TOPICS IN DESCRIPTIVE PRINCIPAL COMPONENT ANALYSIS

A THESIS SUBMITTED TO

THE UNIVERSITY OF KENT AT CANTERBURY

IN THE SUBJECT OF STATISTICS

FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY.

By

Jorge Filipe Campinos Landerset Cadima

September 1992

Abstract

Principal Component Analysis (PCA) is viewed as a *descriptive* multivariate method for a set of n observations on p variables.

Geometric considerations in the inner product spaces associated with such $n \times p$ data sets play a central role throughout this thesis and provide the motivation for the main results. Among these spaces are spaces of matrices, whose geometry is meaningfully discussed in terms of PCA's key concepts.

It is argued that the conventional interpretation of Principal Components, which is based on the magnitude of each variable's loading for that PC, can be misleading. Alternative approaches based on multiple regression are discussed.

The effects on PCA of linear transformations of the data are discussed in general terms, for non-singular and projective transformations. Specific applications are analyzed and a new solution to the problem of removing isometric size from morphometric data is suggested.

Indicators measuring the degree of similarity between the PCA of a data matrix and the PCA of some transformation of that matrix are provided, for various concepts of 'similarity'. Methods for joint multiple comparisons of several such transformations are also suggested, discussed and exemplified.

Finally, a truly scale-invariant alternative to PCA is suggested. At the core of Most Correlated Component Analysis (MCCA) lies a result by Hotelling in his 1933 pioneering paper on PCA. The new method and its performance relative to PCA are discussed in detail. This discussion provides new insights into the information provided by covariance and correlation matrices, as well as a new optimal criterion for PCA.

Preface

This thesis considers a number of different issues within the general framework of Principal Component Analysis (PCA). The common underlying theme which unifies these issues is the role played by geometric considerations in the various linear spaces which are associated with PCA. Such geometric considerations are best understood if we strip PCA of all probabilistic concepts and methods and focus on its linear algebraic foundations. For this reason, we have chosen to work with vectors of observations, as opposed to random variables, and to view PCA as a descriptive, rather than an inferential, technique.

The decision to work within a strictly descriptive context is not at odds with the literature on PCA. Many authors consider an $n \times p$ data matrix (where columns usually correspond to variables and rows to individuals or repeated observations on those variables) as the fundamental object of a Principal Component Analysis (see, for example, Volle (1981) [67], Lebart *et al.* (1982) [40] and (1984) [41], Cailliez and Pagés (1976) [7], among others). The difficulties in obtaining a comprehensive and tractable inferential theory are often cited as reasons for this approach. Chatfield and Collins (1980) [8, (pg.62)] write that:

> If we are prepared to assume that the observations are taken from a multivariate normal distribution, then some sampling theory is available (...). But this theory is of limited practical value, partly because many of the results are for the asymptotic case (as $n \to \infty$), and partly because the normality assumption is often questionable. In any case, the 'sample' may be observations from a complete population. Thus, the modern tendency is to view PCA as a mathematical technique with no underlying statistical model. The principal components obtained from the sample covariance matrix $\mathbf{S}$ are seen as *the* principal components and not as estimates of the corresponding quantities obtained from Σ. (...) Indeed, it is not even necessary to regard $\mathbf{X}$ [the original variables] and $\mathbf{Y}$ [their principal components] as random variables.

On a similar note, Krzanowski (1988) [38, (pg.258)] writes:

> We can thus see that inferential aspects of principal component analysis present a rather confusing picture. Small-sample theory is either very complicated or unavailable, and although large-sample theory can lead to relatively simple results it is not clear how large the sample must be

> before these results are valid. Also, there is ambiguity about the appropriate hypothesis to test. Consequently, it is this writer's view that the most satisfactory uses of principal component analysis are the descriptive ones.

The fact that most practical applications of PCA seem to be descriptive in nature must also be mentioned. Jolliffe (1986) [32, (pg.50)] writes that "although (...) there are many other ways of applying PCs, the original usage as a descriptive, dimension-reducing technique is probably still the most prevalent application".

But we would argue that a further reason can be invoked in making the case for a strictly linear algebraic approach to PCA: that only too often we do not see the geometry of the PCA forest, for the probabilistic trees.

The thesis has seven Chapters.

Chapter I introduces fundamental concepts, notation and terminology. PCA is briefly described, both as originally defined by Hotelling (1933) [27] and in more recent guises. The scatters of points defined in $\mathbb{R}^p$ and $\mathbb{R}^n$ by an $n \times p$ data matrix are discussed and the key concepts of PCA are viewed in that light. The main problems with the method are raised. Given the nature of this Chapter, there is little in its contents that is new, apart from a discussion of Kazmierczak's Logarithmic Analysis [34] in Subsection I.5.3.

Chapter II discusses the relationship between variables and PCs which underlies the 'interpretation' of PCs. The standard approach of associating a PC with a subset of variables on the basis of their loadings' magnitude (which is illustrated in Section II.1) is shown to be potentially misleading, in particular for Covariance Matrix PCs (Section II.2). Furthermore, it is shown that the standard way of approximating a PC by a linear combination of any k given variables (*i.e.*, taking the sum of the terms in the linear combination defining the PC which involve those variables) is sub-optimal and possibly mis-leading (Section II.4). Alternative ways of relating PCs to subsets of the variables and of approximating the PCs are discussed and illustrated with examples (Sections II.3 and II.5). Section II.6 sums up the discussion.

Chapter III introduces several inner product linear spaces of matrices and analyzes the strong relationship between the statistical concepts of PCA and geometric concepts in the matrix spaces associated with $n \times p$ data sets (Section III.1). Indices arising from such geometric considerations, but which also quantify the degree of similarity in the PCAs of two comparable data sets, are proposed in Section III.2. The $p \times p$ covariance (positive semi-definite) matrices, which form a cone in the space of all $p \times p$ matrices, are extensively studied, elaborating on an initial result by Tarazaga (1990) [65] (Section III.3).

Chapter IV discusses the effects on PCA of invertible linear transformations of a data set. After some introductory concepts in Sections IV.1 and IV.2, Section IV.3 goes on to give some general case results. Comparative studies of the PCAs of a data set and some linear transformation of it are carried out in Section IV.4, based on the similarity indices from Chapter III. In particular, the Covariance and Correlation Matrix PCAs of any given data set are compared. Methods for a multiple comparison of the PCAs of several different linear transformations of any given data set are also suggested in Section IV.5. These comparisons are illustrated with examples.

Chapter V discusses the effects on PCA of a particular type of singular linear transformations: projections. The special nature of such linear transformations provides pleasant results which are considered in detail in Section V.1. They are a blend of established results with others which, to the best of the author's knowledge, are new to the field. The results are applied to two different problems: the removal of isometric size from morphometric data – including a new solution to the problem (Section V.2) – and the removal of underlying models from 'sampled functions' or time-series data (Section V.3). Examples of both applications are given.

Chapter VI proposes an alternative Component Analysis: Most Correlated Component Analysis (MCCA). Its main advantage is to produce a set of scale-invariant components. The absence of scale invariance is arguably PCA's main drawback. The key to MCCA lies in a property of Correlation Matrix PCs noted by Hotelling (1933) [27] and in work by Meredith and Millsap (1986) [43]. Apart from giving form to these ideas, the novelty in this Chapter lies in a detailed discussion of the method, which is also compared with PCA (Section VI.2). The advantages and draw-backs of

MCCA are assessed. As a by-product of the discussion, a further optimal property for Principal Components is given (Theorem VI.1).

Finally, **Chapter VII** briefly discusses some ideas for future work.

The computational work for this thesis was carried out at the UKC Computing Centre, on a VAX/VMS 4000/300 and a Sun 4/470 (Unix). The figures and most calculations were produced by the GENSTAT 5 package and the LaTeX text processor was used to produce the text.

Funding for the research which led to this thesis was provided by the Gulbenkian Foundation, during the first ten months, and by the Commission of the European Communities (SCIENCE Programme), during the subsequent 24 months. Without their generous support this thesis would not have been possible.

A special word of thanks must go to Prof. Ian Jolliffe, who effectively supervised the work which led to this thesis, although no longer in an official capacity after taking up a Chair at the University of Aberdeen in April 1992. His fruitful combination of providing many tips and references without straitjacketing the research was very much appreciated, as was his regular and thorough attention to the research work. Also much appreciated were his many comments and suggestions which helped to improve the original draft of the thesis. A word of gratitude also to Prof. Byron Morgan, Mike Bremner, Barry Vowden and Tim Hopkins of the UKC Institute of Mathematics for their cooperation at some stage of this process.

Last but certainly not least, I would like to express my affectionate gratefulness to my wife, Manuela Pires, who has sacrificed so much during the past three years, to my daughter, Maria Cadima, who joined us half-way through this journey, and to my parents, who made all this possible.

Contents

Chapter I

Basic Concepts

We begin by introducing necessary background concepts and by briefly reviewing the nature and purpose of Principal Component Analysis (PCA).

I.1 Data matrices

The object of our attention will be an $n \times p$ **data matrix** $\mathbf{Y}$.

Typically, the p columns of $\mathbf{Y}$ correspond to **variables** and the n rows to **individuals**, that is, elements of a population or sample for which measurements of the p variables are obtained.

We shall denote the measurement on the i-th individual, for the j-th variable, by y_{ij} . The column vector with the n measurements on the j-th variable will be represented by $\mathbf{y}_j \in \mathbb{R}^n$. The p measurements for the i-th individual, that is, the i-th row of $\mathbf{Y}$, will be represented by the column vector $\mathbf{y}_i^{row} \in \mathbb{R}^p$. The generic element, column and row of subsequent matrices will also be represented with an analogous notation.

The issue of why and how the p variables and the n individuals were chosen will not be dealt with.

Given the decision to focus attention on PCA as a *descriptive* method of data analysis, the usually important distinction between a *population* and a *sample* of n individuals is no longer a crucial one. In both cases, the procedures used will be identical and the information they provide will, strictly speaking, relate only to the set of n individuals represented by the rows of $\mathbf{Y}$.

The only restriction that shall be imposed on the nature of the p variables is that they be quantitative variables. Thus, different variables may have different units of measurement. Sometimes, a single variable is observed at p points in time or space and the p columns of $\mathbf{Y}$ are then associated to each of those points, rather than to different variables.

In most applications, the number of variables (p) tends to be smaller, sometimes much smaller, than the number of individuals (n), although many applications in fields such as meteorology (Preisendorfer(1988) [50, (pg.64)]) and chemometrics (Brereton (1990) [6, (pg.13)]) are an exception to this rule . Henceforth **we shall assume that** $p < n$.

In general, no prior assumptions of any relationship between the p variables are made, although when certain relations are known to exist they can often be taken into account, as shall be seen in Chapter V. When there are relations of linear dependence between the p variables, this will affect the rank of the data matrix. It is often assumed that no such multicollinearities exist, that is, that $\text{rank}(\mathbf{Y}) = p$. This assumption does not restrict the number of cases under consideration since redundant variables can be dropped from the data matrix prior to any analysis.

We shall not usually work directly with the *raw data* matrix $\mathbf{Y}$. Instead, the original data will be processed in some way prior to a PCA. Three useful transformations of $\mathbf{Y}$ are:

i). **Column-centering**. Each column of $\mathbf{Y}$ is centered about its mean, i.e., the mean value of column j , $\bar{y}_{.j}$, is subtracted from all elements $\{y_{ij}\}_{i=1}^{n}$ in that column. Column-centering means that the original data matrix is replaced by matrix

$$\mathbf{X} = \mathbf{P}_{\mathbf{1}_n}\mathbf{Y} \tag{I.1.1}$$

where $\mathbf{P}_{\mathbf{1}_n} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ with $\mathbf{I}_n$ the $n \times n$ identity matrix, $\mathbf{1}_n$ the n x 1 matrix of ones and $\mathbf{1}_n'$ its transpose. $\mathbf{P}_{\mathbf{1}_n}$ is a matrix of orthogonal projections onto the orthogonal complement of the subspace spanned by $\mathbf{1}_n$ (see Proposition A.19). The generic element of $\mathbf{X}$ is thus:

$$x_{ij} = y_{ij} - \bar{y}_{.j}$$

This transformation of the data implies that any subsequent analysis will be *insensitive to (possibly different) additive changes of scale in any of the variables.*

ii). **Dividing by $\sqrt{n}$.** Usually accompanied by column-centering, all elements in the data matrix are multiplied by $\frac{1}{\sqrt{n}}$. The new data matrix can be written as:

$$\mathbf{W} = \tfrac{1}{\sqrt{n}}\mathbf{X} \tag{I.1.2}$$

In the next Section we shall see that a matrix in this form is particularly well suited for relating statistical and geometrical concepts. For the moment, we note the notational advantage of writing the **covariance matrix** which is determined by the p columns of $\mathbf{Y}$ as $\mathbf{S} = \mathbf{W}'\mathbf{W}$. It should be stressed that all variances and covariances are defined here with denominator n rather than n-1 as is common for the sample version of these moments. Given that we are not working in an inferential context, there is no need to ensure the unbiasedness of any estimators. Hence, variances are more naturally defined as the mean squared deviations from the mean.

iii). **Standardization** . This means that all entries in any column of $\mathbf{Y}$ are both centered about their means and divided by their standard deviation. In matrix notation, $\mathbf{Y}$ is replaced by:

$$\mathbf{Z} = \mathbf{P}_{\mathbf{1}_n}\mathbf{Y}\mathbf{D}_S^{-\frac{1}{2}} \tag{I.1.3}$$

where

$$\mathbf{D}_S = \mathbf{W}'\mathbf{W} \circ \mathbf{I}_p \tag{I.1.4}$$

is the **Hadamard** (entry-wise) **product** of the covariance matrix $\mathbf{S}$ and the $p \times p$ identity matrix. Thus, $\mathbf{D}_S$ is the diagonal matrix with the variances of all p variables and the power $-\frac{1}{2}$ in (I.1.3) means that the diagonal elements of $\mathbf{D}_S$ are replaced by the reciprocal of their square roots. All p variables in $\mathbf{Z}$ are therefore of unit variance. The **correlation matrix** determined by the original variables in $\mathbf{Y}$ is $\mathbf{R} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}$.

I.2 $\mathbb{R}^p$ and $\mathbb{R}^n$

I.2.1 The dual representation of a data matrix

The rows of any given $n \times p$ data matrix can be viewed as points in the p -dimensional Euclidean space $\mathbb{R}^p$, where each axis corresponds to a variable. Thus, an $n \times p$ data matrix defines a **scatter** or **configuration** of n points in $\mathbb{R}^p$. These n points (and the vectors which they define with the origin) represent the n individuals.

