Technical Report: Criteria for Evaluating Dimension-Reducing Components for Multivariate Data

Daniel Gervini and Valentin Rousson [1]

December 17, 2003

[1]The authors are Postdoctorates, Department of Biostatistics, University of Zürich, Sumatrastrasse 30, CH-8006 Switzerland (E-mail: gervini@ifspm.unizh.ch; rousson@ifspm.unizh.ch). Their research was supported by the Swiss National Science Foundation (grants BE 20.63579.00 and 3200-064047.00/1). The authors thank Prof. Theo Gasser for many fruitful discussions, the referees for their constructive criticism, and the editor and an associate editor for suggestions that improved the presentation.

Abstract

Principal components are the benchmark for linear dimension reduction, but they are not always easy to interpret. For this reason, some alternatives have been proposed in recent years. These methods produce components that, unlike principal components, are correlated and/or have non-orthogonal loadings. We show in this article that the criteria commonly used to evaluate principal components are not adequate for evaluating such components, and propose two new criteria that are more suitable for this purpose.

KEY WORDS: Linear prediction; Principal components; Rotated components; Simple components.

1 Introduction

In multivariate datasets, it is often the case that the variables are highly correlated and provide redundant information. The number of variables is then unnecessarily large, and essentially the same information can be conveyed by fewer dimensions if the variables are wisely combined. In many cases, a lower dimension also helps to visualize patterns in the data that would otherwise go unnoticed.

For these reasons, dimension reduction techniques have played an important role in multivariate analysis. Many of these techniques construct a system of q variables which are linear combinations of the original p variables; the new variables are called components. The most popular of these methods is principal component analysis, which was originally proposed by Hotelling (1933) —a comprehensive and up-to-date reference is Jolliffe (2002). The idea is to sequentially construct a system of components which are uncorrelated and have maximal variance. The component coefficients (called loadings) obtained in this way turn out to be orthogonal. For a given target dimension q, the first q principal components are then the optimal dimension-reducing system because they extract the maximal variability, and they are both statistically non-redundant (uncorrelated) and geometrically non-redundant (orthogonal loadings). For this reason, they are considered the benchmark for linear dimension reduction. Principal components are optimal under different criteria that evaluate uncorrelated or orthogonal components (Rao 1964; Okamoto 1969; McCabe 1984).

In most cases, however, the researcher not only wants to reduce the dimension of the dataset but also wants to obtain components that are interpretable in the context of his research. Unfortunately, principal component loadings sometimes show complicated patterns and are not easy to interpret —see Cadima and Jolliffe (1995) for interesting examples. To improve interpretability, alternative methods that produce components with simpler loadings' patterns have been proposed over

Table 1: Covariance–correlation Matrix of Hearing Loss Data.

	Left Ear				Right Ear			
	500	1000	2000	4000	500	1000	2000	4000
L, 500	41.07	(.78)	(.40)	(.26)	(.70)	(.64)	(.24)	(.20)
L, 1000	37.73	57.32	(.54)	(.27)	(.55)	(.71)	(.36)	(.22)
L, 2000	28.13	44.44	119.70	(.42)	(.24)	(.45)	(.70)	(.33)
L, 4000	32.10	40.83	91.21	384.78	(.18)	(.26)	(.32)	(.71)
R, 500	31.79	29.75	18.64	25.01	50.75	(.66)	(.16)	(.13)
R, 1000	26.30	34.24	31.21	33.03	30.23	40.92	(.41)	(.22)
R, 2000	14.12	25.30	71.26	57.67	10.52	24.62	86.30	(.37)
R, 4000	25.28	31.74	68.99	269.12	18.19	27.22	67.26	373.66

NOTE: Correlations are given in parenthesis.

the years —for example Neuhaus and Wrigley (1954), Kaiser (1958), Hausmann (1982), Kiers (1991), Jolliffe and Uddin (2000), Vines (2000) and Rousson and Gasser (2003); see also Jolliffe (2002, chap. 11) for other proposals. These methods produce components that are no longer uncorrelated and have non-orthogonal loadings, so they are less efficient than principal components at dimension reduction. But how can we quantify precisely the loss of dimension-reducing efficiency of such components? How can we compare the performance of different methods? Remember that most of the optimality criteria mentioned in the preceding paragraph assume that the components are either uncorrelated or have orthogonal loadings, which is no longer the case with these alternative components.

To illustrate the problem with a real dataset, let us consider the audiometric example analyzed in Jackson (1991; chap. 5) and reanalyzed by Vines (2000). The data consists of measurements of lower hearing threshold on 100 men. Observations were obtained, on each ear, at frequencies 500, 1000, 2000 and 4000 Hz, so that eight variables were recorded for each individual. The sample variance of measurements at 4000 Hz turned out to be about 9 times higher than those at

500 Hz, so the variables were standardized before computing the principal components —that is, the principal components were computed on the correlation matrix rather than the covariance matrix. The covariance–correlation matrix is given in Table 1. Jackson (1991) argues that the first four principal components, which explain 87% of the total variability, provide a good approximation to the data. The principal component loadings are given in Table 2. They can be interpreted as follows: the first component is an indicator of average hearing loss, the second one is a contrast between high and low frequency hearing loss, the third one is a contrast between hearing loss at the two highest frequencies, and the fourth one is a contrast between the two ears. Note that some of the components (the third one in particular) are somewhat difficult to interpret at first glance, because some loadings are relatively small but not really close to zero, so it is not clear whether they are significant or not. More clear-cut loadings are obtained with the methods of Vines (2000) and Rousson and Gasser (2003), which produce the same components for this dataset; the loadings are given in Table 2. The simplicity of the loadings' patterns allows unequivocal interpretation of these components, but before deciding to use this system instead of the principal components, the statistician should know how much is lost in terms of dimension-reducing power. The simpler system is no longer uncorrelated, so it does not make sense to simply add up the variances and compare it with the total variance of the original dataset, as it is done with the principal components. What Vines (2000) does is to compare the variance of each simple component with the variance of the corresponding principal component, concluding that "little explanatory power (in terms of variance) is lost by this radical simplification. Furthermore the highest correlation between the first four simple components is only 0.151" (Vines 2000, p. 448.) This is rather vague, however; it would be nice to have a criterion that indicates unambiguously, with a single number (ranging from 0 to 1, say, with 1 being optimal)

Table 2: Component Loadings for Hearing Loss Data.

