and for the stepwise CPCs 4774.06.

Table 3
Eigenvalues for men and women in the Head Dimensions Data obtained by standard and stepwise CPCs, their totals across the groups, and the cumulative totals.

Eigenvalue	Standard				Stepwise			
	Men	Women	Total	Ctotal	Men	Women	Total	Ctotal
1	66.33	62.73	129.06	129.06	66.18	63.73	129.91	129.91
2	6.81	17.05	23.86	212.92	34.00	52.91	86.91	216.82
3	13.22	13.13	26.35	290.44	19.50	49.55	69.05	285.87
4	34.25	49.61	83.86	333.78	14.84	32.96	47.80	333.67
5	16.89	26.46	43.35	360.13	13.05	13.79	26.84	360.51
6	16.94	60.57	77.51	383.99	6.87	16.61	23.48	383.99
Total	154.44	229.55	383.99	383.99	154.44	229.55	383.99	383.99

Table 4
Eigenvalues for the Iris data obtained by standard and stepwise CPCs, their totals across the groups, and the cumulative totals.

Eigenvalue	Standard					Stepwise				
	I	II	III	Total	Ctotal	I	II	III	Total	Ctotal
1	48.46	69.22	14.64	132.33	132.33	46.68	64.66	19.08	130.41	130.41
2	5.54	7.53	12.51	25.58	157.91	7.24	13.10	7.87	28.21	158.62
3	7.47	6.71	2.75	16.93	174.84	7.47	6.59	2.76	16.82	175.44
4	1.01	5.36	1.02	7.39	182.24	1.09	4.49	1.21	6.79	182.24
Total	62.48	88.84	30.92	182.24	182.24	62.48	88.84	30.92	182.24	182.24

Example 2. The stepwise CPCs obtained for Fisher's Iris Data are:

$$Q = \begin{pmatrix} 0.75 & -0.09 & 0.63 & 0.20 \\ 0.44 & 0.79 & -0.33 & -0.26 \\ 0.47 & -0.60 & -0.54 & -0.34 \\ 0.15 & 0.02 & -0.45 & 0.88 \end{pmatrix},$$

and the corresponding eigenvalues obtained as $\mathrm{diag}(Q^\top S_i Q)$ for $i = 1, 2, 3$ are collected in the following column vectors:

$$\begin{pmatrix} 46.68 & 64.66 & 19.08 \\ 7.24 & 13.10 & 7.87 \\ 7.47 & 6.59 & 2.76 \\ 1.09 & 4.49 & 1.21 \end{pmatrix}.$$

These eigenvalues are quite different from the standard ones for the *Iris* data as is evident from Table 4. In the first triple (row) of eigenvalues the larger ones are decreased while the small eigenvalue in the third group is considerably increased. A similar trend is noticeable in the second and the last triples of eigenvalues: the smaller values are increased and the largest eigenvalue is decreased. The third triple of eigenvalues is barely changed. As a result, the total variance explained by the first two stepwise CPCs is higher than that explained by the standard CPCs.

The value of the original CPC objective function (5) for these stepwise CPCs is 1189.25, and the minimum achieved by Flury's CPCs is 1161.18. The first two stepwise CPCs are very similar to those obtained by the common space analysis of $\bar{S}$ (Krzanowski, 1984; Schott, 1988).

Example 3. The training vowel recognition data set is constructed from speaker independent recognition of the 11 steady state vowels of British English. The words (heed, hid, head, had, hard, hud, hod, hoard, hood, who'd, heard) were uttered by each of eight speakers, four male and four female. The sample space is ten-dimensional (for each utterance, there are ten floating-point input values), and the training data set contains 528 ($11 \times 8 \times 6$) records, as each vowel is represented as a six-dimensional vector. The data are available from http://archive.ics.uci.edu/ml/datasets.html under the name: *Connectionist Bench (Vowel Recognition—Deterding Data)*.

The standard deviations are reasonably similar in magnitude: between 0.4793 and 1.1610. Thus, eleven 10×10 covariance matrices are constructed for the vowels each based on 48 records. Then, their CPCs are obtained using the standard and the stepwise approaches. The corresponding eigenvalues for each group (vowel) are depicted as bar diagrams in Fig. 2. The eigenvalues of the standard CPCs (upper panel) appear completely at random within each group. In contrast, the eigenvalues of the stepwise CPCs (bottom panel) are, in general, in decreasing order within each group. In 7 groups the first eigenvalue is the largest, and in 7 groups the first two eigenvalues are largest. The standard CPCs require 1.56 s of CPU time, while the stepwise CPCs require only 0.05 s, even when all 10 CPCs are found.