The effects on this scatter of the transformations of $\mathbf{Y}$ discussed in Section I.1 are:

i). For a column-centered data matrix, the center of gravity of the scatter will be the origin. Column-centering defines a translation of the system of axes in $\mathbb{R}^p$

so that the origin will coincide with the center of gravity of the scatter.

ii). Multiplying each element in the data matrix by $\frac{1}{\sqrt{n}}$ globally contracts the configuration in $\mathbb{R}^p$ by a factor of $\frac{1}{\sqrt{n}}$: each point will move along the vector it defines with the origin until its Euclidean distance from the origin is $\frac{1}{\sqrt{n}}$ times what it was before.

iii). Standardization stretches/contracts the configuration along each axis by a factor of $\frac{1}{s_j}$ where s_j is the j-th variable's standard deviation. The effects are not trivial and the general 'shape' of the scatter will change (unless all the s_j's are equal).

An alternative way of viewing the $n \times p$ data matrix is as a scatter or configuration of p points in $\mathbb{R}^n$, with each axis associated to an individual and each of the p points/vectors representing a variable (column of the data matrix).

The effect on this scatter of column-centering is slightly more complex than with the scatter in $\mathbb{R}^p$. On the other hand, standardization becomes easier to visualize:

i). It is not hard to see that a vector $\mathbf{x} \in \mathbb{R}^n$ is centered about its mean if and only if $\mathbf{1}_n{}'\mathbf{x} = 0$. Thus, column-centering constrains all p columns of the data matrix, and therefore any vector in the subspace of $\mathbb{R}^n$ which they span, to be orthogonal (with the usual inner product) to the vector $\mathbf{1}_n$ and any scalar multiple of it. Column-centering represents a **projection** of the p points in $\mathbb{R}^n$ onto the orthogonal complement of the line which bisects the first orthant of

$\mathbb{R}^n$, *i.e.*, which forms equal angles with the positive parts of all n axes (see also Appendix A,in particular Proposition A.19).

ii). As with the scatter in $\mathbb{R}^p$, multiplying the data by $\frac{1}{\sqrt{n}}$ globally contracts the scatter in $\mathbb{R}^n$.

iii). *Standardization* corresponds to multiplying the vector which the j-th point defines with the origin by the scalar $\frac{1}{s_j}$. Thus, we *change the magnitude but not the directions of the* p *vectors.*

This dual interpretation of an $n \times p$ data matrix will be of great importance in the discussion that follows. Its significance has been highlighted by many authors such as Kendall (1980) [35] (who speaks of the First and Second Kind of Spatial Representation), Lebart *et al* (1982) [40] and (1984) [41], Gower (1966) [19], Cailliez and Pagés (1976) [7], Escoufier (1987) [13], Volle (1981) [67], Krzanowski (1988) [38] (who speaks of the 'Object space' and the 'Variable space') and Ramsay (1982) [51], to name but a few.

We shall normally use the concise terminology "$\mathbb{R}^p$" and "$\mathbb{R}^n$", when referring to these two spaces. Occasionally, we shall also use the more frequent and suggestive terms "**variable space**" and "**individual space**" or "**column space**" and "**row space**" of the data matrix.

I.2.2 Statistical and geometrical concepts

A key factor for the importance of this dual representation is the *interplay of geometrical and statistical concepts in both spaces*, in particular with transformation (I.1.2).

In fact, the usual *inner product* between any two columns of matrix $\mathbf{W}$, $(<\mathbf{w}_i, \mathbf{w}_j>= \mathbf{w}_i'\mathbf{w}_j)$ is the *covariance* between the original variables from which they were derived (the variables $\mathbf{y}_i$ and $\mathbf{y}_j$) . The *norm* or length defined by that inner product (that is, the ℓ_2-norm, $\|\mathbf{w}_i\| = \sqrt{<\mathbf{w}_i, \mathbf{w}_j>}$) is the *standard deviation* of the corresponding variable $(\mathbf{y}_j)$. And the *cosine of the angle* between any two columns $(\cos(\mathbf{w}_i, \mathbf{w}_j) = \frac{<\mathbf{w}_i,\mathbf{w}_j>}{\|\mathbf{w}_i\|\cdot\|\mathbf{w}_j\|})$ is the *correlation* between the respective variables $(\mathbf{y}_i$ and $\mathbf{y}_j)$.

When similar geometric concepts are applied directly to the columns of a column-centered data matrix $\mathbf{X}$ (I.1.1), the corresponding statistical concepts will appear with a scaling factor involving n or $\sqrt{n}$ (except the correlation which is insensitive to scalings). When the geometric concepts are applied to the columns of the original data matrix $\mathbf{Y}$, the statistical moments must be replaced by their non-central counterparts and, again, scaling factors may be required.

The need for scaling factors in the statistical interpretation of inner products involving the columns of $\mathbf{Y}$ and $\mathbf{X}$ (and of $\mathbf{Z}$ in transformation (I.1.3)) is unfortunate. All the more so since the columns of the matrices $\mathbf{X}$ and $\mathbf{W}$ are not conceptually different from a linear-algebraic point of view, unlike what happens with the columns of $\mathbf{Y}$ (which are any vectors in $\mathbb{R}^n$) and the columns of a column-centered matrix $\mathbf{X}$ (which are orthogonal to the vector $\mathbf{1}_n$). In other words, whereas data vectors which are, or are not, centered about their means are easily identifiable as linear algebraic concepts, the statistically more convenient columns of $\mathbf{W}$ (I.1.2) have no similar linear algebraic property which sets them apart from the columns of $\mathbf{X}$. This problem can be overcome by working with an alternative inner product in $\mathbb{R}^n$, defined by:

$$< \mathbf{a}, \mathbf{b} >_{\frac{1}{n}\mathbf{I}_n} = \mathbf{a}' \left(\tfrac{1}{n}\mathbf{I}_n \right) \mathbf{b} = \tfrac{1}{n}\mathbf{a}'\mathbf{b} \qquad \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n \qquad \text{(I.2.1)}$$

Using this inner product, the norm of a vector in $\Re(\mathbf{1}_n)^{\perp}$ is merely the standard deviation of the variable it represents. The inner product of two vectors in the same subspace is the covariance of the variables which they represent. It is therefore a *natural inner product in* $\mathbb{R}^n$ *for statistical purposes.*

However, since the changes involved are not fundamental, we shall proceed for the time being with the more familiar Euclidean inner product, using the columns of matrix $\mathbf{W}$ instead of $\mathbf{X}$ whenever necessary.

Like any other $n \times p$ matrices, the data matrices described in Section I.1 and their transposes define linear mappings from $\mathbb{R}^p$ to $\mathbb{R}^n$ and vice-versa. Thus, for any vector $\mathbf{a} \in \mathbb{R}^p$, $\mathbf{Xa}$ is a vector of $\mathbb{R}^n$. And for any vector $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{X}'\mathbf{z}$ is a vector in $\mathbb{R}^p$. (For a fuller discussion see Ramsay (1982) [51]).

An intuitive grasp of certain aspects relating to a data set may be best obtained by one or the other of the scatters of points it defines in $\mathbb{R}^n$ and $\mathbb{R}^p$. Thus, adding or deleting individuals from a data set corresponds to adding or deleting points from the scatter in $\mathbb{R}^p$, whereas in $\mathbb{R}^n$ there is a more fundamental change in the dimensionality of the space itself. Conversely, adding or deleting a variable to the data set is best visualized as adding or deleting a point to the scatter in $\mathbb{R}^n$. We already saw that when different multiplicative changes of scale are introduced for different variables (which is the implication of standardization or, for example, when different kinds of measurements are converted from Standard Imperial units to the metric system) the effect is best visualized in terms of the configuration in $\mathbb{R}^n$.

In many multivariate statistics techniques, including PCA, *linear combinations of variables* play a key role. These new variables can be represented in $\mathbb{R}^n$ as the resultants of those linear combinations of the p vectors representing the original variables. We have seen how this will provide a notion of the new variable's standard deviation (from its length) and correlation (cosine of the angle formed) with any other variable. *Geometric intuition therefore becomes a valuable aid in understanding and working with such techniques.*

I.3 Matrix decompositions

Two general matrix theory results which will be extensively used below are the Spectral Decomposition of a (necessarily square) symmetric matrix and the Singular Value Decomposition of any matrix.

I.3.1 The Spectral Decomposition

It is well known (see, for example, Horn and Johnson (1985) [25, (Section 4.1)] for a more detailed discussion) that any (real) symmetric m x m matrix $\mathbf{M}$ (that is, such that $\mathbf{M} = \mathbf{M}'$) always has m *real* **eigenvalues** $\{\lambda_i\}_{i=1}^{m}$ and a complete set of m (real) orthonormal **eigenvectors** $\{\mathbf{a}_i\}_{i=1}^{m}$ and that, furthermore, $\mathbf{M}$ can be written as:

$$\mathbf{M} = \sum_{i=1}^{m} \lambda_i \mathbf{a}_i \mathbf{a}_i' \qquad (I.3.1)$$

This equation is known as the **Spectral Decomposition** of the symmetric matrix $\mathbf{M}$ and can be re-written in matrix notation as:

$$\mathbf{M} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}' \qquad (I.3.2)$$

where $\mathbf{\Lambda}$ is the diagonal matrix whose i-th diagonal element is λ_i and $\mathbf{A}$ is the matrix whose i-th column is $\mathbf{a}_i$. The orthonormality requirements on the eigenvectors imply that $\mathbf{A}$ is an **orthogonal matrix** (that is, $\mathbf{A}'\mathbf{A} = \mathbf{I}_m$).

Symmetric matrices for which all eigenvalues are non-negative are called **positive semi-definite matrices** (**positive definite** if all eigenvalues are strictly positive).

Positive semi-definite (**p.s.d.**) matrices are particularly relevant for our purposes since all covariance and correlation matrices, in fact, *all matrices of the form* $\mathbf{A}'\mathbf{A}$ *or* $\mathbf{A}\mathbf{A}'$, *for any* n×p *matrix* $\mathbf{A}$, *are positive semi-definite.* (see Horn and Johnson (1985) [25, (pg.407)]).

For future reference, we note some well established facts concerning the Spectral Decomposition and **p.s.d.** matrices:

i). If $\mathbf{a}_i$ is an eigenvector of $\mathbf{M}$, then any scalar multiple $\alpha\mathbf{a}_i$ is also an eigenvector of $\mathbf{M}$. The requirement that the eigenvectors have unit norm removes this source of indeterminacy, but only up to $\alpha = \pm 1$, that is, up to reflections of the vector about the origin. This remaining ambiguity, due to what we shall call the *sign-switching* of any vector $\mathbf{a}_i$ in the decomposition (I.3.1), implies that it is the absolute value and the pattern of signs in the coefficients of $\mathbf{a}_i$, rather than the signs themselves, which are determined. We will usually omit explicit references to this indeterminacy (in keeping with general practice).

ii). If all m eigenvalues of $\mathbf{M}$ are different, then the spectral decomposition (I.3.1) is unique, apart from sign-switching. In the case of equal eigenvalues, there is no longer uniqueness. If $\{\mathbf{a}_j\}_{j=1}^{k}$ are any k eigenvectors of $\mathbf{M}$ with a common eigenvalue, then any linear combination of the $\{\mathbf{a}_j\}_{j=1}^{k}$ will also be an eigenvector of $\mathbf{M}$ with the same eigenvalue. In that case, it is the subspace spanned by those eigenvectors, rather than the vectors themselves, which is uniquely determined.

Any orthonormal basis for that subspace can be used in (I.3.1).

iii). Even when all eigenvalues of $\mathbf{M}$ are different, the matrix form of the spectral decomposition (equation (I.3.2)) will only be unique (apart from sign-switching) if some convention is agreed to, regarding the order in which the eigenpairs $\{(\lambda_i, \mathbf{a}_i)\}_{i=1}^{m}$ are placed in the columns of matrices $\mathbf{A}$ and $\mathbf{\Lambda}$. The usual convention, which we shall follow, is to order the eigenpairs according to the size of the eigenvalues, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$.

iv). Positive semi-definite matrices are often defined by the alternative (equivalent) criterion that $\mathbf{M}$ is **p.s.d.** if and only if $\boldsymbol{\beta}'\mathbf{M}\boldsymbol{\beta} \geq 0$, $\forall \boldsymbol{\beta} \in \mathbb{R}^m$. If $\mathbf{M}$ is a covariance matrix, *i.e.*, if $\mathbf{M} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ for some column-centered data matrix $\mathbf{X}$, then the quadratic forms $\boldsymbol{\beta}'\mathbf{M}\boldsymbol{\beta} = \frac{1}{n}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \frac{1}{n}||\mathbf{X}\boldsymbol{\beta}||^2$ are merely the variances of $\mathbf{X}\boldsymbol{\beta}$, that is, of the linear combination of the columns of $\mathbf{X}$ defined by the coefficients of vector $\boldsymbol{\beta}$. These must clearly be non-negative.

We are unlikely to encounter a real-life data set whose (full rank) covariance matrix has *exactly* equal eigenvalues. However, covariance matrices with approximately equal eigenvalues are more frequent. In an inferential setting, *i.e.* , when $\mathbf{S}$ is a sample covariance (or correlation) matrix which estimates its population counterpart, this may result from a population covariance (or correlation) matrix which does have equal eigenvalues that are perturbed by sampling variability. In both an inferential or a descriptive setting, nearly equal eigenvalues of $\mathbf{S}$ may also mean that measurement errors in obtaining the data are perturbing the otherwise equal eigenvalues. These situations therefore warrant special attention in any analysis. However, we shall follow the widespread practice of *ignoring the case of equal eigenvalues* in most theoretical discussions.

If $\mathbf{M}$ is an *invertible* (full rank) symmetric matrix, with Spectral Decomposition (I.3.2), then its inverse can be written as:

$$\mathbf{M}^{-1} = \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}' \tag{I.3.3}$$

where $\boldsymbol{\Lambda}^{-1}$ is the inverse of $\boldsymbol{\Lambda}$, *i.e.*, the diagonal matrix whose diagonal elements are the reciprocals of those of $\boldsymbol{\Lambda}$. Furthermore, (I.3.3) is a spectral decomposition of $\mathbf{M}^{-1}$, although without the convention mentioned in point iii) above.

More generally, if $\mathbf{M}$ is a *positive definite* matrix, we can speak of the **k-th power of $\mathbf{M}$** for any real number k, defined as:

$$\mathbf{M}^{k} = \mathbf{A}\boldsymbol{\Lambda}^{k}\mathbf{A}' \tag{I.3.4}$$

If $k > 0$, then $\mathbf{M}$ can be taken to be a positive *semi*-definite matrix.

For integer values of k, this definition coincides with taking the repeated multiplications of $\mathbf{M}$ or $\mathbf{M}^{-1}$. For $k = \frac{1}{2}$ we have what is known as the **positive (semi)-definite square root of $\mathbf{M}$** (see Horn and Johnson (1985) [25, (pg.406)]).

The matrices $\mathbf{M}^{k}$ are always positive (semi)-definite.