	Principal components				Simple components				Varimax components			
Variable	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
L, 500	.40	-.32	.16	-.33	.35	-.35	.00	-.35	.60	.03	-.09	.15
L, 1000	.42	-.23	-.05	-.48	.35	-.35	.00	-.35	.67	-.03	.11	-.03
L, 2000	.37	.24	-.47	-.28	.35	.35	-.50	-.35	.29	.02	.61	-.19
L, 4000	.28	.47	.43	-.16	.35	.35	.50	-.35	.13	.70	-.01	-.10
R, 500	.34	-.39	.26	.49	.35	-.35	.00	.35	.03	.02	-.16	.74
R, 1000	.41	-.23	-.03	.37	.35	-.35	.00	.35	.07	-.02	.15	.58
R, 2000	.31	.32	-.56	.39	.35	.35	-.50	.35	-.26	-.01	.75	.21
R, 4000	.25	.51	.43	.16	.35	.35	.50	.35	-.13	.71	.02	.09

the dimension-reducing power of a system of components.

Some of the authors mentioned earlier (Neuhaus and Wrigley 1954; Kaiser 1958; Kiers 1991; Jolliffe and Uddin 2000) have already proposed non-standard criteria to evaluate components. The problem is that these authors aim at simplicity, so they propose criteria that measure some sort of simplicity of the system rather than its dimension-reducing power. As a result, it does not make much sense to use, for example, the quartimax criterion of Neuhaus and Wrigley (1954) to evaluate varimax components, which by definition maximize the different simplicity criterion of Kaiser (1958). Consider, for example, the varimax components for the Hearing Loss example, given in Table 2. They are harder to interpret than either the principal components or the simple system, and are highly correlated. Varimax components offer only disadvantages in this example, yet they are considered optimal (by definition) under the criterion of Kaiser (1958), while principal and simple components are considered suboptimal. It is clear, then, that we need criteria that evaluate the dimension-reducing power of components regardless of the notion of simplicity that the researcher prefers. In this article we are going to review some of the existing criteria and propose two new ones, because we found that none of the existing criteria is completely adequate for this task.

2 Necessary and desirable properties of criteria for evaluating components

From the discussion in Section 1, we conclude that criteria for evaluating components should assign optimal value to the principal components, since they are the most efficient dimension-reducing system in terms of variability extraction and non-redundancy of information. These criteria should also be applicable to systems of components that may not be uncorrelated and may not have orthogonal loadings, because most alternatives to principal components do not. For example, both simple and varimax components for the Hearing Loss example are correlated, so we cannot simply add up the variances of the components and divide it by the sum of variances of the principal components; a more elaborate criterion, that takes correlations into account, is necessary. Specifically, correlations between components should be penalized, because they imply redundancy of information.

A criterion for evaluating dimension-reducing components, then, should satisfy at least two conditions:

1. *Generality*. The criterion has to be applicable to a broad range of components, with the only restriction of unit-norm and linearly independent loadings —the least restrictive assumptions that rule out artificial cases. Under these general conditions, and for a given target dimension q, the criterion must be maximized by the first q principal components.

2. *Uniqueness.* The criterion must be maximized *only* by the principal components under the conditions mentioned above.

The Uniqueness condition might seem too strong, but it guarantees that correlations between components and deviations from orthogonality are penalized.

Other properties may be useful, even desirable, but we do not think that they are strictly necessary. For instance:

- *Additivity*. Many criteria can be naturally expressed as a sum of q terms, indicating the contribution of each component towards the overall dimension reduction. This is a good thing, but we are mainly interested in evaluating systems of components as a whole, rather than individual contributions of the components.
- *Invariance under permutation of components*. Since we are evaluating systems as a whole, a criterion that assigns different values to two systems of components which are just a permutation of one another is not very appealing; therefore, permutation invariance is desirable. In practice, however, the components are computed in a sequential way, so that a natural ordering is given by construction and alternatives consisting merely on permutations are normally not contemplated.

The next Section reviews some existing criteria, focusing on those that satisfy the property of Generality. It turns out that none of them satisfies the property of Uniqueness. This motivates our introduction in Section 4 of two new criteria that satisfy both properties.

3 Existing criteria

Before we start reviewing the existing criteria, let us introduce some notation. Consider a random vector $\mathbf{x} \in \mathbb{R}^p$, that without loss of generality will be assumed to have zero mean. A linear dimension reduction technique will produce a system of components $\mathbf{y} = \mathbf{A}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbb{R}^{p\times q}$ is called the loading matrix and $q \le p$. The principal components are defined as follows. Let $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{x})$ and $\mathbf{\Sigma} = \mathbf{\Gamma}^\top \mathbf{\Lambda}\mathbf{\Gamma}$ be the eigenvalue decomposition of $\mathbf{\Sigma}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \ge \ldots \ge \lambda_p > 0$ and $\mathbf{\Gamma} \in \mathcal{O}(p)$, where $\mathcal{O}(p)$ denotes the family of $p \times p$ orthogonal matrices ($\mathcal{O}(p, q)$ will denote the family of $p \times q$ orthogonal matrices). The elements of $\mathbf{z} = \mathbf{\Gamma}^\top \mathbf{x}$ are called the principal components of $\mathbf{x}$. Note that $\mathrm{cov}(\mathbf{z}) = \mathbf{\Lambda}$, so that the principal components are uncorrelated. $\mathbf{\Gamma}_q$ will indicate the loading matrix consisting of the first q columns of $\mathbf{\Gamma}$, and $\mathbf{\Lambda}_q = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$. Note that although the components $\mathbf{z}$ are unique, the loading matrix $\mathbf{\Gamma}$, and consequently $\mathbf{\Gamma}_q$, is only determined up to column sign reversal and exchange of columns with identical eigenvalues, so it is not unique. But to avoid unnecessary complications in phraseology, we will talk about "the unique" matrix $\mathbf{\Gamma}_q$.

There are essentially three approaches to dimension reduction: prediction (find $\mathbf{y} = \mathbf{A}^\top \mathbf{x}$ that provides the best linear prediction of $\mathbf{x}$), variability maximization (find $\mathbf{y}$ with the largest possible variance among linearly independent combinations of $\mathbf{x}$), and correlation (find $\mathbf{y}$ that is maximally correlated with $\mathbf{x}$). The review that follows is organized in three subsections corresponding to these approaches.