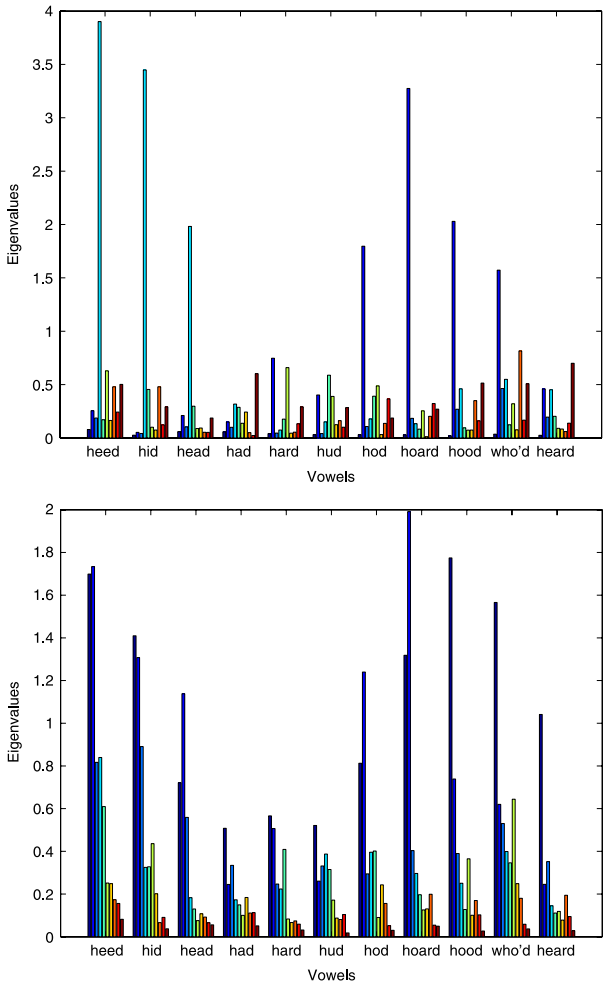

Fig. 2. Eigenvalues for *vowel* training data obtained by standard and stepwise CPCs (bottom panel).

6. Concluding remarks

Stepwise CPCs related to partial CPCs (Flury, 1988; Schott, 1999) will be considered elsewhere, as well as the possible relation of the stepwise CPCs to the population quantities of the original CPC model. The proposed stepwise procedure may also be applied to robust scatter matrices as an alternative to the approaches studied by Boente et al. (2006) and Schyns et al. (2010).

Acknowledgements

The author thanks the reviewers and Prof. Chris Jones for their careful reading of the manuscript and helpful comments.

References

Absil, P.-A., Mahony, R., Sepulchre, R., 2008. Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton.
Boente, G., Pires, A.M., Rodrigues, I.M., 2006. General projection-pursuit estimators for the common principal components model: influence funtions and Monte Carlo study. Journal of Multivariate Analysis 97, 124–147.
Faddeev, D.K., Faddeeva, V.N., 1963. Computational Methods of Linear Algebra. W.H. Freeman and Company, San Francisco.
Flury, B., 1984. Common principal components in *k* groups. Journal of the American Statistical Association 79, 892–898.
Flury, B., 1988. Common Principal Components and Related Multivariate Models. John Wiley & Sons, New York.
Golub, G.H., Van Loan, Ch.F., 1996. Matrix Computations, 3rd ed. The John Hopkins University Press, Baltimore, London.
Horn, R.A., Johnson, C.R., 1986. Matrix Analysis. Cambridge University Press, Cambridge, UK.
Jolliffe, I.T., 2002. Principal Component Analysis, 2nd ed. Springer-Verlag, New York.
Krzanowski, W.J., 1979. Between-groups comparison of principal components. Journal of the American Statistical Association 74, 703–707.
Krzanowski, W.J., 1984. Principal component analysis in the presence of group structure. Applied Statistics 33, 164–168.
Schott, J.R., 1988. Common principal component subspaces in two groups. Biometrika 75, 229–236.
Schott, J.R., 1999. Partial common principal component subspaces. Biometrika 86, 899–908.

Schyns, M., Haesbroeck, G., Critchley, F., 2010. RelaxMCD: smooth optimisation for the minimum covariance determinant estimator. Computational Statistics and Data Analysis 54, 843–857.
Trendafilov, N.T., 2006. The dynamical system approach to multivariate data analysis, a review. Journal of Computational and Graphical Statistics 15, 628–650.
Trendafilov, N.T., Jolliffe, I.T., 2006. Projected gradient approach to the numerical solution of the SCoTLASS. Computational Statistics and Data Analysis 50, 242–253.