I.3.2 The Singular Value Decomposition

The other major matrix decomposition we will encounter is the **Singular Value Decomposition (SVD)** of a generic $n \times p$ matrix $\mathbf{Y}$ (see Mardia, Kent and Bibby (1979) [42] , Jolliffe (1986) [32] , Rao (1973) [55] for a fuller discussion).

A generic $n \times p$ matrix $\mathbf{Y}$ of rank r can always be written as:

$$\mathbf{Y} = \sum_{i=1}^{r} \sigma_i \boldsymbol{\alpha}_i \mathbf{a}_i' \tag{I.3.5}$$

where $\{\sigma_i\}_{i=1}^{r}$ is a set of positive scalars whose squares are the (common) non-zero eigenvalues of $\mathbf{Y'Y}$ and $\mathbf{YY'}$; $\{\boldsymbol{\alpha}_i\}_{i=1}^{r}$ is a set of orthonormal vectors in $\mathbb{R}^n$ which are eigenvectors of $\mathbf{YY'}$ in the same order as the corresponding eigenvalues σ_i^2 ; $\{\mathbf{a}_i\}_{i=1}^{r}$ is a set of orthonormal vectors in $\mathbb{R}^p$ which are eigenvectors of $\mathbf{Y'Y}$, again ordered as their corresponding eigenvalues σ_i^2 . This is known as the **Singular Value Decomposition of matrix Y** . The vectors $\{\boldsymbol{\alpha}_i\}_{i=1}^{r}$ and $\{\mathbf{a}_i\}_{i=1}^{r}$ are called the **left** and **right** (respectively) **singular vectors of** $\mathbf{Y}$ and the scalars $\{\sigma_i\}_{i=1}^{r}$ are called the **singular values** of matrix $\mathbf{Y}$.

The SVD can also be expressed in matrix form as:

$$\mathbf{Y} = \boldsymbol{\alpha} \boldsymbol{\Sigma} \mathbf{A}' \tag{I.3.6}$$

where the r columns of $\boldsymbol{\alpha}$[1] and $\mathbf{A}$ are (respectively) the left and right singular vectors of $\mathbf{Y}$ and $\boldsymbol{\Sigma}$ is a diagonal matrix whose diagonal elements are the singular

[1] Capital α in the Greek alphabet is, of course, A. The use of a large α is preferred here for notational coherence. Similar use shall be made of other Greek letters whose upper cases coincide with those of Roman letters.

values. As with the Spectral Decomposition of a **p.s.d.** matrix, it is *assumed that the singular values of* $\mathbf{Y}$ (hence its singular vectors) *are ordered according to their size,* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$.

It should be noted that:

i). The characterization of the columns of $\boldsymbol{\alpha}$ and $\mathbf{A}$ as eigenvectors of $\mathbf{YY}'$ and $\mathbf{Y}'\mathbf{Y}$ implies comments similar to those for the spectral decomposition of a **p.s.d.** matrix concerning the uniqueness of the SVD. Thus, if $\mathbf{Y}$ has two or more equal singular values, then the corresponding singular vectors (right and left) can be replaced by any other orthonormal bases for the subspaces they span. The issue of sign-switching becomes slightly more complex, since it is still true that any $\mathbf{a}_i$ may be replaced by $-\mathbf{a}_i$, but only if the corresponding left singular vector $\boldsymbol{\alpha}_i$ is also replaced by $-\boldsymbol{\alpha}_i$. Thus, singular vectors *as a pair* are insensitive to sign-switching.

ii). *If* $\mathbf{Y}$ *is a* (symmetric) *positive definite matrix* (but not if $\mathbf{Y}$ is a negative definite or an indefinite symmetric matrix, *i.e.*, if the quadratic form $\boldsymbol{\beta}'\mathbf{Y}\boldsymbol{\beta}$ can take only negative or both positive and negative values for vectors $\boldsymbol{\beta} \in \mathbb{R}^m$) then $\mathbf{Y}$'s *Singular Value and Spectral Decomposition coincide*, since we can always take $\boldsymbol{\alpha} = \mathbf{A}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$. This implies that its singular and eigen values coincide (see Horn and Johnson (1985) [25, (pg.417)]). The same is essentially true for positive *semi*-definite matrices although in that case the above definitions will have to be made compatible either by defining matrix $\mathbf{A}$ in the spectral decomposition to be of type p x r and matrix $\mathbf{\Lambda}$ of type r x r or by using the definition of the SVD given in Horn and Johnson (1985) [25] where $\boldsymbol{\alpha}$ and $\mathbf{A}$ are both orthogonal matrices of types $n \times n$ and $p \times p$ and $\mathbf{\Sigma}$ is a rectangular matrix of type $n \times p$ with non-zero entries only for elements σ_{ii} $(i = 1, ..., r)$. Given that our concern for symmetric matrices centers essentially on **p.s.d.** matrices, we could actually omit any reference to the Spectral Decomposition and speak only of the Singular Value Decomposition.

iii). the *'best' rank* k *$(k < r)$ approximation to matrix* $\mathbf{Y}$, by which we mean a rank k matrix $\hat{\mathbf{Y}}_k$ which minimizes the **Frobenius norm** of the difference $\hat{\mathbf{Y}}_k - \mathbf{Y}$ (defined for any $n \times p$ matrix $\mathbf{A}$ as $||\mathbf{A}||_F = \sqrt{tr(\mathbf{A}'\mathbf{A})} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}^2}$ — see Horn and Johnson (1985) [25, (pg.291)]) is obtained by taking the first k terms in $\mathbf{Y}$'s SVD as given by (I.3.5) (see Jolliffe (1986) [32, (pg.38)]).This amounts to deleting the last $r - k$ columns of $\boldsymbol{\alpha}$ and $\mathbf{A}$ and the last $r - k$ rows and columns of $\mathbf{\Sigma}$ in equation (I.3.6).

As in the last Subsection, replacing $\mathbf{Y}$'s diagonal matrix of singular values with its inverse, in (I.3.6) will give rise to another matrix bearing an interesting relation to $\mathbf{Y}$. In fact, if we take $\mathbf{Z} = \boldsymbol{\alpha}\ \mathbf{\Sigma}^{-1}\mathbf{A}'$, then the transpose of matrix $\mathbf{Z}$ is commonly known as the **Moore-Penrose generalized inverse of Y** (see Basilevski (1983) [4, (pg.288)]) :

Definition I.1 *If* $\mathbf{Y}$ *is an* n×p *matrix with Singular Value Decomposition* $\mathbf{Y} = \boldsymbol{\alpha}\boldsymbol{\Sigma}\mathbf{A}'$, *then the* p×n *matrix* $\mathbf{Y}^-$ *given by:*

$$\mathbf{Y}^- = \mathbf{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}' \tag{I.3.7}$$

is known as the **Moore-Penrose generalized inverse** *of* $\mathbf{Y}$.

Since every matrix has an SVD, every matrix has a Moore-Penrose generalized inverse. It can also be shown (Horn and Johnson (1985) [25, (pg.421)]) that the generalized inverse is unique. Some properties of the Moore-Penrose generalized inverse (often used as its definition) are:

Proposition I.2 *Let* $\mathbf{Y}$ *be an* n×p *matrix and* $\mathbf{Y}^-$ *its Moore-Penrose generalized inverse. Then, the following are true:*

i). $\mathbf{Y}\mathbf{Y}^-$ *and* $\mathbf{Y}^-\mathbf{Y}$ *are symmetric.*

ii). $\mathbf{Y}\mathbf{Y}^-\mathbf{Y} = \mathbf{Y}$.

iii). $\mathbf{Y}^-\mathbf{Y}\mathbf{Y}^- = \mathbf{Y}^-$.

iv). If $\mathrm{n} = \mathrm{p}$ *and* $\mathbf{Y}$ *is invertible, then* $\mathbf{Y}^- = \mathbf{Y}^{-1}$.

We also have, from the definition:

$$\mathbf{Y}^-\mathbf{Y} = \mathbf{A}\mathbf{A}' = \mathbf{P}_{\Re(\mathbf{A})} \tag{I.3.8}$$

and

$$\mathbf{Y}\mathbf{Y}^{-} = \boldsymbol{\alpha}\boldsymbol{\alpha}' = \mathbf{P}_{\Re(\boldsymbol{\alpha})} \qquad \text{(I.3.9)}$$

The requirements of orthonormality on the columns of $\mathbf{A}$ and $\boldsymbol{\alpha}$ imply that $\mathbf{P}_{\Re(\mathbf{A})}$ (in equation (I.3.8)) and $\mathbf{P}_{\Re(\boldsymbol{\alpha})}$ (in equation (I.3.9)) are, respectively, *the orthogonal projectors on the subspaces spanned by the columns of* $\mathbf{A}$ *(in* $\mathbb{R}^p$ *) and of* $\boldsymbol{\alpha}$ *(in* $\mathbb{R}^n$ *)* (see Proposition A.19). If $\mathbf{Y}$ is of rank p , the matrix (I.3.8) will be the identity matrix and the columns of $\boldsymbol{\alpha}$, $\mathbf{Y}$ and $\mathbf{Y}^{-'}$ all span the same subspace of $\mathbb{R}^n$, since the columns of $\mathbf{Y}$ and $\mathbf{Y}^{-'}$ are merely (invertible) linear combinations of the columns of $\boldsymbol{\alpha}$. Therefore, the orthogonal projector (I.3.9) can be written as (see also Appendix A):

$$\boldsymbol{\alpha}\boldsymbol{\alpha}' = \mathbf{Y}\mathbf{Y}^{-} = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'$$

The relation which the above equation suggests can be directly confirmed to be true (for $\mathbf{Y}$ of rank p):

$$\mathbf{Y}^{-} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}' \qquad \text{(I.3.10)}$$

In this case, the generalized inverse of $\mathbf{Y}$, when pre-multiplying any vector $\mathbf{z} \in \mathbb{R}^n$, will give the p-dimensional vector of regression coefficients for the multiple linear regression of $\mathbf{z}$ on the columns of $\mathbf{Y}$ (see, for example, Searle (1982) [59, (pg.366)]).

I.4 Principal Component Analysis

The term **Principal Component Analysis** was coined by Hotelling in his pioneering paper of 1933 [27]. Hotelling's derivation of Principal Components is no longer the standard way of introducing the subject, but it is useful for a fuller understanding of the essence of the method.

We shall follow through Hotelling's reasoning, re-phrasing the inferential statistics and probabilistic terminology and concepts with their descriptive statistics and linear algebraic counterparts. This will enable us to introduce the key concepts with which we shall work from now on.

The connections with other approaches to PCA will then be briefly discussed.

I.4.1 Hotelling's approach

Hotelling's approach is strongly influenced by **Factor Analysis** (for a discussion of Factor Analysis and its relation to PCA, see Chapter 7 of Jolliffe (1986) [32]) and linear regression. Given p variables $\{\mathbf{x}_j\}_{j=1}^{p}$, he asks the standard Factor Analysis question ([27, (pg. 417)]):

> whether some more fundamental set of independent variables exists, perhaps fewer in number than the $\mathbf{x}$'s, which determine the values that the $\mathbf{x}$'s will take.

These new "fundamental" variables are called **components** by Hotelling, who denotes them as $\{\gamma_i\}_{i=1}^{p}$.

Given the ultimate goal of determining the values of the variables from those of the components, and with a linear regression in mind, Hotelling assumes that *the variables are linear combinations of the components.* Thus, he considers :

$$\mathbf{x}_j = \sum_{i=1}^{p} c_{ij}\gamma_i \ , \qquad \forall j = 1, ..., p \tag{I.4.1}$$

where the $\{c_{ij}\}_{i,j=1}^{p}$ are the (unknown) coefficients in the linear combinations.

Replacing Hotelling's random variables with n-dimensional vectors, the p equations in (I.4.1) become:

$$\mathbf{x}_j = \mathbf{\Gamma}\mathbf{c}_j \ , \qquad \forall j = 1, ..., p \tag{I.4.2}$$

where $\mathbf{\Gamma}$ is the $n{\times}p$ matrix whose i-th column is the i-th component γ_i and $\mathbf{c}_j$ is the p-dimensional column vector $\mathbf{c}_j = (c_{1j}, c_{2j}, ..., c_{pj})$.

The p equations (I.4.2) can be summed up in the matrix equation:

$$\mathbf{X} = \mathbf{\Gamma}\mathbf{C} \tag{I.4.3}$$

where $\mathbf{x}_j$ and $\mathbf{c}_j$ are the j-th columns of the $n{\times}p$ matrix $\mathbf{X}$ and the $p{\times}p$ matrix $\mathbf{C}$, respectively.

Hotelling assumes that the p variables are centered about their means, *i.e.*, that $\mathbf{1}_n{}'\mathbf{X} = \mathbf{0}$. From (I.4.3) we then have $\mathbf{1}_n{}'\mathbf{\Gamma}\mathbf{C} = \mathbf{0}$. Even though nothing has yet been said about the invertibility of $\mathbf{C}$, this equation suggests that we assume that the p components are also centered about their means.

Hotelling then strengthens his original requirement for a "fundamental set of independent variables" by saying [27, (pg.417)] :

> we shall consider only normally distributed systems of components having zero correlations and unit variances.

The multinormality assumption (which equates independence with uncorrelatedness and implies that the variables will also have a multinormal distribution - see Anderson (1958) [1, (Section 2.4)]) is not meaningful in our setting. The uncorrelatedness and unit variance requirements mean that we impose the additional constraint:

$$\tfrac{1}{n}\mathbf{\Gamma}'\mathbf{\Gamma} = \mathbf{I}_p \tag{I.4.4}$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix.

It also implies that the p components $\{\gamma_i\}_{i=1}^{p}$ form an orthogonal basis of the p-dimensional subspace of $\mathbb{R}^n$ which they span. Unfortunately, we cannot speak of an ortho*normal* basis. With the standard inner product in $\mathbb{R}^n$, unit variance corresponds to a norm of $\sqrt{n}$, not to a unit norm. This is part of the price we must pay for working with the usual, rather than the statistically natural, inner product. But it is not a major problem and when necessary, we shall re-scale the components to have unit (conventional) norm.

From equations (I.4.3) and (I.4.4), we have:

$$\mathbf{C} = \tfrac{1}{n}\mathbf{\Gamma}'\mathbf{X} \tag{I.4.5}$$

The generic element c_{ij} of matrix $\mathbf{C}$ can now be seen to be the *covariance of the (as of yet undetermined) i-th component with the j-th variable* :

$$c_{ij} = \tfrac{1}{n} < \boldsymbol{\gamma}_i, \mathbf{x}_j > \tag{I.4.6}$$

But Hotelling notes [27, (pg.420)] that this still leaves an "infinity of possible modes of resolution of our variables into components", that is, an infinity of possible criteria to determine the "fundamental" variables satisfying equations (I.4.3) and (I.4.4).