3.1 Prediction Approach

The best linear predictor of $\mathbf{x}$ based on $\mathbf{y} = \mathbf{A}^\top \mathbf{x}$, in the sense of minimizing $E(\|\mathbf{x} - \mathbf{B}\mathbf{y}\|^2)$, is $\widehat{\mathbf{B}}\mathbf{y}$ with $\widehat{\mathbf{B}} = \mathbf{\Sigma}\mathbf{A}(\mathbf{A}^\top \mathbf{\Sigma}\mathbf{A})^{-1}$. Therefore, the matrix $\mathbf{A}$ producing the optimal predictor is $\widehat{\mathbf{A}}$ that minimizes

$$E(\|\mathbf{x} - \mathbf{\Sigma}\mathbf{A}(\mathbf{A}^\top \mathbf{\Sigma}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{x}\|^2) = \mathrm{tr}(\mathbf{\Sigma}) - \mathrm{tr}(\mathbf{\Sigma}\mathbf{A}(\mathbf{A}^\top \mathbf{\Sigma}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{\Sigma}),$$

or equivalently, $\widehat{\mathbf{A}}$ that maximizes $\operatorname{tr}(\mathbf{\Sigma A}(\mathbf{A}^\top\mathbf{\Sigma A})^{-1}\mathbf{A}^\top\mathbf{\Sigma})$. Rao (1964) shows that $\widehat{\mathbf{A}} = \mathbf{\Gamma}_q$ is the maximizer in $\mathcal{O}(p,q)$, and the maximum is $\operatorname{tr}(\mathbf{\Lambda}_q) = \sum_{k=1}^q \lambda_k$. Then our first criterion is

$$\operatorname{BLP}(\mathbf{A}) = \frac{\operatorname{tr}(\mathbf{\Sigma A}(\mathbf{A}^\top\mathbf{\Sigma A})^{-1}\mathbf{A}^\top\mathbf{\Sigma})}{\sum_{k=1}^q \lambda_k}, \tag{1}$$

where BLP stands for "best linear prediction" (all criteria in this paper is standardized so that the optimum is 1). This can be rewritten as an additive criterion if one so wishes.

The value of BLP only depends on the subspace spanned by the columns of $\mathbf{A}$, rather than on the actual matrix $\mathbf{A}$. Consequently, any full-rank transformation of $\mathbf{A}$ is equivalent for this criterion. In the Appendix we prove that any matrix of the form $\widehat{\mathbf{A}} = \mathbf{\Gamma}_q\mathbf{C}$ with non-singular $\mathbf{C} \in \mathbb{R}^{q\times q}$ maximizes (1). Therefore BLP satisfies the property of Generality, but not Uniqueness; it fails to discriminate between systems of components that are obviously not equivalent from a practical point of view. Therefore, this criterion is not adequate for our purposes.

3.2 Variability Maximization Approach

Finding a system of uncorrelated components with largest variance is probably the most familiar approach to dimension reduction. The q components with largest variance are the ones that carry most of the information of the original data, while the others vary little about zero. In fact, if $\mathbf{x}$ lies on a q-dimensional subspace of $\mathbb{R}^p$ with probability 1, then the variance of the last $p-q$ components is exactly zero. The total variance of the system can be defined as either the trace or the determinant of the covariance matrix (the latter is usually known as generalized variance). Using $\operatorname{tr}(\operatorname{cov}(\mathbf{y})) = \operatorname{tr}(\mathbf{A}^\top\mathbf{\Sigma A})$ is more common, and if the components are assumed to be uncorrelated, the maximization can be carried out in a sequential way, by maximizing $\mathrm{var}(y_j) = \mathbf{a}_j^\top \boldsymbol{\Sigma} \mathbf{a}_j$ subject to the restrictions $\|\mathbf{a}_j\| = 1$ and $\mathrm{cov}(y_j, y_k) = \mathbf{a}_j^\top \boldsymbol{\Sigma} \mathbf{a}_k = 0$ for all $k < j$. The optimal $\mathbf{y}$ turns out to be the vector of the first q principal components. It is interesting to note that the optimal loading matrix comes out orthogonal, although this was not an original restriction. If one imposes the restriction $\mathbf{A} \in \mathcal{O}(p, q)$ from the beginning (but not uncorrelation) then the maximizers of $\mathrm{tr}(\mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A})$ turn out to be the matrices of the form $\boldsymbol{\Gamma}_q \mathbf{R}$ with $\mathbf{R} \in \mathcal{O}(q)$, that is, the rotations of the first q principal components. It is clear, however, that maximizing the trace only makes sense when either one of the restrictions of uncorrelation or orthogonality is imposed, which violates the property of Generality.

On the other hand, the generalized variance $\det(\mathrm{cov}(\mathbf{y}))$ is maximized by $\boldsymbol{\Gamma}_q$ under the unique restriction of unit-norm loadings, without assuming orthogonality or uncorrelation (see Okamoto 1969). Then the criterion

$$\mathrm{GV}(\mathbf{A}) = \left(\frac{\det(\mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A})}{\prod_{k=1}^{q} \lambda_k} \right)^{\frac{1}{q}}$$

satisfies the property of Generality, in contrast with the trace criterion. But unfortunately GV is invariant under rotation of the components and then it does not satisfy the property of Uniqueness. This implies, for example, that GV cannot discriminate between principal components and the varimax rotation. Although GV is more informative than BLP, it is still not good enough for our purposes.

3.3 Correlation Approach

The third approach to dimension reduction consists in finding components that are maximally correlated with the data, using measures of matrix correlation based on the sample data matrix. Given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, let $\mathbf{X}$ be the $n \times p$ data matrix and $\mathbf{Y} = \mathbf{X}\mathbf{A}$ be the $n \times q$ matrix of components. Robert and Escoufier

(1976) measure the closeness between "data configurations" $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{Y}\mathbf{Y}^\top$ as $\operatorname{corr}(\mathbf{X}\mathbf{X}^\top, \mathbf{Y}\mathbf{Y}^\top)$, where $\operatorname{corr}(\mathbf{A}, \mathbf{B})$ is the inner-product matrix correlation $\langle \mathbf{A}, \mathbf{B}\rangle / (\langle \mathbf{A}, \mathbf{A}\rangle^{\frac{1}{2}} \langle \mathbf{B}, \mathbf{B}\rangle^{\frac{1}{2}})$ with $\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{tr}(\mathbf{A}^\top \mathbf{B})$. If $\mathbf{S} = \mathbf{X}^\top \mathbf{X}/n$ denotes the sample covariance matrix, it is not difficult to see that

$$\operatorname{corr}(\mathbf{X}\mathbf{X}^\top, \mathbf{Y}\mathbf{Y}^\top) = \frac{\operatorname{tr}(\mathbf{S}\mathbf{A}\mathbf{A}^\top \mathbf{S})}{\{\operatorname{tr}(\mathbf{S}^2) \operatorname{tr}((\mathbf{A}^\top \mathbf{S}\mathbf{A})^2)\}^{\frac{1}{2}}}. \tag{2}$$

This is known as the RV–coefficient of Robert and Escoufier (1976), who showed that (2) is uniquely maximized by the first q sample principal components among all uncorrelated components. It is possible to relax the restriction of uncorrelation at the price of losing uniqueness (this trade-off is unavoidable, since $\operatorname{corr}(\mathbf{X}\mathbf{X}^\top, \mathbf{Y}\mathbf{Y}^\top)$ is invariant under rotation of components). Assuming only that the loadings have norm one, all the maximizers of $\operatorname{corr}(\mathbf{X}\mathbf{X}^\top, \mathbf{Y}\mathbf{Y}^\top)$ are the rotations of the first q principal components (the proof is given in the Appendix). Therefore, replacing the sample covariance matrix by the population covariance matrix, from (2) we deduce a criterion that satisfies the property of Generality, but not Uniqueness:

$$\mathrm{RV}(\mathbf{A}) = \frac{\operatorname{tr}(\mathbf{A}^\top \boldsymbol{\Sigma}^2 \mathbf{A})}{\{\sum_{j=1}^{q} \lambda_j^2\}^{\frac{1}{2}} \{\operatorname{tr}((\mathbf{A}^\top \boldsymbol{\Sigma}\mathbf{A})^2)\}^{\frac{1}{2}}}.$$

Cadima and Jolliffe (2001) consider two measures of correlation, which they use for variable selection based on principal components. One is the matrix correlation between the data matrix $\mathbf{X}$ and its projection on the components, $\mathbf{P}_\mathbf{Y}\mathbf{X}$, where $\mathbf{P}_\mathbf{Y} = \mathbf{Y}(\mathbf{Y}^\top \mathbf{Y})^{-1}\mathbf{Y}^\top$. After some algebra we have that $\operatorname{corr}(\mathbf{X}, \mathbf{P}_\mathbf{Y}\mathbf{X}) = \{\operatorname{tr}(\mathbf{S}\mathbf{A}(\mathbf{A}^\top \mathbf{S}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{S}) / \operatorname{tr}(\mathbf{S})\}^{\frac{1}{2}}$. Clearly this is equivalent to (1) after replacing $\mathbf{S}$ with $\boldsymbol{\Sigma}$, so we do not arrive at any new criterion in this way.

The second measure proposed by Cadima and Jolliffe is the matrix correlation between projectors $\mathbf{P}_\mathbf{Y}$ and $\mathbf{P}_\mathbf{Z}$, where $\mathbf{Z} = \mathbf{X}\mathbf{G}_q$ are the first q sample principal components ($\mathbf{G} \in \mathcal{O}(p)$ is the matrix of eigenvectors of $\mathbf{S}$, and $\mathbf{L}$ will denote the diagonal matrix of eigenvalues). After some algebra we obtain

$$\mathrm{corr}(\mathbf{P}_{\mathbf{Y}}, \mathbf{P}_{\mathbf{Z}}) = \frac{\mathrm{tr}((\mathbf{A}^\top \mathbf{S}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{S}_q \mathbf{A})}{q},$$

where $\mathbf{S}_q = \mathbf{G}_q \mathbf{L}_q \mathbf{G}_q^\top$. Replacing sample quantities $\mathbf{S}$ and $\mathbf{S}_q$ by their population equivalents, we have a new criterion

$$\mathrm{CP}(\mathbf{A}) = \frac{\mathrm{tr}((\mathbf{A}^\top \boldsymbol{\Sigma}\mathbf{A})^{-1}\mathbf{A}^\top \boldsymbol{\Sigma}_q \mathbf{A})}{q},$$

where CP stands for "correlation of projectors". Note that CP is close (although not exactly equivalent) to BLP, and then it shares all its advantages and disadvantages. In particular, the maximizers are all the matrices of the form $\widehat{\mathbf{A}} = \boldsymbol{\Gamma}_q \mathbf{C}$ with invertible $\mathbf{C} \in \mathbb{R}^{q \times q}$ (the proof is similar to that of (1) and hence omitted).

4 New criteria

We saw in the previous section that the existing criteria do not satisfy the condition of Uniqueness. None of them can discriminate between rotations of principal components, and BLP does not even discriminate between arbitrary full-rank transformations. In this section we propose two new criteria that satisfy the Uniqueness property.

Our first proposal is, essentially, a sum of variances corrected for correlations. The idea is that if a new component $y_k = \mathbf{a}_k^\top \mathbf{x}$ is added to a system of $k-1$ components, an indicator of the *real* contribution of y_k to the total variance of the system is the residual variance of the linear prediction of y_k given the first $k-1$ components. Adding all these residual variances together gives

$$\sum_{k=1}^{q} (\mathbf{a}_k^\top \boldsymbol{\Sigma}\mathbf{a}_k - \mathbf{a}_k^\top \boldsymbol{\Sigma}\mathbf{A}_{(k-1)}(\mathbf{A}_{(k-1)}^\top \boldsymbol{\Sigma}\mathbf{A}_{(k-1)})^{-1}\mathbf{A}_{(k-1)}^\top \boldsymbol{\Sigma}\mathbf{a}_k), \quad (3)$$

where $\mathbf{A}_{(k)} = (\mathbf{a}_1, \ldots, \mathbf{a}_k)$. Note that (3) is just the sum of variances of the components if they are uncorrelated (because $\mathbf{A}_{(k-1)}^\top \mathbf{\Sigma} \mathbf{a}_k = 0$), otherwise it is strictly smaller. Therefore (3) penalizes correlations, as we wanted. Moreover, the unique maximizer of (3) among full-rank matrices with unit-norm columns is $\mathbf{\Gamma}_q$ (see the Appendix for a proof). Then the criterion

$$\mathrm{CSV}(\mathbf{A}) = \frac{\sum_{k=1}^q (\mathbf{a}_k^\top \mathbf{\Sigma} \mathbf{a}_k - \mathbf{a}_k^\top \mathbf{\Sigma} \mathbf{A}_{(k-1)} (\mathbf{A}_{(k-1)}^\top \mathbf{\Sigma} \mathbf{A}_{(k-1)})^{-1} \mathbf{A}_{(k-1)}^\top \mathbf{\Sigma} \mathbf{a}_k)}{\sum_{k=1}^q \lambda_k},$$

where CSV stands for "corrected sum of variances", satisfies both the Generality and the Uniqueness properties (and is also additive).