Reasoning that if we are to, again, regress variables on components, then we should ([27, (pg.421)]):

> begin with a component $\boldsymbol{\gamma}_1$ whose contributions to the variances of the $\mathbf{x}$'s have as great a total as possible; that we next take a component $\boldsymbol{\gamma}_2$, independent of $\boldsymbol{\gamma}_1$, whose contribution to the residual variance is as great as possible; and that we proceed in this way to determine the components, not exceeding [p] in number and perhaps neglecting those whose contributions to the total variance are small. This we shall call the *method of principal components.*

The variance of the j-th variable is obtained from the norm squared of the j-th column of matrix $\mathbf{C}$:

$$\tfrac{1}{n}\|\mathbf{x}_j\|^2 = \tfrac{1}{n}\mathbf{x}_j'\mathbf{x}_j = \tfrac{1}{n}\mathbf{c}_j'\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\mathbf{c}_j = \mathbf{c}_j'\mathbf{c}_j = \|\mathbf{c}_j\|^2 = \sum_{i=1}^{p} c_{ij}^2 \tag{I.4.7}$$

Hotelling calls each c_{ij}^2 the *contribution of* $\boldsymbol{\gamma}_i$ *to the variance of* $\mathbf{x}_j$. Thus, the sum of $\boldsymbol{\gamma}_i$'s contributions to the variances of all $\mathbf{x}_j$'s , as required by the above criterion, is given by the norm squared of the i-th row of matrix $\mathbf{C}$:

$$||\mathbf{c}_i^{row}||^2 = \sum_{j=1}^{p} c_{ij}^2 \tag{I.4.8}$$

If we now depart from Hotelling's reasoning, the problem becomes easy to solve in our linear algebraic context.

From equation (I.4.5), each row of matrix $\mathbf{A}$ can be written as:

$$\mathbf{c}_i^{row} = \frac{1}{n}\boldsymbol{\gamma}_i'\mathbf{X} \ , \quad (i = 1, ..., p) \tag{I.4.9}$$

Hence, the first step in the method of Principal Components (maximizing $||\mathbf{c}_1^{row}||^2$) becomes:

$$\max_{\mathbf{c}\in \mathbb{R}^p} ||\mathbf{c}||^2 = \max_{\substack{\boldsymbol{\gamma}\in\Re(\mathbf{1}_n)^{\perp} \\ \boldsymbol{\gamma}'\boldsymbol{\gamma}=n}} \frac{1}{n^2}\boldsymbol{\gamma}'\mathbf{X}\mathbf{X}'\boldsymbol{\gamma} = \max_{\boldsymbol{\gamma}\in \mathbb{R}^n} \frac{\boldsymbol{\gamma}'[\frac{1}{n}\mathbf{X}\mathbf{X}']\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{\gamma}} \tag{I.4.10}$$

The last equation follows since if $\boldsymbol{\gamma}$ has any component in $\Re(\mathbf{1}_n)$ it will contribute nothing to $\boldsymbol{\gamma}'\mathbf{X}\mathbf{X}'\boldsymbol{\gamma}$, as that component vanishes with the product $\boldsymbol{\gamma}'\mathbf{X}$.

The last expression in (I.4.10) is the well-known *Rayleigh-Ritz ratio variational characterization of the first eigenpair of* $\frac{1}{n}\mathbf{X}\mathbf{X}'$ (see Horn and Johnson (1985) [25, (pg.176)], for example).

The remaining Principal Components follow in a similar fashion. The residual variance of the j-th variable, after removing $\boldsymbol{\gamma}_1$'s contribution, is given by $\frac{1}{n}\mathbf{x}_j'\mathbf{x}_j - c_{1j}^2$. From equation (I.4.7) we can still consider c_{ij}^2, $(i > 1)$ as the contribution of $\boldsymbol{\gamma}_i$

to this residual variance of $\mathbf{x}_j$. Thus, the total contribution of γ_2 to the residual variances of all $\mathbf{x}$'s is still given by $||\mathbf{c}_2^{row}||^2$ in equation (I.4.8). We are therefore back to the maximization problem (I.4.10) with the added constraint that the solution $\gamma \in \mathbb{R}^n$ must be orthogonal (uncorrelated) to the previously determined component γ_1. As is known (see, for example, Jolliffe (1986) [32] or Anderson (1958) [1] with the appropriate changes resulting from our descriptive context) this new problem is solved by the second eigenpair of $\frac{1}{n}\mathbf{XX}'$.

As was noted previously , Hotelling's PCs must , strictly speaking, be the $\sqrt{n}$-norm eigenvectors of $\frac{1}{n}\mathbf{XX}'$. This resulted from a subjective, but convenient, assumption of scale for the components (unit-variance). In our context, an alternative assumption is convenient: to work with unit-*norm* vectors. We therefore proceed to re-scale the PCs accordingly:

$$\mathbf{\Psi} = \frac{1}{\sqrt{n}}\mathbf{\Gamma} \tag{I.4.11}$$

Using the notation of Section I.1, equations (I.4.3) up to (I.4.5) can now be re-written as:

$$\mathbf{W} = \mathbf{\Psi C} \tag{I.4.3$'$}$$

$$\mathbf{\Psi}'\mathbf{\Psi} = \mathbf{I}_p \tag{I.4.4$'$}$$

$$\mathbf{C} = \mathbf{\Psi}'\mathbf{W} \tag{I.4.5$'$}$$

The new unit-norm PCs $\{\psi_j\}_{j=1}^{p}$ now form an orthonormal set (with the usual inner product).

Thus, the p unit-norm *Principal Components* for the variables $\{\mathbf{x}_j\}_{j=1}^p$ (and for the original uncentered variables $\{\mathbf{y}_j\}_{j=1}^p$) are the *first* p *unit-norm eigenvectors of the matrix* $\mathbf{WW}'$. The part of the sum of variances of the variables which is accounted for by each PC is the corresponding eigenvalue of $\mathbf{WW}'$.

From the previous discussion and from standard linear algebra or matrix results it is possible to state several properties and characteristics of PCs:

i). The non-uniqueness of unit-norm eigenvectors that results from *sign-switching* implies that each PC is actually a pair of vectors going off from the origin into opposite directions in $\mathbb{R}^n$. Sometimes it is convenient to think of a PC as the *line* (subspace) which these vectors span.

ii). It is customary to express the part of the total variance accounted for by any PC as a percentage or, alternatively, as a *proportion*:

$$\pi_i = \frac{\lambda_i}{\sum_{i=1}^p \lambda_i} \tag{I.4.12}$$

These quantities will be referred to as the **relative eigenvalues** of the matrix $\mathbf{WW}'$. The use of the relative eigenvalues, which take values between zero and one, not only makes for easier interpretations but also ensures that the results of a PCA are *insensitive to a global multiplicative change of scales in the variables.* In fact, it is easily confirmed that if $(\lambda, \mathbf{a})$ is an eigenpair of a matrix $\mathbf{M}$, then $(\alpha\lambda, \mathbf{a})$ is the corresponding eigenpair of matrix $\alpha\mathbf{M}$, for any scalar α. But the relative eigenvalues of $\mathbf{M}$ and $\alpha\mathbf{M}$ are identical.

iii). The $n \times p$ matrix of components $\mathbf{\Gamma}$ is, due to the orthogonality requirement (I.4.4), always of rank p. A standard matrix result (Horn and Johnson (1985), [25, (pg.13)]) together with equation (I.4.3), guarantees that if $\mathbf{X}$ is also of rank p, then $\mathbf{C}$ is non-singular. In that case, we have:

$$\mathbf{\Gamma} = \mathbf{X}\mathbf{C}^{-1} \tag{I.4.13}$$

which enables us to write the components as linear combinations of the variables.

iv). For any $n \times p$ matrix $\mathbf{Y}$, the rank of $\mathbf{Y}\mathbf{Y}'$ equals the rank of $\mathbf{Y}$ (see, for example, [25, (pg.13)]). Since the rank of a symmetric matrix (as is $\mathbf{W}\mathbf{W}'$) equals the number of its non-zero eigenvalues ([25, (pg.175)]), the number of PCs which actually contribute something to the total variance (*i.e.*, corresponding to non-zero eigenvalues) equals the rank of $\mathbf{X}$. Hence, the subspace spanned by $\{\mathbf{x}_j\}_{j=1}^{p}$ is of dimensionality $k = \text{rank}(\mathbf{X})$ and the first k PCs will form an orthogonal basis for that subspace.

v). Hotelling actually worked with the *additional assumption of unit-variance variables*. This would correspond to using standardized data and to determining the first p eigenpairs of $\frac{1}{n}\mathbf{Z}\mathbf{Z}'$ rather than $\mathbf{W}\mathbf{W}'$ (where $\mathbf{Z}$ is given by (I.1.3)). Together with the unit-variance components assumption, it implies that the generic element of matrix $\mathbf{C}$ (given by equation (I.4.6)) is the *correlation* between the i-th PC of the standardized data and the j-th variable. We have not introduced the unit-variance requirement so far because it is not crucial for the discussion, but it will be further discussed in the final Section of this Chapter.

vi). Equations (I.4.1) and (I.4.6) combine to give:

$$\mathbf{x}_j = \sum_{i=1}^{p} \tfrac{1}{n} < \boldsymbol{\gamma}_i, \mathbf{x}_j > \boldsymbol{\gamma}_i \tag{I.4.14}$$

This can be re-written in terms of unit-norm PCs as:

$$\mathbf{x}_j = \sum_{i=1}^{p} < \boldsymbol{\psi}_i, \mathbf{x}_j > \boldsymbol{\psi}_i \tag{I.4.15}$$

This second equation is a general expression for any vector $\mathbf{x}_j$ in a subspace of any finite-dimensional inner product linear space – and, indeed, in any Hilbert space (see Goffman and Pedrick (1965) [18, (Section 4.4)]) – when the $\{\boldsymbol{\psi}_i\}_{i=1}^{p}$ form an orthonormal basis for the subspace containing $\mathbf{x}_j$. The coefficients $< \boldsymbol{\psi}_i, \mathbf{x}_j >$ are then known as the **Fourier coefficients** of the vector $\mathbf{x}_j$ with respect to the orthonormal basis $\{\boldsymbol{\psi}_i\}_{i=1}^{p}$ for the given inner product (see, for example, Shilov (1961) [61, (pg.143)] or Goffman and Pedrick (1965) [18, (pg.175)]).

Equation (I.4.14) can also be interpreted as the Fourier expansion of $\mathbf{x}_j$ with respect to the basis $\{\boldsymbol{\gamma}_i\}_{i=1}^{p}$ and the inner product (I.2.1).

The geometric interpretation of the Fourier coefficients

$$< \boldsymbol{\psi}_i, \mathbf{x}_j > = \|\mathbf{x}_j\| \cos(\boldsymbol{\psi}_i, \mathbf{x}_j) = \|\mathbf{x}_j\| \cos(\boldsymbol{\gamma}_i, \mathbf{x}_j)$$

(where $\cos(\mathbf{a}, \mathbf{b})$ denotes the cosine of the angle in $\mathbb{R}^n$ between the vectors $\mathbf{a}$ and $\mathbf{b}$) is that they are the *coefficients of the orthogonal projection of vector* $\mathbf{x}_j$ *onto the direction of the vector* ψ_i *(or* γ_i*)* - see also Shilov (1961) [61, (pg.263)]. They are given by the generic elements of matrix $\sqrt{n}\mathbf{C}$ (see (I.4.6)).

vii). The previous remark and the properties of orthogonal bases imply that if we take any subset $\mathcal{G}$ of k (with $k < p$) unit-norm PCs and orthogonally project the p variables in $\mathbb{R}^n$ onto the subspace spanned by those components, the p new projected variables will be given (using also (I.4.5')) by:

$$\hat{\mathbf{X}} = \mathbf{\Psi}_{\mathcal{G}}(\sqrt{n}\mathbf{C}_{\mathcal{G}}) = \mathbf{\Psi}_{\mathcal{G}}\mathbf{\Psi}_{\mathcal{G}}'\mathbf{X} \qquad \text{(I.4.16)}$$

where $\mathbf{\Psi}_{\mathcal{G}}$ is the n x k submatrix of $\mathbf{\Psi}$ whose k columns are the unit-norm PCs in the subset $\mathcal{G}$ and $\mathbf{C}_{\mathcal{G}}$ is the k x p submatrix of $\mathbf{C}$ given by those rows of $\mathbf{C}$ corresponding to PCs in $\mathcal{G}$ (see also Proposition A.19). In PCA it is common practice to consider subsets $\mathcal{G}$ given by the *first* k PCs of $\mathbf{X}$, which account for most of the total variance. Equation (I.4.16) then gives the orthogonal projection of the original variables onto this k-dimensional Principal Subspace.

The reasoning behind equation (I.4.16) is valid for any set of orthonormal vectors $\{\psi_i\}_{i=1}^{k}$ in $\mathbb{R}^n$, even if they are not *Principal* components. This will be considered again in Chapter VI .

The above solution to Hotelling's formulation of the method of Principal Components is a short-cut to his own solution which became possible due to the decision to work in the context of descriptive statistics, replacing random variables with n-dimensional vectors.

The essence of Hotelling's solution lay in the fact that the PCs can also be determined from the eigenpairs of the $p\times p$ covariance matrix of the data $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$.

The tie-up with our previous discussion is well-known. If $\mathbf{Z}$ is an m x q matrix; if $\mathbf{V} = \mathbf{Z}'\mathbf{Z}$; if $\{\lambda_i\}_{i=1}^{q}$ are the eigenvalues of $\mathbf{V}$ and if $\{\mathbf{a}_i\}_{i=1}^{q}$ are the set of q associated orthonormal eigenvectors of $\mathbf{V}$, then $\{\mathbf{Za}_i\}_{i=1}^{q}$ is a set of orthogonal eigenvectors of $\mathbf{W} = \mathbf{ZZ}'$, whose corresponding eigenvalues are also $\{\lambda_i\}_{i=1}^{q}$. In fact, $\mathbf{Va}_i = \lambda_i\mathbf{a}_i$ is equivalent to:

$$\mathbf{Z}'\mathbf{Za}_i = \lambda_i\mathbf{a}_i$$

$$\Rightarrow \mathbf{ZZ}'(\mathbf{Za}_i) = \lambda_i(\mathbf{Za}_i) \qquad (\text{I}.4.17)$$

Taking $\mathbf{Z} = \frac{1}{\sqrt{n}}\mathbf{X}$ we obtain the link used by Hotelling. If $\{\mathbf{a}_j\}_{j=1}^{p}$ is a set of p orthonormal eigenvectors of the covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$, with eigenvalues $\{\lambda_j\}_{j=1}^{p}$, then $\{\mathbf{Xa}_j\}_{j=1}^{p}$ is a set of eigenvectors of matrix $\frac{1}{n}\mathbf{XX}'$ with the same eigenvalues. The p vectors $\{\mathbf{Xa}_j\}_{j=1}^{p}$ form an orthogonal set, but no longer with unit norms, since:

$$< \mathbf{Xa}_i, \mathbf{Xa}_j >= \mathbf{a}_i'\mathbf{X}'\mathbf{Xa}_j = n\lambda_j\mathbf{a}_i'\mathbf{a}_j = n\lambda_j\delta_{ij}$$

Thus, the j-th vector $\mathbf{Xa}_j$ has norm $\sqrt{n \cdot \lambda_j}$ and is merely a rescaling of Hotelling's PCs which (assuming all λs are non-zero) is obtained as:

$$\gamma_j = \frac{1}{\sqrt{\lambda_j}}\mathbf{Xa}_j = \frac{1}{\sqrt{\lambda_j}}\sum_{i=1}^{p} a_{ij}\mathbf{x}_i \quad (j = 1, ..., p) \tag{I.4.18}$$

These p equations can be written in matrix form as:

$$\boldsymbol{\Gamma} = \mathbf{XA}\boldsymbol{\Lambda}^{-\frac{1}{2}} \tag{I.4.19}$$

where $\boldsymbol{\Gamma}$ and $\mathbf{X}$ are, as before, the matrices of $\sqrt{n}$-norm components and of the data, $\mathbf{A}$ is the $p \times p$ matrix whose p columns are the eigenvectors $\{\mathbf{a}_i\}_{i=1}^{p}$ of $\mathbf{S}$ and $\boldsymbol{\Lambda}^{-\frac{1}{2}}$ is the diagonal matrix whose i-th diagonal element is $\frac{1}{\sqrt{\lambda_i}}$.

The orthonormality constraint on the eigenvectors of $\mathbf{S}$ implies that $\mathbf{A}$ is an orthogonal matrix (*i.e.*, $\mathbf{A}'\mathbf{A} = \mathbf{I}_p$), hence invertible. The last equation then implies (again if $\mathbf{X}$ is of rank p) that:

$$\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{A}'$$

and from (I.4.5) the matrix of coefficients can be written as:

$$\mathbf{C} = \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{A}' \tag{I.4.20}$$

The elements of the eigenvectors $\{\mathbf{a}_j\}_{j=1}^{p}$ — which are the coefficients of the linear combination of the columns of $\mathbf{X}$ producing the PCs of norm $\sqrt{n \cdot \lambda_j}$ — are known as the **loadings** of those PCs.

Since the key quantities to be determined in this approach are the eigenpairs $\{(\lambda_j, \mathbf{a}_j)\}_{j=1}^{p}$ of the covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$, PCA as defined above is also referred to as **Covariance Matrix PCA**. With Hotelling's original assumption of unit variance variables, the eigenpairs of the *correlation* matrix determined by $\mathbf{X}$ will be the key quantities. We then speak of a **Correlation Matrix PCA**.

This approach has the advantage of working with the eigenvectors of a $p \times p$ matrix, rather than with those of the normally larger $n \times n$ matrices. Quite often the value of n is very large and computing the eigenpairs of the $n \times n$ matrix can become a heavy computational burden. It is fortunate that in many other circumstances – as will be seen throughout this thesis – *many concepts of relevance for PCA in the space* $\mathbb{R}^n$ *can be studied by looking at associated concepts in the usually lower-dimensional space* $\mathbb{R}^p$.

I.4.2 Alternative approaches to PCA

More recent presentations of PCA (see, for example, Jolliffe (1986) [32, (Section I.2)] and Preisendorfer (1988) [50, (Section I.C)] for more detailed historical reviews of PCA) have almost invariably inverted Hotelling's original formulation of the variables as linear combinations of the components. In choosing to begin by defining the components as linear combinations of the variables, they have shed the legacy of Factor Analysis.

This has also entailed a shift towards replacing Hotelling's original criterion for determining PCs with alternative criteria, which will often produce a re-scaled version of Hotelling's PCs.

The most popular among these alternative criteria seem to be:

i). Looking for the *uncorrelated linear combinations of the* p *original variables which successively have greatest variance*, subject to the constraint that the (p-dimensional) vectors of coefficients be of unit-norm. This is the approach with which PCA is introduced by authors such as Anderson (1958) [1], Jolliffe (1986) [32], Mardia, Kent and Bibby (1979) [42], Chatfield and Collins [8], among others.

This approach, which is a favourite when working with random variables, in our context implies that we seek the set of successively *longest* possible orthogonal vectors in the subspace spanned by $\{\mathbf{x}_j\}_{j=1}^{p}$, subject to the above constraints on the coefficients. This solution is given by the vectors $\{\mathbf{X}\mathbf{a}_i\}_{i=1}^{p}$ where $\{\mathbf{a}_i\}_{i=1}^{p}$ are the eigenvectors of $\mathbf{S}$.

ii). Working with the configurations in Euclidean space defined by the $n \times p$ data matrices and seeking the *system of orthonormal vectors which are the successively "best-fitting" vectors to the configurations.* "Best-fitting" is taken to mean minimizing the sum of squared distances from each point in the scatter to the line defined by the selected vectors. This is essentially the approach taken by Pearson [48] in his 1901 paper. It should be noted that this approach (which Anderson (1958) [1, (pg.278-279)] also formulates in a probabilistic context) actually defines *two* distinct problems, depending on whether the configuration which is considered is the one in $\mathbb{R}^p$ or the one in $\mathbb{R}^n$. In $\mathbb{R}^p$ the "best-fitting" vectors are the (unit-norm) eigenvectors of the covariance matrix $\mathbf{S} = \mathbf{W}'\mathbf{W}$, whereas in $\mathbb{R}^n$ they are the (unit-norm) eigenvectors of $\mathbf{W}\mathbf{W}'$. The two sets of vectors are related as was seen in equation (I.4.17).

This is the approach with which PCA is introduced by, among others, Volle (1981) [67], Cailliez and Pagés (1976) [7] and Lebart *et. al.* (1979) [40] and (1984) [41]. It can also be formulated as looking for the "best-fitting" k-dimensional subspace (with $1 < k \leq p$) but in that case the first k eigenvectors of $\mathbf{W}\mathbf{W}'$ and $\mathbf{W}'\mathbf{W}$ are only one in an infinity of solutions, since any other orthonormal basis for the k-dimensional subspace will do. This indeterminacy can be eliminated by adopting a step-by-step approach: start with $k = 1$, fix the (unique) solution obtained, move up to $k = 2$ to obtain a second vector for the basis, and so on.

Other criteria optimized by PCs are considered in Okamoto (1969) [47] and can be used to define PCA. A further optimal property of PCs is considered in Chapter VI (Theorem VI.1).

Four sets of orthogonal vectors arise from these various approaches. Three are merely re-scalings of a set of p orthogonal vectors in $\mathbb{R}^n$ and the last is an orthonormal basis of $\mathbb{R}^p$:

- Hotelling's original Principal Components $\{\gamma_j\}_{j=1}^{p}$. They are the first p $\sqrt{n}$-norm eigenvectors of $\frac{1}{n}\mathbf{X}\mathbf{X}'$. We shall refer to them as **Hotelling's $\sqrt{n}$-norm Principal Components** or **unit-variance Principal Components** of the data.

- The global re-scaling of Hotelling's PCs given by the vectors of unit (standard) norm $\{\psi_j\}_{j=1}^{p}$. These will be called **unit-norm PCs or Principal Axes (PAs) in $\mathbb{R}^n$** of the data.

- The individual re-scalings of Hotelling's PCs given by the vectors $\{\mathbf{X}\mathbf{a}_j\}_{j=1}^{p}$. These will be called **natural length PCs**, **natural variance PCs** or simply **Principal Components** of the data. The variance of the j-th such PC is λ_j, the j-th eigenvalue of the covariance matrix of the data. Its norm is $\sqrt{n \cdot \lambda_j}$. These vectors are commonly known in the PCA literature as the vectors of **PC scores** for the n individuals.

- The eigenvectors of the covariance matrix $\mathbf{S} = \mathbf{W}'\mathbf{W}$, which form an orthonormal basis in $\mathbb{R}^p$ and which will henceforth be called the **(unit-norm) Principal Axes (PAs) in $\mathbb{R}^p$** of the data. Under the mapping from $\mathbb{R}^p$ to $\mathbb{R}^n$ defined by the matrix $\mathbf{X}$, their image is the set of natural length PCs.

The name Principal Components is used in connection with all of the above (see, for example, Krzanowski (1988) [38, (pg.64)] for the case of the vectors in $\mathbb{R}^p$).

It should be stressed that, by taking in equation (I.4.17) $\mathbf{Z} = \frac{1}{\sqrt{n}}\mathbf{X}'$, we can also obtain re-scaled PAs in $\mathbb{R}^p$ from the unit-norm PCs in $\mathbb{R}^n$. In fact, if $\{(\lambda_j, \boldsymbol{\psi}_j)\}_{j=1}^p$ are the non-zero-eigenvalue eigenpairs of $\mathbf{WW}'$, then $\{(\lambda_j, \frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\psi}_j)\}_{j=1}^p$ are eigenpairs of the covariance matrix of $\mathbf{X}$. Hence, we can speak of a fifth set of vectors connected with PCA:

- the **natural length PAs in $\mathbb{R}^p$** of the data, given by $\{ \frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\psi}_j\}_{j=1}^p$, where $\{\boldsymbol{\psi}_j\}_{j=1}^p$ are the unit-norm eigenvectors of $\mathbf{WW}'$ with non-zero eigenvalues. These natural length PAs in $\mathbb{R}^p$ are merely individual re-scalings of the unit-norm PAs in $\mathbb{R}^p$ of the data. The norm of the j-th natural length PA in $\mathbb{R}^p$ is $\sqrt{\lambda_j}$, thus incorporating information on its importance. These PAs are often used with Correlation Matrix PCA (see, for example, Jackson (1991) [29, (pg.16)]).

For data sets where the p columns of $\mathbf{X}$ represent p observations of functions at different values of their domain (rather than p different variables) and the n rows are independent sets of observations of those functions, the very notion of linear combinations of the p columns is usually not meaningful. In this context "PCA" usually refers *only* to the "Analysis in $\mathbb{R}^p$" (to use the terminology of Volle (1981) [67] or Lebart *et al* (1979) [40] and (1984) [41]), and the eigenvectors of the covariance matrix – which can be interpreted as other functions observed at p points – become the main center of interest. Such data sets will be considered in Section V.3.

A similar comment applies to "PCA in Function Spaces" which involves a "data set" of n functions – the p points become a continuum – and where interest focuses on the function space. For a discussion of these cases, see Ramsay (1982) [51] and Besse and Ramsay (1986) [5].

The term PCA is also used in connection with an analogous method, but without the assumption of variables centered about their means. We then speak of a **Non-centered PCA**. It is known (see Jolliffe (1986) [32, (pg.227)]) that a Non-centered PCA finds the successively "best-fitting" vectors crossing the origin of $\mathbb{R}^p$ *which is no longer previously forced to coincide with the center of gravity of the data configuration.* In $\mathbb{R}^n$ it involves working with any vectors and not just vectors which are orthogonal to $\mathbf{1}_n$. This corresponds to replacing references to the variances of the variables with references to their non-central second order moments. Non-centered PCA has been advocated as the appropriate variant of PCA in certain contexts (see, for example, Noy-Meir (1973) [46]).

Equally, we speak of a **Doubly-centered PCA** when both the columns *and* the rows of the data matrix are centered about their means (possibly with the global mean added back in). If only the row means (but not the column means) are subtracted from the data values we speak of a **Row-centered PCA**.

I.4.3 PCA and the Singular Value Decomposition

The most compact way of introducing PCA in our context is through the Singular Value Decomposition of the matrix $\mathbf{W} = \frac{1}{\sqrt{n}}\mathbf{X}$.

Using previous notation and based on the discussion in this and in the previous Sections, it is now clear that the SVD of the matrix $\mathbf{W}$ is given (in matrix form) by:

$$\mathbf{W} = \mathbf{\Psi}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{A}' \tag{I.4.21}$$

The first matrix in the decomposition, $(\mathbf{\Psi})$, has the *unit-norm PCs of* $\mathbf{X}$ in its columns. The second matrix, $(\mathbf{\Lambda}^{\frac{1}{2}})$ gives the *standard deviations associated to those PCs*. And the third matrix, $(\mathbf{A})$, has the *unit-norm Principal Axes in* $\mathbb{R}^p$ *of* $\mathbf{X}$ as its columns.

Thus, the fundamental concepts of PCA correspond to matrices in the SVD of matrix $\mathbf{W}$.

It can be readily seen that post-multiplying a data matrix by an orthogonal $p \times p$ matrix does not change its PCs or their natural lengths. If $\mathbf{B}$ is such an orthogonal matrix, the SVD of $\mathbf{WB}$ is simply:

$$\mathbf{WB} = \mathbf{\Psi}\mathbf{\Lambda}^{\frac{1}{2}}(\mathbf{A}'\mathbf{B})$$

since the product of orthogonal matrices is again an orthogonal matrix. Thus, 'rotating the data in $\mathbb{R}^p$' does not affect the analysis in $\mathbb{R}^n$.

From the results of Section I.3, we also know that the 'best' rank k approximation to $\mathbf{W}$, in the sense of being the closest rank k matrix to $\mathbf{W}$ with the Frobenius norm, is given by deleting the last $r-k$ columns of $\mathbf{\Psi}$ and $\mathbf{A}$ and the last $r-k$ rows and columns of $\mathbf{\Lambda}^{\frac{1}{2}}$. But this has the same effect as taking, in equation (I.4.16), $\mathcal{G}$ to be the subset formed by the first k unit-norm PCs. Hence, the above Frobenius norm criterion is merely stating that the 'best' rank k approximation to $\mathbf{W}$ is obtained by orthogonally projecting the p variables in $\mathbb{R}^n$ onto the subspace spanned by the first k PCs. This ties up with Pearson's approach to PCA in its 'k-dimensional subspace' form.

The SVD of $\mathbf{W}$ is therefore an ideal tool to discuss many aspects of PCA and it will be used extensively.

I.5 Problems with PCA

There are several well-known problems relating to Principal Component Analysis and its scope of application.

I.5.1 Units of measurement

A first problem arises if we seek to make all the algebraic relations of the previous Section meaningful and coherent also from the point of view of their units of measurement.