However, CSV is not invariant under permutation of components. In practice this is not very problematic, but it is not hard to construct an invariant criterion if one wishes to. One possibility is to simply take the maximum of CSV among all permutations of the components. Another possibility is to define a "symmetrically corrected sum of variances"

$$\mathrm{SCSV}(\mathbf{A}) = \frac{\sum_{k=1}^q (\mathbf{a}_k^\top \mathbf{\Sigma} \mathbf{a}_k - \mathbf{a}_k^\top \mathbf{\Sigma} \mathbf{A}_{-k} (\mathbf{A}_{-k}^\top \mathbf{\Sigma} \mathbf{A}_{-k})^{-1} \mathbf{A}_{-k}^\top \mathbf{\Sigma} \mathbf{a}_k)}{\sum_{k=1}^q \lambda_k}, \qquad (4)$$

where $\mathbf{A}_{-k}$ is the $p \times (q-1)$ matrix obtained after deleting the k-th column of $\mathbf{A}$. The numerator of (4) is the sum of the residual variances of the linear predictors of y_k given the other $q-1$ components. Note that $\mathrm{SCSV}(\mathbf{A}) = \mathrm{CSV}(\mathbf{A})$ if the system is uncorrelated and $\mathrm{SCSV}(\mathbf{A}) < \mathrm{CSV}(\mathbf{A})$ otherwise. Then, SCSV is also uniquely maximized by $\mathbf{\Gamma}_q$ among full-rank matrices with unit-norm columns. This criterion also satisfies the properties of Generality and Uniqueness, plus additivity and invariance under permutation of components. But it penalizes correlations more strongly than CSV and then it can be overly pessimistic in some situations. Besides, it is not a sequential criterion: it "looks into the future", subtracting from $\mathrm{var}(y_k)$ correlations with components y_j with $j > k$. It must be noted, however, that the properties of invariance under permutation of components and sequentiality are at odds with each other (except, of course, in the case of uncorrelated components, where one just takes the sum of variances). At this point we cannot envisage a criterion that is simultaneously permutation invariant, sequential, and penalizes correlations so as to satisfy the Uniqueness property. But in practice sequentiality seems to be preferable over permutation invariance, so that we tend to favor the CSV criterion.

5 Examples

5.1 Toy example

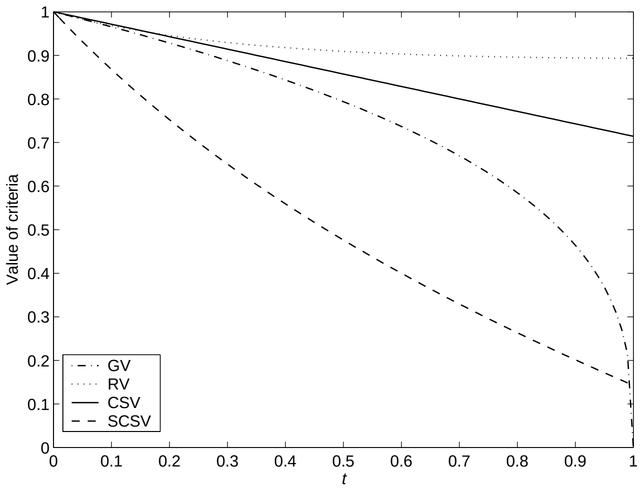
Consider $\mathbf{\Sigma} = \mathbf{\Lambda} = \mathrm{diag}(4, 2, 1, 1/2)$, $q = 3$, and the family of loading matrices

$$\mathbf{A}(t) = \begin{pmatrix} 1 & \sqrt{t} & 0 \\ 0 & \sqrt{1-t} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \; t \in [0, 1].$$

The optimal solution $\mathbf{\Gamma}_3$ corresponds to $t = 0$. For $t > 0$ the correlation between the first two components increases with t, and $t = 1$ corresponds to a degenerate system of rank two. After straightforward computation we obtain: $\mathrm{BLP}(t) = 1$ if $t \neq 1$, $\mathrm{CP}(t) = 1$ if $t \neq 1$, $\mathrm{GV}(t) = (1-t)^{\frac{1}{3}}$, $\mathrm{RV}(t) = (21+12t)/\sqrt{21}\sqrt{21+40t+4t^2}$, $\mathrm{CSV}(t) = \{4+2(1-t)+1\}/7$ and $\mathrm{SCSV}(t) = [\{4-8t/(1+t)\}+2(1-t)+1]/7$ (for better insight we wrote out the three terms of CSV and SCSV, instead of giving more compact expressions).

The non-constant criteria are plotted in Figure 1. A comparison reveals that BLP and CP are not informative at all: they assign 100% optimality to all sys-

Figure 1: The four non-constant optimality criteria for the Toy Example, plotted as functions of t.



tems with $t < 1$, because $\mathbf{A}(t)$ spans the same subspace as $\mathbf{\Gamma}_3$ for such t's. The other four criteria judiciously downweight correlations, although they do so with different strength. GV is close to CSV for moderate values of t, but decreases quickly to zero as the system approaches degeneracy. For this criterion, a single redundant component outweighs the benefits of all the other components; this penalty for correlations seems to be too strong. In contrast, RV penalizes correlations so timidly that all systems are at least 89% optimal. On the other hand, the new criterion CSV is tougher than RV for penalizing correlations but not as tough as SCSV. The reason is that CSV looks at the system in a sequential way and downweights the second component only, while SCSV considers the system as a whole and penalizes all the terms involving correlated components. In this example, for the extreme case $t = 1$ criterion SCSV does not acknowledge any contribution from the first two components. It is true that SCSV, unlike GV, does not downweight the third component and then it does acknowledge some merit of $\mathbf{A}(1)$. But all things considered, we feel that SCSV penalizes correlations too

Table 3: Covariance-correlation Matrices for Components of Hearing Loss Data.

Principal components				Simple components				Varimax components			
3.93	0	0	0	3.86	(-.12)	(-.09)	(-.14)	1.85	(.27)	(.42)	(.73)
0	1.62	0	0	-.30	1.59	(.15)	(.05)	.49	1.71	(.41)	(.24)
0	0	.98	0	-.18	.19	.99	(.03)	.75	.71	1.74	(.35)
0	0	0	.46	-.18	.04	.02	.45	1.30	.40	.60	1.68

NOTE: Correlations are given in parenthesis.

strongly; the more compromising CSV criterion seems to represent better the real dimension-reducing power of a the correlated systems.

5.2 Hearing Loss

Let us apply the criteria reviewed in Section 3 and the new criteria proposed in Section 4 to the Hearing Loss Data presented in the Introduction. Remember that we have three alternative systems of components, shown in Table 2: the optimal principal components, the less optimal but better interpretable simple components, and the highly correlated and not very meaningful varimax components.

For this dataset, the varimax rotation evenly redistributes the total variance among components and reintroduces high correlations, as shown in Table 3. The varimax components are clearly unattractive in this example. Yet BLP, GV and RV criteria assign maximum optimality to those components (remember that this is always so, because these criteria are invariant under rotations). On the other hand, the proposed CSV and SCSV criteria assign values .79 and .61 to these components, which is far from optimal and a more realistic evaluation of the system's performance.

That the simple components are better at dimension reduction than the varimax components is evident from Table 3. The variances are closer to the principal component variances and the correlations are relatively small. BLP, GV and RV values are high for this system (.99, .97 and .99, respectively), but this is hardly surprising, since these criteria tend to err on the optimistic side, assigning high values to any reasonable system of components. What is more interesting, the more demanding CSV and SCSV criteria also assign high values to this system: .98 and .95, respectively. We conclude that the correlated system of simple components incurs only a 2% (respectively 5%) loss of dimension-reducing power compared to principal components. This indicates that simple components are a good alternative for this dataset.

5.3 Reflex Measurements

This is a real-data example taken from Jolliffe (200, p.58). The data consist of measurements of strength of reflexes at ten sites of the body, for 143 individuals. The variables come in five pairs, corresponding to right and left measurements on triceps, biceps, wrists, knees and ankles. The correlation matrix for these data is given in Table 1. Note that all of these correlations are positive and that "within-pair" correlations are particularly high (between .88 and .98). For this dataset, a natural and simple way to reduce dimension is to take the five averages between paired measurements. Hence we will take $q = 5$ as target dimension and will compare the performance of different systems of five components.

The first five principal components are given in Table 5. Overall, this system accounts for 97% of the variability of the original data. Within the system, the components account for 54, 21, 11, 9 and 5 percent of the variability, respectively. This is an example where principal components are easy to interpret: the first component is roughly an average of the ten variables, the second component is a contrast between leg- and arm-related measurements, and the other three are contrasts within these two groups.

Table 4: Correlation Matrix of Reflex Measurement Data.

Triceps		Biceps		Wrists		Knees		Ankles	
1	.98	.60	.71	.55	.55	.38	.25	.22	.20
	1	.62	.73	.57	.57	.40	.28	.21	.19
		1	.88	.61	.56	.48	.42	.19	.18
			1	.68	.68	.53	.47	.23	.21
				1	.97	.33	.27	.16	.13
					1	.33	.27	.19	.16
						1	.90	.40	.39
							1	.41	.40
								1	.94
									1

NOTE: First column under each heading corresponds to left side of the body.

The varimax components for this dataset are also shown in Table 5. These components are, roughly, the five averages of paired variables. Within the system, each component accounts for 20% of the variability, but far from being uncorrelated, the largest correlation between components is now .68. This is a clear example where the varimax rotation offers only disadvantages: the variances are evenly redistributed and correlations are reintroduced in the system. Yet criteria BLP, CP, GV and RV assign 100% optimality to this system. In contrast, the new criteria appropriately rate the varimax system as suboptimal: we have $\mathrm{CSV} = .75$ and $\mathrm{SCSV} = .58$.

There is another version of the varimax rotation that produces uncorrelated components (Jolliffe 1995), which is given in Table 6. This system offers little advantage over the classical orthogonal varimax system. It is true that the system is uncorrelated, but the standardization of the principal components performed before the rotation resulted in great loss of explanatory power. BLP and CP criteria assign 100% optimality to this system, but $\mathrm{GV} = .73$ and $\mathrm{RV} = .75$. The new criteria (which come down to the sum of variances for these uncorrelated

Table 5: Component Loadings for Reflex Measurements Data.

Variable	Principal components					Varimax components				
Triceps, L	.35	-.18	-.18	.49	.27	.71	.01	-.01	-.01	.01
Triceps, R	.36	-.19	-.15	.47	.27	.69	-.01	.01	.00	.02
Biceps, L	.36	-.13	.14	.04	-.71	-.08	.00	-.04	-.06	.81
Biceps, R	.39	-.14	.09	.05	-.41	.08	.00	.04	.06	.58
Wrists, L	.34	-.24	-.14	-.51	.16	-.02	-.01	-.01	.69	.04
Wrists, R	.34	-.22	-.17	-.52	.23	.01	.01	.00	.72	-.02
Knees, L	.30	.29	.50	.02	.24	.06	-.01	.69	.00	-.01
Knees, R	.27	.35	.54	-.07	.18	-.06	.01	.72	.00	.02
Ankles, L	.20	.53	-.41	-.03	-.07	.01	.70	.00	.01	-.01
Ankles, R	.19	.54	-.40	-.02	-.10	.00	.71	.00	.01	.01

Table 6: Component Loadings for Reflex Measurements Data (continued).

Variable	Uncorrelated varimax					Simple components				
Triceps, L	.66	-.03	-.02	-.11	-.15	.41	.00	.00	-.50	.29
Triceps, R	.63	-.05	-.00	-.10	-.14	.41	.00	.00	-.50	.29
Biceps, L	-.35	.02	-.26	-.31	.80	.41	.00	.00	.00	-.58
Biceps, R	-.14	-.01	-.13	-.15	.52	.41	.00	.00	.00	-.58
Wrists, L	-.12	-.02	-.02	.64	-.09	.41	.00	.00	.50	.29
Wrists, R	-.08	-.00	.00	.67	-.16	.41	.00	.00	.50	.29
Knees, L	-.05	-.17	.66	-.01	-.12	.00	.50	.50	.00	.00
Knees, R	.07	-.15	.67	-.00	-.08	.00	.50	.50	.00	.00
Ankles, L	.02	.68	-.11	.00	-.01	.00	.50	-.50	.00	.00
Ankles, R	.03	.69	-.12	-.02	.01	.00	.50	-.50	.00	.00

components) give a still lower optimality rate: $\mathrm{CSV} = \mathrm{SCSV} = .51$.

Simple components, more meaningful than either version of the varimax rotation, are obtained using the method of Gasser and Rousson (2003). The loading matrix is given in Table 6. Interpretation is even simpler than for principal components. The first component is the average of arm-related variables and the second component is the average of leg-related variables. The other three components are clear-cut contrasts. The correlation between the first two “block components” is .39; the other correlations are all smaller than .20 (which is low compared to the correlations of the varimax system). For this simple system we have, rounding up to the third decimal place, $\mathrm{BLP} = 1$, $\mathrm{GV} = .99$, $\mathrm{CP} = .998$, $\mathrm{RV} = 1$, $\mathrm{CSV} = .947$ and $\mathrm{SCSV} = .836$. For BLP, GV, CP and RV, this system is as good as the varimax. According to the new criteria, however, this system is clearly superior. Besides rating the whole system, it is possible to assess the individual contribution of the components. For for uncorrelated components CSV, the components account for 49, 25, 12, 9 and 6 percent of the total criterion, respectively; for SCSV, the corresponding figures are 44, 27, 12, 10, and 7 percent. The first component is by far the most relevant, explaining almost half of the dimension-reducing power of the system, while the second component explains one fourth; the other three components (the contrasts) do not add much. The new criteria, then, provide an adequate evaluation not only of the whole system but also of the individual components.