We begin by considering Hotelling's approach of viewing variables as linear combinations of components. Recalling (from equation (I.4.6)) that the coefficients $c_{ij} = \frac{1}{n} < \boldsymbol{\gamma}_i, \mathbf{x}_j >$ in these linear combinations are the covariances between the i-th component and the j-th variable, we see that both sides of equation (I.4.14) are only compatible in their units *if the components* $\{\boldsymbol{\gamma}_i\}_{i=1}^{p}$ *are dimensionless*. This, in turn, implies that the j-th element of $\mathbf{c}_i^{row}$ (equation (I.4.9)) has the same units as the j-th variable. In other words, the j-th column of matrix $\mathbf{C}$ (I.4.5) has the units of the j-th column of the data matrix $\mathbf{X}$. A quick check using the general formula for matrix inverses then reveals that the j-th row of $\mathbf{C}^{-1}$ has the reciprocal units of the j-th column of $\mathbf{X}$, thereby making equation (I.4.13) coherent from the point of view of physical dimensions. A further, somewhat minor, problem does occur if we consider the criterion used to obtain the components (see equation (I.4.8)). In fact, the norm of each row of $\mathbf{C}$ only has meaningful units *if all* p *variables have common physical dimensions* (including the case of adimensionality). In this case, equation (I.4.10) implies that the eigenvalues associated with each PC also have units, these being the squares of the (common) units of the variables.

This latter restriction plays a much more central role in the converse approach of defining components as linear combinations of variables. Assuming that the coefficients in such linear combinations are adimensional scalars, it is obvious that any linear combination of the variables is only coherent from the point of view of its units if all variables have common units of measurement. However, these units of measurement are then carried over to the natural length PCs. Since Hotelling's unit-variance PCs are obtained by dividing these natural length PCs by the square root of their associated eigenvalue (equation (I.4.18)), we are back to the adimensionality of Hotelling's PCs.

Thus, for both approaches, full coherence in terms of units of measurement is associated with the restriction that all variables have common physical dimensions. This discussion is naturally related to the decision on whether to carry out a Covariance or a Correlation Matrix PCA, since the standardization of any set of variables is one way of ensuring that they will henceforth have common (no) units of measurement.

We can plausibly brush aside this problem, focusing on the algebraic aspects of PCA rather than on the compatibility of physical dimensions. But more fundamental problems also exist as will now be seen.

I.5.2 Scale dependence

In previous sections, it was noted that PCA was insensitive to a global multiplicative change of scales and to (possibly different) additive changes of scales in the variables. However, PCA is *not* insensitive to different multiplicative changes of scale in each of the variables.

In fact, if the SVD of a re-scaled data matrix (as in (I.1.2)) is given by: $\mathbf{W} =$

$\mathbf{\Psi\Sigma A}'$, then post-multiplying $\mathbf{W}$ by a generic $p \times p$ diagonal matrix $\mathbf{D}_p$ will not leave the natural length PCs (the product $\sqrt{n}\mathbf{\Psi\Sigma}$) unchanged unless $\mathbf{D}_p$ is also an orthogonal matrix. But this can only happen when $\mathbf{D}_p$ has all diagonal entries equal to ± 1, that is, when the variables are simply left unchanged or reflected about the origin.

This is not surprising since a change of scales will re-shape the configurations of points in both $\mathbb{R}^n$ and $\mathbb{R}^p$. The length of the p vectors in $\mathbb{R}^n$ will, as we saw in Section I.2 , change. As can be easily visualized, this implies that the longest possible vector which can be obtained by a linear combination of the p vectors (with a unit-norm vector of coefficients for the linear combination), that is, the first PC, will in general have a different direction and magnitude.

A similar reasoning reveals that PCA is also not insensitive to different weightings of different individuals, that is, to pre-multiplying $\mathbf{W}$ by a diagonal matrix of different non-negative diagonal elements adding up to 1.

But the sensitivity to changes of scale in the variables is qualitatively more serious since attaching weights to individuals usually corresponds to an explicit and conscious decision by the analyst. In contrast, scales are often arbitrary conventions. For example, the Principal Components of a set of variables in different SI units will differ from the PCs of the same set of variables in metric system units. In the words of Chatfield and Collins (1980) [8, (pg.70)] :

> The practical outcome of the above result is that principal components are generally changed by scaling and that they are therefore not a unique characteristic of the data.

Kendall (1980) [35, (pg.19)] considers that this "distinctive feature of component analysis (...) seriously affects the use which we make of it in practice".

It should be noted that the issue of scale dependence also affects the related problem of determining the Principal Axes in $\mathbb{R}^p$. If our main interest lies with PCs (in $\mathbb{R}^n$), this is only of passing concern. But if our main interest is, say, to obtain a low-dimensional approximate representation of the data in $\mathbb{R}^p$, the implication is that – not unexpectedly – these representations are also scale-dependent.

However, for those data sets of a 'sampled function' nature, in which the main concern is with the Principal Axes in $\mathbb{R}^p$, this scale dependence is usually *not* a major source of trouble. Due to the 'functional' nature of the data, it is unlikely that different changes of scales at each of the p points of observation will be justified. In other words, even if the scales are globally arbitrary, they are not locally arbitrary. Hence, the invariance of results for any plausible change of scales is guaranteed.

Except for such cases where the scales of all p variables are not (individually) arbitrary, *scale dependence is undoubtedly the major drawback in PCA.*

I.5.3 The quest for invariance to re-scalings

Both problems mentioned above can be avoided if the data set is always transformed, prior to a PCA, into a set of dimensionless data.

Correlation Matrix PCA

By far the most common way of doing so is to standardize the data, as described in (I.1.3), that is, to carry out a Correlation Matrix PCA. It has already been mentioned that Hotelling (1933) [27] defined PCA with this additional assumption, although he did allow for the possible generalization to covariance matrices (pg.421).

Standardization amounts to a mapping of all possible re-scalings of a given $n \times p$ data matrix into a single data matrix, which is then subjected to PCA. More formally, we can say that the set of all $n \times p$ data matrices is partitioned into **equivalence classes**, with two matrices $\mathbf{X}_1, \mathbf{X}_2$ belonging to the same equivalence class if and only if there exists some diagonal matrix D with positive diagonal elements such that $\mathbf{X}_2 = \mathbf{X}_1\mathbf{D}$. Given a data matrix $\mathbf{X}$, a PCA is carried out, not on $\mathbf{X}$ itself, but on an element $\mathbf{XD}$ of $\mathbf{X}$'s equivalence class. This representative element is picked by taking $\mathbf{D}$ as given in equation (I.1.4).

But this approach raises other questions. A first issue is that *standardization is not the only way of making a data set dimensionless.* Gower (1966) [19, (pg.327)] makes the point that "other normalizers could be used, for example, the variate mean (when zero is not arbitrarily located), or the range, or even the cube root of the sample third moment". In fact, dividing each column's entries by any quantity with the same units of measurement will lead to dimensionless data. This means that the choice of representative from each equivalence class is somewhat arbitrary. Since PCA is scale dependent, different choices will produce different sets of principal components.

A second, more fundamental, problem is that standardization (like any of the alternatives that have just been mentioned) "avoids rather than solves the scaling problem", to quote Chatfield and Collins (1980) [8, (pg.71)]. In fact, *the principal components of the data matrix chosen to represent each equivalence class are not principal components for the other matrices in the class*, in the sense that they do not meet the requirements discussed in the previous Section. We have, in effect, discarded the original data set and replaced it with a more convenient data set, but in so doing, we can no longer directly relate the information obtained by a PCA to the original data set. Thus, the problem of scale dependence continues to plague us, although in a different form.

Logarithms and Kazmierczak's Analysis

Similar problems affect an alternative approach by Kazmierczak (1985) [34] which simultaneously addresses the problem of scale dependence and dependence on a set of weights which may be associated with each individual (row of the data matrix). This approach produces the same results if the original data matrix is replaced by the data matrix $\mathbf{D}_n\mathbf{Y}\mathbf{D}_p$ where $\mathbf{D}_n$ is an $n \times n$ diagonal matrix and $\mathbf{D}_p$ a $p \times p$ diagonal matrix, with all diagonal elements positive.

This method, which Kazmierczak called **Logarithmic Analysis**, can only be used – in its original form – when all the data values are positive. The method consists of the following three steps:

i). replacing each element of $\mathbf{Y}$ by its logarithm, *i.e.*, y_{ij} becomes $l_{ij} = \log y_{ij}$ (hence the restriction to positive data values).

ii). subtracting from each resulting element its column and row mean, and adding back the global mean of all elements in the matrix. Thus, each l_{ij} is replaced by $v_{ij} = l_{ij} - \bar{l}_{i\cdot} - \bar{l}_{\cdot j} + \bar{l}_{\cdot\cdot}$, with obvious notation.

iii). carrying out a PCA of the resulting matrix $\mathbf{V}$ with generic element v_{ij}

The invariance to re-weighting of rows and re-scaling of columns of $\mathbf{Y}$ is achieved by the first two steps. In fact, if each y_{ij} was replaced by $a_i y_{ij} b_j$ (where a_i, b_j are the diagonal elements of $\mathbf{D}_n$ and $\mathbf{D}_p$ respectively), then taking logarithms converts the new element into a sum $\log a_i + \log y_{ij} + \log b_j$ and the second stage will then cancel all terms involving the a's and the b's. When only the re-scaling of columns is of concern, a simple column-centering of the log-transformed data suffices to ensure the invariance. This is one of the reasons why log-transforms are popular in certain fields (see, for example, [31, (pg.497)]).

One specific drawback of Logarithmic Analysis is that it destroys PCA's insensitivity to *additive* changes of scale. We cannot even consider adding a preliminary step of column-centering the data matrix $\mathbf{Y}$, since this would produce negative data values.

It can be seen, however, that the logarithms do not really play a crucial role in ensuring Logarithmic Analysis' invariance to re-scaling and re-weighting. In fact, the properties of logarithms show that the generic element of matrix $\mathbf{V}$, as obtained in the second step of the above procedure is:

$$v_{ij} = \log \left(\frac{y_{ij}}{\left(\frac{\sqrt[n]{\prod_{i=1}^{n} y_{ij}} \cdot \sqrt[p]{\prod_{j=1}^{p} y_{ij}}}{\sqrt[np]{\prod_{i=1}^{n} \prod_{j=1}^{p} y_{ij}}} \right)} \right) \tag{I.5.1}$$

The invariance property is already guaranteed in the argument of the logarithm, as can be easily verified (not least because the logarithmic function is bijective in its domain). Likewise, logarithms are not required to make the data adimensional.

We can therefore drop step (i) of the procedure and suitably modify step (ii). Dispensing with logarithms allows for negative data values to be considered as well, as long as the geometric means of the rows, columns, and whole matrix which appear in the denominator of (I.5.1) are taken to refer to the absolute values of the data.

In other words, Logarithmic Analysis can be replaced by a more general procedure – which we shall refer to as **Kazmierczak Analysis** – consisting of two steps:

i). replacing each element y_{ij} of the data matrix $\mathbf{Y}$ with:

$$v_{ij} = \frac{y_{ij}}{\left(\frac{G_i^{row} . G_j^{col}}{G^{tot}} \right)} \tag{I.5.2}$$

where G_i^{row},G_j^{col} , G^{tot} are the geometric means of the absolute values of row i, column j and the whole matrix, respectively.

ii). carrying out a PCA of the resulting matrix **V**.

An important restriction on the data sets that can be considered remains, since any zero values will make expression (I.5.2) meaningless. If a prior column-centering stage (which is now acceptable in principle) is required, the restriction is that no data value can equal its column mean.

A quick look at the transformation (I.5.2) shows that, since the overall geometric mean is a constant feature, the relative importance of the y_{ij}'s that are in "large" rows and/or columns is decreased and the relative importance of the y_{ij}'s that are in "small" rows and/or columns is increased. In other words, Kazmierczak Analysis focuses essentially on the relative magnitude of each data observation y_{ij} in relation to the corresponding row (individual) and column (variable) sizes, as a means of by-passing the effects of re-weighting rows and re-scaling columns.

For our purposes, the crucial aspect to note is that both Logarithmic and Kazmierczak Analyses, like the other methods described in this Section, "avoid rather than solve the scaling problem". As with Correlation Matrix PCA, the components obtained are only principal components for the transformed data and not for the original data. Incidentally, the data transformations in both Logarithmic and Kazmierczak Analysis reduce the rank of the data matrix being analyzed.

The problems raised in this Section will be addressed in later Chapters. In Chapter IV we shall consider ways of comparing the results of PCAs of a given data set which has been made adimensional in different ways. This paves the way for a fuller understanding of the implications of each choice. In Chapter VI we address the whole issue of obtaining a truly scale-independent set of components for a given data set.

Chapter II

Interpreting PCs

PCA as a descriptive method for exploratory data analysis is probably at its most useful when the new "fundamental set of independent variables" – as Hotelling called the Principal Components – or at least the first few of these PCs can be interpreted as meaningful quantities or indicators for the problem which is being studied.

II.1 The standard approach

The interpretation of PCs, as it is usually carried out, is based on the notion that it is the p original variables, or some subset of them, which can provide 'meaning' to a PC. Thus, attention is focused on the linear combinations of the original variables which define the natural length PCs.

A *first stage* towards interpretation consists in assessing whether a given PC can be adequately approximated by fewer than p variables. This is done by looking at the coefficients (loadings) of each variable in the linear combination and discarding those terms for which the loadings are of small magnitude. A *second stage* then focuses on either the PC itself or the 'truncated PC', whichever emerged from the first stage. This vector is now viewed as either a weighted average or a contrast of those variables which were retained, depending on whether or not their loadings are all of the same sign. A 'meaningful' PC is a PC for which the variables retained or the variables on either side of a contrast can be viewed as having a common feature of relevance to the problem which is being studied. The specificities of each particular problem are undoubtedly of crucial importance in this second stage.

Sometimes these two stages are preceded by transformations of the PCs ('rotations') which raise a number of issues. Such transformations, which are discussed at length in Chapter 8 of Jackson (1991) [29] , shall not be dealt with here.

An example of interpretation of PCs is provided on page 20 of Kendall (1980) [35]. The data set consists of $n = 20$ samples of soil on which $p = 4$ measurements are taken: silt content ($\mathbf{y}_1$), clay content ($\mathbf{y}_2$), organic matter ($\mathbf{y}_3$) and acidity ($\mathbf{y}_4$). The first three variables are percentages, whilst the fourth is in the pH scale. Hence, the scale of the fourth variable is not comparable with those of the other variables. We saw in Chapter I that this should suggest a Correlation, rather than a Covariance, Matrix PCA. But we proceed with Kendall's approach. The data set, its covariance and correlation matrices and the loadings for the centered variables $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ on the four PCs are all reproduced in Tables II.1 to II.3 (with the appropriate replacement of n -1 with n in the denominator of variances and covariances).

Sample	Silt content (%)	Clay content (%)	Organic matter (%)	Acidity (pH)
number	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$
1	13.0	9.7	1.5	6.4
2	10.0	7.5	1.5	6.5
3	20.6	12.5	2.3	7.0
4	33.8	19.0	2.8	5.8
5	20.5	14.2	1.9	6.9
6	10.0	6.7	2.2	7.0
7	12.7	5.7	2.9	6.7
8	36.5	15.7	2.3	7.2
9	37.1	14.3	2.1	7.2
10	25.5	12.9	1.9	7.3
11	26.5	14.9	2.4	6.7
12	22.3	8.4	4.0	7.0
13	30.8	7.4	2.7	6.4
14	25.3	7.0	4.8	7.3
15	31.2	11.6	2.4	6.5
16	22.7	10.1	3.3	6.2
17	31.2	9.6	2.4	6.0
18	13.2	6.6	2.0	5.8
19	11.1	6.7	2.2	7.2
20	20.7	9.6	3.1	5.9

Table II.1: Kendall's soil data

	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$
$\mathbf{y}_1$	75.751	(0.674)	(0.213)	(0.024)
$\mathbf{y}_2$	21.265	13.129	(-0.196)	(0.013)
$\mathbf{y}_3$	1.450	-0.555	0.611	(0.079)
$\mathbf{y}_4$	0.105	0.024	0.031	0.250

Table II.2: Covariance (and Correlation) matrix for Kendall's soil data

The j-th column in Table II.3 gives the loadings for the j-th PC of the data in Table

Variable	Eigenvectors			
	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_4$
$\mathbf{x}_1$	0.956	-0.288	-0.059	0.006
$\mathbf{x}_2$	0.294	0.945	0.142	-0.018
$\mathbf{x}_3$	0.015	-0.154	0.979	-0.136
$\mathbf{x}_4$	0.001	-0.002	0.137	0.991
Eigenvalues	82.308	6.739	0.448	0.246
Relative eigenvalues	0.9172	0.0751	0.0050	0.0027

Table II.3: Eigenpairs of the soil data's Covariance Matrix

II.1. The i-th row in Table II.3 gives the loadings associated with the i-th variable for all p PCs. The fact that the loadings in the main diagonal of Table II.3 are all very large is a well-known consequence of the fact that the variances in Table II.2 are of very different sizes (see Jolliffe (1986) [32, (pg.46)]).

Kendall considers the first PC and says ([35, (pg.22)]):

> Can we attribute any meaning to this component, or is it just an artifact conjured up by the analysis? This is the most difficult question to answer in component analysis and one to which there is rarely a simple unambiguous reply to be given; From the size of the coefficients of [$\mathbf{x}_3$] and [$\mathbf{x}_4$] it looks as if this component, if it has any 'reality', is simply the physical quality of the soil, expressible in terms of silt and clay (...) and independent of organic matter and acidity. We might provisionally call it silt content. The second component accounts for very much less, only about 7.5 percent of the variation and the coefficients would lead us to

regard it as dominated by $[\mathbf{x}_2]$, namely clay.

Most authors would probably also focus on the very high loadings in the main diagonal of Table II.3 and consider that each PC is associated with a single variable. This conclusion would follow, for example, from using the criterion in Jeffers (1967) [30] of retaining, for each PC, those variables whose loading exceeds 70% of the highest loading for that PC (regardless of sign).

Interpretation of PCs is, to some extent, subjective and several authors (not least Kendall as we shall see below) have expressed reservations about it. Some of these criticisms question the very idea of seeking to interpret PCs.

Jolliffe (1986) [32, (pg.51)] points out that:

> There is no reason, *a priori*, why a mathematically derived linear function of the original variables (which is what PCs are) should have a simple interpretation. It is remarkable how often it seems to be possible to interpret the first few PCs, though it is probable that some interpretations owe a lot to the analyst's ingenuity and imagination.

Kendall points out ([35, (pg.23)]) that even if a 'meaningful' interpretation is found, the scale dependence of PCs which was discussed in Section I.5 will make the interpretation pointless if the scales are arbitrary. This specific criticism is not so much a criticism of interpretation of PCs as a fundamental criticism of PCA itself. We shall address this problem in Chapter VI.

Chatfield and Collins (1980) [8] also consider the scale problem, among other difficulties. Clearly unimpressed by the fairly common practice of transforming ('rotating') PCs to make for easier interpretations they say ([8, (pg.73)]) that "in trying to identify components in this sort of way, the PCA often seems to be an end in itself" and conclude that "it is rather dangerous to try and read too much meaning into components".

These criticisms raise important points for deciding whether the interpretation of a given PC is a valid exercise. They qualify, rather than reject, the practice of interpreting PCs. But, with few exceptions, (for example, Jackson (1991) [29], who mentions the problem in Section 7.3) they have not focused on what we have called the first stage of the process, that is, on the use of PC loadings to decide whether the interpretation of the PC can be based only on a subset of k of the original variables and the subsequent replacement of the PC with a 'truncated PC'. It is to this stage of the process that we now turn our attention.

II.2 Loadings and PCs

When 'near-zero' loadings are ignored for the purpose of interpretation they are, in effect, being replaced with exactly zero loadings. The resulting 'truncated PC' is then a vector in the subspace of $\mathbb{R}^n$ spanned by the k retained variables.

The rationale for ignoring 'near-zero' loadings is that these will correspond to only minor displacements in the direction of the variables which they multiply and can therefore be safely discarded. In other words, the PC is 'almost' on the subspace spanned by the k variables with largest magnitude loadings. Replacing that PC with the 'truncated PC' which *is* in that subspace should be a valid approximation which may make any subsequent interpretations easier.

We will see in this Section that this premise is not generally true for Covariance Matrix PCs since it does not take into account the variances (*i.e.*, the lengths in $\mathbb{R}^n$) of each variable. A less serious problem with the method, for both Covariance and Correlation Matrix PCs, is that it does not take into account the pattern of correlations (*i.e.*, the relative positions in $\mathbb{R}^n$) of the variables.

In order to proceed with this discussion, it is essential to have some means of gauging how close a PC and an approximation to the PC actually are. The criterion we shall follow is to require that the two vectors be strongly correlated, that is, that the angle between them (in $\mathbb{R}^n$) be small.

The use of correlations as a measure of good approximation has, among others, the following advantages:

i). it provides an easily interpretable measure of closeness with values whose magnitude lies between zero and one and which is comparable for different approximations of different PCs (unlike, say, the distance – norm of the difference – between a PC and its approximation);

ii). it is relevant for a similar discussion in an inferential context, where similar conclusions will follow;

iii). it makes the precise definition of PCs which is used (unit-norm, unit variance, natural length) irrelevant;

iv). it has a track record in this context. Both Jolliffe (1986) [32, (pg.229)] and Jackson (1991) [29, (pg.145;pg.361)] use correlations as a criterion for an acceptable approximation to a given PC or to a vector of loadings of a given PC. (Jackson also required that any approximation to a PC "should be simple to use" and should have small correlations with all other approximations to any of the remaining PCs in a given data set. We shall not worry about this second requirement. It can be implicitly incorporated by requiring sufficiently high correlations between each PC and its approximation).

Let us then return to Kendall's soil data (Table II.1) and reconsider the association between PCs and variables.

The conclusion that the second PC is 'dominated' by clay content ($\mathbf{x}_2$) should mean that $\mathbf{x}_2$ and the second PC are highly correlated. But that is not the case.

It is well known (see, for example, paragraph 8.2.4 in Mardia, Kent and Bibby (1979) [42]) and can easily be deduced from equations (I.4.6) and (I.4.20) that the correlation between the i-th variable and the j-th PC is given by:

$$\rho_{ij} = a_{ij}\sqrt{\frac{\lambda_j}{s_{ii}}} \qquad \text{(II.2.1)}$$

where a_{ij} is the generic element of the matrix $\mathbf{A}$, whose j-th column gives the loadings for the j-th PC ; λ_j is the eigenvalue associated with that PC ; and s_{ii} is the variance of the i-th variable.

Incidentally, the ambiguity of sign-switching discussed in Section I.4 also implies that for any given PC, it is the sign patterns of the correlations (II.2.1), rather than the signs themselves, that are relevant.

In Kendall's soil data, the correlation between the second PC and the second variable is only 0.667 – far from the convincing performance which the very high loading (0.945) would have us believe. It implies that the angle in $\mathbb{R}^n$ between the PC and the variable which is supposed to 'dominate' it exceeds 47°. Even by relatively flexible standards this must mean that there is more to the second PC that just clay content. All the more so, since $\mathbf{x}_2$ actually has a higher correlation with the *first* PC (0.7353) than with the second PC (despite having a loading of only 0.294 for that first PC!). A 2-dimensional plot of $\mathbf{x}_2$ *vs.* the second PC will also convince us that this rather low correlation is indeed reflecting a lack of any particular relation in this n-point scatter, rather than some *non-linear* relation.

It is instructive to take a closer look at the correlations between each PC and all the variables, which are given in Table II.4.

	PC1	PC2	PC3	PC4
$\mathbf{x}_1$	0.9963	-0.0859	-0.0045	0.0004
$\mathbf{x}_2$	0.7353	0.6772	0.0261	-0.0025
$\mathbf{x}_3$	0.1737	-0.5107	0.8376	-0.0862
$\mathbf{x}_4$	0.0239	-0.0101	0.1840	0.9826

Table II.4: Correlations between variables and PCs for Kendall's soil data

These correlations suggest several comments:

i). Although $\mathbf{x}_1$'s loading for the first PC (0.956) is only slightly larger than $\mathbf{x}_2$'s loading for the second PC (0.945), the correlation of the former pair of vectors is very much higher (0.9963). Thus, it is justifiable to view the first PC as a 'silt content' PC. The general conclusion is that *similar loadings, even very large ones, may correspond to a very different importance of the respective variables in approximating different PCs.* The reason for this is apparent from equation (II.2.1): we must also consider the ratio of the standard deviations of the PC and the variable.

ii). The converse of the above general conclusion is also true: *very different loadings may correspond to a similar importance of a given variable, for different PCs.* This is the case with $\mathbf{x}_2$ in the first and second PCs.

iii). The correlation between the second PC and $\mathbf{x}_3$ is not much less significant than the same PC's correlation with $\mathbf{x}_2$. This would not have been guessed by looking at the magnitude of the loadings of $\mathbf{x}_3$ (-0.154) and $\mathbf{x}_2$ (0.945). In fact, by most standards, the loading of the third variable would probably be considered too small to warrant any attention. Again, equation (II.2.1) explains this apparently strange behaviour: $\mathbf{x}_3$'s standard deviation is also significantly smaller than that of $\mathbf{x}_2$. The general conclusion is that *a given PC may be similarly correlated with a high- and with a low-loading variable.*

iv). The previous conclusion also suggests that *variables with similar loadings for a given PC can be very differently correlated with that PC.* Although there is no particularly convincing example of that in this data set, the case of variables $\mathbf{x}_2$ and $\mathbf{x}_4$ in the third PC provides some evidence for this possibility.

v). The variable ($\mathbf{x}_1$) with the second highest magnitude loading (-0.228) for the second PC is almost orthogonal to that PC (its correlation is -0.0859). From this and the previous comment we can draw the general conclusion that *the order of the loadings need not correspond to the order of the variables' correlation with the PC.* Again, from equation (II.2.1) it is obvious that the relative size of the variables' standard deviation must also be considered.

vi). The fourth PC is strongly correlated with the fourth variable (acidity). Thus, *the interpretability of a PC does not necessarily depend on the proportion of the total variance it accounts for* (0.0027 in the case of the fourth PC).

The central point that is being made here is that the variances of variables and components must always be taken into account when judging the loadings. This point can also be visualized in $\mathbb{R}^n$, with the soil data's first PC as an example. The vector for the first variable (silt content) is almost two and a half times longer than the vector corresponding to $\mathbf{x}_2$ (clay content). Its loading is more than three times larger than that of $\mathbf{x}_2$. Thus, the linear combination of these two variables which is defined by those loadings amounts to the sum of two vectors one of which is more than 8 times longer than the other and will therefore dominate the sum. But if the same loadings were swapped, we would be talking about a sum of two vectors of approximately equal size. The resultant would be approximately the bisector of the angle between the two vectors.

The consequences of this for the *first* PC of a Covariance Matrix PCA deserve to be singled out. As is known, the largest coefficients of the first eigenvector of a covariance matrix correspond to the variables with largest variances. Combining large loadings with long vectors implies that *the first PC of a Covariance Matrix PCA* is indeed associated with the large-loading variables, but *to an even larger degree than the loadings alone appear to indicate.*

The case of Correlation Matrix PCA should also be mentioned. Since the variance of all variables is then 1, all vectors representing variables are of equal length in $\mathbb{R}^n$. Many of the problems raised above will then not occur. In particular, for any given PC, the order of each variable's loading and of its correlation with the PC will coincide. And if one variable's loadings is twice that of another, their correlations with the PC will also be in the same proportion. But eigenvalues will continue to play an important role in equation (II.2.1). Thus, equal loadings for different PCs may still correspond to very different roles of their associated variables in approximating those PCs. And a near-one loading in a near-zero-eigenvalue PC need not mean that the PC is dominated by that variable. They could even be nearly orthogonal, all depending on the relative size of $\sqrt{\lambda_j}$. It is for this reason that the loadings in Correlation Matrix PCA are sometimes multiplied by $\sqrt{\lambda_j}$, thus in effect giving the correlation of the respective variable with the PC. These are the 'natural length PAs in $\mathbb{R}^p$' which were mentioned in Subsection I.4.2. This amounts to re-scaling PCs so that their norm is proportional to their variance, rather than their standard deviation. For Covariance Matrix PCs, it is not possible to re-scale loadings in a similarly simple fashion. From equation (II.2.1) we see that different loadings in the same PC would have to be multiplied by different scalars in order to obtain the correlations.

From what has been said above it is already clear that loadings can be misleading when judging the role played by each variable in approximating a Covariance Matrix PC. However, the above discussion has only addressed part of the problem. Correlations between individual variables and any given PC are appropriate indicators to confirm or reject interpretations of a PC 'dominated' by a single variable. But they are not sufficient to assess the validity of approximations based on a choice of $k > 1$ variables. Unless the variables themselves are orthogonal (in which case a PCA is pointless) it is not possible to determine how close a PC is to its 'truncated' approximations merely by looking at the correlations between individual variables and PCs.

This can again be illustrated with Kendall's soil data. If clay content ($\mathbf{x}_2$) is not, on its own, sufficient for a good approximation of the second PC the obvious question is whether a single additional variable can do the job. If the criterion to 'recruit' a new variable is the strongest correlation (of either sign) with the PC, then organic matter ($\mathbf{x}_3$) must be our choice. The new 'truncated PC' will then be $\mathbf{v} = 0.945\mathbf{x}_2 - 0.154\mathbf{x}_3$. But the correlation between $\mathbf{v}$ and the second PC is only marginally better: 0.6896. Curiously, if we follow the usual criterion of loadings and choose to include the term in $\mathbf{x}_1$ instead, the correlation between the PC and this alternative approximation $\mathbf{v} = -0.288\mathbf{x}_1 + 0.945\mathbf{x}_2$ improves dramatically to 0.9992.