6 Conclusion

A number of alternatives to principal components have been proposed recently, that sacrifice some of the dimension-reducing power of the principal components in exchange for simplicity of the loadings and better interpretability. This calls for criteria that are able to evaluate the performance of correlated and/or non-orthogonal systems of components. We have shown in this paper that the existing criteria are not appropriate for this, because they do not handle correlations and lack of orthogonality in adequate ways. The examples in Section 5 reveal that these criteria often assign full or almost full optimality to systems that are too far from the principal components. In contrast, the new criteria proposed in Section 4 can discriminate well between "good" and "bad" suboptimal systems. Of these two criteria we tend to favor CSV, although both are consistent in their evaluations if the systems are not too far from optimal. For these reasons, we think that our proposals are a significant improvement over existing criteria.

Appendix

Maximization of (1): Let $\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Gamma}\mathbf{\Lambda}^{\frac{1}{2}}$ and $\mathbf{P} = \mathbf{\Sigma}^{\frac{1}{2}\top}\mathbf{A}(\mathbf{A}^\top\mathbf{\Sigma}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{\Sigma}^{\frac{1}{2}}$. We can rewrite (1) as $\mathrm{tr}(\mathbf{\Lambda}\mathbf{P})$. Since $\mathbf{P}$ is a projection matrix of rank q, $\mathbf{P} = \mathbf{B}\mathbf{B}^\top$ for some $\mathbf{B} \in \mathcal{O}(p,q)$ and then $\mathrm{tr}(\mathbf{\Lambda}\mathbf{P}) = \mathrm{tr}(\mathbf{B}\mathbf{\Lambda}\mathbf{B}^\top)$, which has a unique maximizer $\widehat{\mathbf{B}} = \mathbf{I}_q$, where $\mathbf{I}_q$ are the first q columns of the identity matrix; see Property A1 of Jolliffe (1986, p. 9) for a proof. Then $\widehat{\mathbf{P}} = \mathbf{I}_q\mathbf{I}_q^\top$, which occurs if and only if $\mathbf{\Gamma}_q\mathbf{\Gamma}_q^\top\widehat{\mathbf{A}} = \widehat{\mathbf{A}}$. This is satisfied only by matrices of the form $\widehat{\mathbf{A}} = \mathbf{\Gamma}_q\mathbf{C}$ with $\mathbf{C} \in \mathbb{R}^{q\times q}$ non-singular. ■

Maximization of (2): We will show that all the maximizers of $RV(\mathbf{A})$ under the restrictions $\|\mathbf{a}_k\| = 1$, $k = 1,\ldots,q$, and $\mathbf{A}$ of full rank, are of the form $\widehat{\mathbf{A}} = \mathbf{\Gamma}_q\mathbf{R}$ with $\mathbf{R} \in \mathcal{O}(q)$. The Lagrangian of this problem is $F(\mathbf{A},\mathbf{L}) = \mathrm{tr}(\mathbf{A}^\top\mathbf{\Sigma}^2\mathbf{A})/\{\mathrm{tr}(\mathbf{A}^\top\mathbf{\Sigma}\mathbf{A})^2\}^{\frac{1}{2}} + \mathrm{tr}\{\mathbf{L}(\mathbf{I}-\mathbf{A}^\top\mathbf{A})\}$, where $\mathbf{L}$ is diagonal with non-negative elements. The derivative with respect to $\mathbf{A}$, which we computed using differentials (Magnus and Neudecker 1999), is

$$\frac{\partial F}{\partial \mathbf{A}} = \frac{2\boldsymbol{\Sigma}^2\mathbf{A}}{\{\mathrm{tr}(\mathbf{A}^\top\boldsymbol{\Sigma}\mathbf{A})^2\}^{\frac{1}{2}}} - \frac{2\,\mathrm{tr}(\mathbf{A}^\top\boldsymbol{\Sigma}^2\mathbf{A})\boldsymbol{\Sigma}\mathbf{A}\mathbf{A}^\top\boldsymbol{\Sigma}\mathbf{A}}{\{\mathrm{tr}(\mathbf{A}^\top\boldsymbol{\Sigma}\mathbf{A})^2\}^{\frac{3}{2}}} - 2\mathbf{A}\mathbf{L}.$$

After setting this derivative to zero, premultiplying by $\mathbf{A}^\top$ and taking traces, we obtain that $\mathrm{tr}(\mathbf{A}^\top\mathbf{A}\mathbf{L}) = 0$; since the diagonal elements of $\mathbf{A}^\top\mathbf{A}$ equal 1 by assumption, and the elements of $\mathbf{L}$ are nonnegative, we deduce that $\mathbf{L} = 0$. Therefore, the maximizers of F satisfy the equation

$$\frac{\boldsymbol{\Sigma}\mathbf{A}}{\mathrm{tr}(\mathbf{A}^\top\boldsymbol{\Sigma}^2\mathbf{A})} = \frac{\mathbf{A}\mathbf{A}^\top\boldsymbol{\Sigma}\mathbf{A}}{\mathrm{tr}(\mathbf{A}^\top\boldsymbol{\Sigma}\mathbf{A})^2}.$$

Premultiplying by $\mathbf{A}^\top$ again and using that the diagonal of $\mathbf{A}^\top\mathbf{A}$ is 1, we actually deduce that $\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} = \mathbf{I}$ and $\mathrm{tr}(\widehat{\mathbf{A}}^\top\boldsymbol{\Sigma}^2\widehat{\mathbf{A}}) = \mathrm{tr}(\widehat{\mathbf{A}}^\top\boldsymbol{\Sigma}\widehat{\mathbf{A}})^2$. Then $F(\widehat{\mathbf{A}}, \widehat{\mathbf{L}}) = \{\mathrm{tr}(\widehat{\mathbf{A}}^\top\boldsymbol{\Sigma}^2\widehat{\mathbf{A}})\}^{\frac{1}{2}}$, which among $\mathcal{O}(p, q)$ matrices is uniquely maximized by $\widehat{\mathbf{A}}$ of the form given above. ■

Maximization of (3)*:* We will show that $\boldsymbol{\Gamma}_q$ is the unique maximizer of (3) under the constraints $\|\mathbf{a}_k\| = 1$, $k = 1, \ldots, q$. The Lagrangian of this optimization problem is

$$\begin{aligned} F(\mathbf{a}_1, \ldots, \mathbf{a}_q, l_1, \ldots, l_q) &= \sum_{k=1}^{q}(\mathbf{a}_k^\top\boldsymbol{\Sigma}\mathbf{a}_k - \mathbf{a}_k^\top\boldsymbol{\Sigma}\mathbf{A}_{(k-1)}\mathbf{B}_{(k-1)}\mathbf{A}_{(k-1)}^\top\boldsymbol{\Sigma}\mathbf{a}_k) \\ &\quad + \sum_{k=1}^{q} l_k(1 - \mathbf{a}_k^\top\mathbf{a}_k), \end{aligned}$$

where $\mathbf{B}_{(k)} = (\mathbf{A}_{(k)}^\top\boldsymbol{\Sigma}\mathbf{A}_{(k)})^{-1}$. The partial derivatives are

$$\begin{aligned} \frac{\partial F}{\partial \mathbf{a}_k} &= 2\boldsymbol{\Sigma}\mathbf{a}_k - 2l_k\mathbf{a}_k - 2\boldsymbol{\Sigma}\mathbf{A}_{(k-1)}\mathbf{B}_{(k-1)}\mathbf{A}_{(k-1)}^\top\boldsymbol{\Sigma}\mathbf{a}_k \\ &\quad +2\sum_{j=k+1}^{q}(\mathbf{e}_k^\top\mathbf{B}_{(j-1)}\mathbf{A}_{(j-1)}^\top\boldsymbol{\Sigma}\mathbf{a}_j)(\boldsymbol{\Sigma}\mathbf{A}_{(j-1)}\mathbf{B}_{(j-1)}\mathbf{A}_{(j-1)}^\top\boldsymbol{\Sigma}\mathbf{a}_j - \boldsymbol{\Sigma}\mathbf{a}_j) \end{aligned} \qquad (5)$$

for $k = 1, ..., q$, where $\mathbf{e}_k$ is the vector containing a one in the kth coordinate and zeros elsewhere. A necessary condition for $\mathbf{A}$ to maximize (3) is that all the partial derivatives (5) be equal to zero. We can solve for $\mathbf{A}$ in a sequential way. For $k = q$ (5) yields $2\mathbf{\Sigma a}_q - 2l_q\mathbf{a}_q - 2\mathbf{\Sigma A}_{(q-1)}\mathbf{B}_{(q-1)}\mathbf{A}_{(q-1)}^\top\mathbf{\Sigma a}_q = 0$, and premultiplying by $\mathbf{A}_{(q-1)}^\top$ we obtain that $l_q\mathbf{A}_{(q-1)}^\top\mathbf{a}_q = 0$. Hence $\mathbf{A}_{(q-1)}^\top\mathbf{a}_q = 0$, because $l_q \neq 0$. Thus, the last component should be orthonormal to the first $(q-1)$ components. Proceeding in this way for $k = (q-1), \ldots, 2$, we finally obtain that the maximizer $\widehat{\mathbf{A}}$ has to be orthonormal. It is easy to see that the components given by $\widehat{\mathbf{A}}$ have to be uncorrelated: if $\mathbf{A}^\top\mathbf{\Sigma A} \neq \mathbf{I}$ then $F(\mathbf{a}_1, \ldots, \mathbf{a}_q, l_1, \ldots, l_q) < \sum_{k=1}^q \mathbf{a}_k^\top\mathbf{\Sigma a}_k$, but $\sum_{k=1}^q \mathbf{a}_k^\top\mathbf{\Sigma a}_k \leq \sum_{k=1}^q \lambda_k$ for $\mathbf{A} \in \mathcal{O}(p,q)$ and $F(\gamma_1, \ldots, \gamma_q, l_1, \ldots, l_q) = \sum_{k=1}^q \lambda_k$, hence an orthogonal $\mathbf{A}$ such that $\mathbf{A}^\top\mathbf{\Sigma A} \neq \mathbf{I}$ is necessarily suboptimal. In other words, $\widehat{\mathbf{A}} \in \mathcal{O}(p,q)$ and $\widehat{\mathbf{A}}^\top\mathbf{\Sigma}\widehat{\mathbf{A}} = \mathbf{I}$, which is only satisfied by matrices consisting of q different eigenvectors of $\mathbf{\Sigma}$. Among these, the maximum is attained at $\mathbf{\Gamma}_q$. ■

References

[1] Cadima, J. and Jolliffe, I. T. (1995), "Loadings and Correlations in the Interpretation of Principal Components," *Journal of Applied Statistics*, 22, 203–214.

[2] Hausmann, R. (1982), "Constrained Multivariate Analysis," in *Optimisation in Statistics*, eds. S. H. Zanckis and J. S. Rustagi, Amsterdam: North-Holland, pp. 137–151.

[3] Hotelling, H. (1933), "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24, 417–441 and 498–520.

[4] Jackson, J. E. (1991), *A User's Guide to Principal Components* (1st ed.), New York: Wiley.

[5] Jolliffe, I. T. (2002), *Principal Component Analysis* (2nd ed.), New York: Springer.

[6] Jolliffe, I. T. (1995), "Rotation of Principal Components: Choice of Normalization Constraints," *Journal of Applied Statistics*, 22, 29–35.

[7] Jolliffe, I. T. and Uddin, M. (2000), "The Simplified Component Technique: An Alternative to Rotated Principal Components," *Journal of Computational and Graphical Statistics*, 9, 689–710.

[8] Kaiser, H. F. (1958), "The Varimax Criterion for Analytic Rotation in Factor Analysis," *Psychometrika*, 23, 187–200.

[9] Kiers, H. A. L. (1991), "Simple Structure in Component Analysis Techniques for Mixtures of Qualitative and Quantitative Variables," *Psychometrika*, 56, 197–212.

[10] Magnus, J. R. and Neudecker, H. (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics* (Revised ed.), New York: Wiley.

[11] McCabe, G. P. (1984), "Principal Variables," *Technometrics*, 26, 137–145.

[12] Neuhaus, J. and Wrigley, C. (1954), "The Quartimax Method: an Analytical Approach to Orthogonal Simple Structure," *British Journal of statistical Psychology*, 7, 81–91.

[13] Okamoto, M. (1969), "Optimality of Principal Components," in *Multivariate Analysis II*, ed. P. R. Krishnaiah, New York: Academic Press, pp. 673–685.

[14] Rao, C. R. (1964), "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhya*, Ser. A, 26, 329–358.

[15] Robert, P. and Escoufier, Y. (1976), "A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient," *Applied Statistics*, 25, 257–265.

[16] Rousson, V. and Gasser, Th. (2003), "Simple Component Analysis," unpublished.

[17] Vines, S. K. (2000), "Simple Principal Components," *Applied Statistics*, 49, 441–451.