Thus, if loadings alone had already been seen to be an unreliable means of determining a good 'truncated PC', it is now clear that the correlations between that PC and each variable will not be adequate either, whenever the number of variables needed to ensure a good approximation exceeds one.

That neither approach is satisfactory is confirmed by looking at the third PC in the same example. Although this PC would probably not be interpreted in a practical application - due to the small percentage of total variance it accounts for (0.5%) - we have already seen that there is nothing to prevent us from following through the same steps.

Judging from loadings alone, this third PC would appear to be dominated by organic matter ($\mathbf{x}_3$). But the correlation between these two variables (0.8376), although much better than that between $\mathbf{x}_2$ and the second PC, is still unsatisfactory if we are to speak of a PC 'dominated' by $\mathbf{x}_3$. The angle between them in $\mathbb{R}^n$ is approximately 33°. (But see also Jackson (1991) [29, (pg.147)] for a more lenient attitude towards correlations of this order of magnitude).

If we are to add new terms to the 'truncated PC' according to their loadings, the most sensible choice would appear to be the terms in both $\mathbf{x}_2$ and $\mathbf{x}_4$, whose loadings are nearly equal. If we are to add the term in the variable second best correlated with the PC, then $\mathbf{x}_4$ alone seems adequate. In both cases, the triplet $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ seems to be the best 3-variable subset for approximating the third PC. However, the correlation between $\mathbf{v} = 0.142\mathbf{x}_2 + 0.979\mathbf{x}_3 + 0.137\mathbf{x}_4$ and the third PC actually *drops* to 0.7923, a worse approximation than with $\mathbf{x}_3$ alone.

On the other hand, if the term in $\mathbf{x}_4$ is replaced with the term in $\mathbf{x}_1$, the correlation with the PC grows to a substantial 0.9948 (an angle of just under 6° in $\mathbb{R}^n$). The implication is that $\mathbf{x}_1$ plays a significant role in ensuring a good approximation of the third PC despite its negligible loading (-0.0590) and correlation with that PC (-0.0045). The size (standard deviation) of this first variable is having its say.

The magnitude of the correlations between each PC of Kendall's soil data set and the sum of the terms in the linear combinations which define them, involving all two and three variable subsets, are given in Table II.5. These values complement those in Table II.4 to give a complete picture of the correlations between each PC and all its possible approximations via 'truncated PCs'.

Variables	PC1	PC2	PC3	PC4
$\mathbf{x}_1,\mathbf{x}_2$	1.0000	0.9992	0.0380	0.0037
$\mathbf{x}_1,\mathbf{x}_3$	0.9962	0.1092	0.7790	0.0844
$\mathbf{x}_1,\mathbf{x}_4$	0.9963	0.0859	0.0290	0.9740
$\mathbf{x}_2,\mathbf{x}_3$	0.7388	0.6900	0.7850	0.0822
$\mathbf{x}_2,\mathbf{x}_4$	0.7354	0.6772	0.0502	0.9760
$\mathbf{x}_3,\mathbf{x}_4$	0.1740	0.5104	0.8448	0.9950
$\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3$	1.0000	1.0000	0.9948	0.1012
$\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_4$	1.0000	0.9992	0.0677	0.9771
$\mathbf{x}_1,\mathbf{x}_3,\mathbf{x}_4$	0.9962	0.1092	0.7878	0.9913
$\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4$	0.7388	0.6900	0.7923	0.9938

Table II.5: Magnitude of Correlations between PCs and all 2- and 3-term 'truncated PCs' for Kendall's soil data

A general conclusion to be drawn from looking at the soil data's third PC is that *adding new terms to a 'truncated PC' might actually make the approximation worse.*

In order to understand this apparently strange conclusion, let us generalize equation (II.2.1) to give the correlation between any PC, given as $\mathbf{X}\mathbf{a}_j$, and a 'truncated PC' given by $\mathbf{X}\mathbf{a}_{\mathcal{Z}}$, where $\mathbf{a}_{\mathcal{Z}}$ is the vector $\mathbf{a}_j$ with some subset of p -k of its coefficients set equal to zero. We have:

$$r_t = \operatorname{corr}(\mathbf{X}\mathbf{a}_j, \mathbf{X}\mathbf{a}_{\mathcal{Z}}) = \frac{\mathbf{a}_j'\mathbf{S}\mathbf{a}_{\mathcal{Z}}}{\sqrt{\lambda_j}\cdot\sqrt{\mathbf{a}_{\mathcal{Z}}'\mathbf{S}\mathbf{a}_{\mathcal{Z}}}} = \frac{\sqrt{\lambda_j}\cdot\mathbf{a}_j'\mathbf{a}_{\mathcal{Z}}}{\sqrt{\mathbf{a}_{\mathcal{Z}}'\mathbf{S}\mathbf{a}_{\mathcal{Z}}}} \qquad \text{(II.2.2)}$$

where $\mathbf{S}$ is the covariance matrix determined by $\mathbf{X}$.

The numerator in (II.2.2) is $\sqrt{\lambda_j}$ times the sum of squares of the loadings that have been retained. Discarding the smallest magnitude loadings, as is usually done, will maximize the numerator (for any fixed number k of retained variables) and thus contribute to maximize the correlation. But this advantage must be weighted against the consequences in the denominator, which are far more complex. General trends which help to minimize $\mathbf{a}_{\mathcal{Z}}'\mathbf{S}\mathbf{a}_{\mathcal{Z}} = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{i\mathcal{Z}}a_{j\mathcal{Z}}s_{ij}$ – and therefore maximize (II.2.2) – are to retain loadings for which the product $a_{i\mathcal{Z}}a_{j\mathcal{Z}}s_{ij}$ is negative, or else *small* magnitude loadings and loadings which correspond to 'small' rows/columns of $\mathbf{S}$ (that is, to low-variance variables and variables which are nearly uncorrelated with the rest).

That the results of these different effects are not straightforward is what we have been seeing.

Formula (II.2.2), like its particular case (II.2.1) when $a_{\mathcal{Z}}$ has a single non-zero entry, does not require prior knowledge of the data set. Knowing the covariance matrix for the data is all that is really required.

As was seen before, the covariances between the variables retained will play a role in determining the denominator of (II.2.2). Thus, some tricky effects can be expected even when we are dealing with a Correlation Matrix PCA.

To summarize, we have seen that loadings are not reliable to determine whether some subset of the p original variables can provide an acceptable 'truncated PC', in particular for Covariance Matrix PCA. Likewise, correlations between individual variables and PCs are not appropriate except when judging the adequacy of single-variable approximations.

In the next Section we shall consider a possible alternative.

II.3 Corrected loadings

In Section II.2 we saw how the varying size of the p vectors representing the variables in $\mathbb{R}^n$ played a major part in distorting the impressions given by the loadings of Covariance Matrix PCs.

A possible way around this problem is to re-write the PC as a linear combination of the *standardized* – rather than the original – variables. In other words, the j-th PC is re-written as:

$$\mathbf{X}\mathbf{a}_j = \mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}(\mathbf{D}_S^{\frac{1}{2}}\mathbf{a}_j)$$

where $\mathbf{D}_S$ is the diagonal matrix of variances defined in equation (I.1.4). The vector $\mathbf{D}_S^{\frac{1}{2}}\mathbf{a}_j$, which is *not* an eigenvector of the correlation matrix $\mathbf{R} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}\mathbf{D}_S^{-\frac{1}{2}}$, is now a vector of coefficients for a linear combination of vectors of equal size – the columns of $\mathbf{Z} = \mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}$. For greater ease in assessing the role of these new coefficients, we can re-scale the vector to have unit-norm. Thus, we obtain a vector of **corrected loadings:**

$$\mathbf{a}_j{}^c = \frac{\mathbf{D}_S^{\frac{1}{2}}\mathbf{a}_j}{||\mathbf{D}_S^{\frac{1}{2}}\mathbf{a}_j||} \qquad \text{(II.3.1)}$$

The fact that the set of p vectors of corrected loadings $\{\mathbf{a}_j^c\}_{j=1}^p$ is no longer orthonormal and that the corresponding PCs which they define have all been re-scaled by different factors will not be a problem. Corrected loadings are to be used merely as a numerical artifact in attempting to incorporate the influence of different-sized vectors into the coefficients.

For Correlation Matrix PCA, equation (II.3.1) will produce no changes in the loadings. For Covariance Matrix PCA, the corrections in the loadings will tend to be greater for data sets with greater differences in the variances of the variables.

The corrected loadings for Kendall's soil data are given in Table II.6.

	PC1	PC2	PC3	PC4
$\mathbf{z}_1$	0.992	-0.590	-0.486	0.108
$\mathbf{z}_2$	0.127	0.807	0.485	-0.128
$\mathbf{z}_3$	0.001	-0.028	0.724	-0.207
$\mathbf{z}_4$	0.000	0.000	0.065	0.964

Table II.6: Corrected loadings for the PCs of Kendall's soil data

Some of the problems raised during the discussion in Section II.2 have been overcome. The significant role of the first variable in approximating the second PC is now apparent. Equally, for the third PC, it is clearly the triplet ($\mathbf{x}_1$,$\mathbf{x}_2$,$\mathbf{x}_3$) which deserves our attention. The dominant variables in the first and fourth PCs retain high corrected loadings, in accordance with their crucial role in approximating these PCs.

Apparently, retaining the variables whose corrected loadings are above some relatively small threshold – say 0.25 in magnitude – will produce the k-variable subsets we would probably have chosen by inspecting Tables II.4 and II.5.

But a closer look at the corrected loadings will show that not all has turned out well.

For example, unlike what happens with Correlation Matrix PCs, the corrected loadings of any given PC are not proportional to the PC's correlations with each variable. In fact, from equation (II.3.1) we have:

$$a_{ij} = a_{ij}^{c} \cdot \frac{||\mathbf{D}_S^{\frac{1}{2}}\mathbf{a}_j||}{\sqrt{s_{ii}}}$$

Hence, re-writing equation (II.2.1) in terms of corrected loadings gives, for the correlation between the i-th variable and the j-th PC:

$$\rho_{ij} = a_{ij}^{c} \frac{k_j}{s_{ii}}$$

with $k_j = \sqrt{\lambda_j} \cdot ||\mathbf{D}_S^{\frac{1}{2}}\mathbf{a}_j||$. Not even the rankings of the corrected loadings and correlations need be the same.

Additionally, the order of the corrected loadings' magnitudes need not imply that the corresponding variables' contribution to the PC are ranked in the same order. For example, Table II.5 tells us that if we add a single extra term to improve on the approximation to the third PC provided by $\mathbf{x}_3$, it should be the term in $\mathbf{x}_4$. But that is the term with the smallest magnitude corrected loading.

Thus, corrected loadings may have a role to play in fulfilling our original goal: to detect whether there are negligible displacements of a given PC in the direction of certain variables. Near-zero corrected loadings may indicate that such is the case. But two words of warning are immediately obvious: (i) we should not try to read any more than this into corrected loadings, for they do not reflect the fine print in the relations between PCs and variables ; (ii) any choice of a k -variable subset of the p variables to approximate a given PC should always be cross-checked by computing the correlation between the PC and the vector chosen to approximate it.

II.4 Regressing PCs on variables

So far, we have only considered the role of loadings in deciding which k-variable subset should be associated to each PC. Given such a subset, the PC has then been approximated by the 'truncated PC', that is, by the sum of those terms of $\mathbf{X}\mathbf{a}_j = \sum_{i=1}^{p} a_{ij}\mathbf{x}_i$ corresponding to the k selected variables. We shall see in this Section that this choice will *not*, in general, give us the best approximate PC using those k variables. An alternative way of approximating the PC will be considered.

At the heart of this alternative method lies the realization that in seeking to approximate a given variable (the PC) with a linear combination of some set of k other variables we are, in effect, confronting the problem which the time-honoured method of **multiple linear regression** addresses.

Given the subset of k variables, the multiple linear regression equation will provide the 'best' approximation to the PC in the least squares sense. In our linear algebraic context, this approximation is the orthogonal projection of the PC on the subspace of $\mathbb{R}^n$ spanned by the k variables (see, for example, Draper and Smith (1981) [11, (pg.201)]).

The correlation between the PC and its projection is the **multiple correlation coefficient** of the k variables with the PC. Its square is the R^2 statistic often used in regression to judge the adequacy of the fit. This provides further justification for our decision to use the correlation between a PC and an approximating vector to judge its usefulness.

It is known that the multiple correlation coefficient maximizes the correlation between a given variable and any linear combination of another set of variables (see, for example, Anderson (1958) [1, (pg.32)], who actually uses this property to define the multiple correlation coefficient). This implies that the regression equation is also the best approximation to the PC (for the given set of k variables) by the criterion we have chosen to use. It therefore implies that the 'truncated PCs' with which we have worked in the previous Section are sub-optimal. Except when $k = 1$, the projected vectors will generally provide approximations which are better correlated with the PCs than the 'truncated PCs' obtained by setting some loadings to zero.

In order to have some idea of how significant these improvements may be, we shall compute the multiple correlation coefficients of all four PCs of the soil data set with all 10 two- and three-variable subsets of the four original variables.

We recall (see, for example, Anderson (1958) [1, (Section 4.4.1)]) that the multiple correlation coefficient of a variable $\mathbf{u}$ with a set of k other variables $\{\mathbf{t}_i\}_{i=1}^{k}$ is given by:

$$r_m = \sqrt{\frac{\mathbf{s}'\mathbf{S}_t^{-1}\mathbf{s}}{s_{uu}}}$$

where $\mathbf{S}_t$ is the covariance matrix of the k variables $\{\mathbf{t}_i\}_{i=1}^{k}$, s_{uu} is the variance of the variable $\mathbf{u}$ and $\mathbf{s}$ is the k x 1 vector of covariances between $\mathbf{u}$ and each $\mathbf{t}_i$.

If $\mathbf{u}$ is taken to be the j-th PC, $\mathbf{Xa}_j$, and $\{\mathbf{t}_i\}_{i=1}^{k}$ a subset of k of the original p variables whose indices form the set $\mathcal{G}$, *i.e.*, $\{\mathbf{x}_i\}_{i\in\mathcal{G}}$, then we know from equation (II.2.1) that the correlation between the PC and the variable $\mathbf{x}_i$ is given by $a_{ij} \cdot \sqrt{\frac{\lambda_j}{s_{ii}}}$. Their covariance will then be:

$$\mathrm{cov}(\mathbf{Xa}_j, \mathbf{x}_i) = \lambda_j a_{ij}$$

If we now define $\mathbf{a}_j{}^{\mathcal{G}}$ to be the k-dimensional vector whose k entries are the entries of $\mathbf{a}_j$ corresponding to indices in $\mathcal{G}$, we see that $\mathbf{s} = \lambda_j \mathbf{a}_j{}^{\mathcal{G}}$. The variance s_{uu} is merely λ_j. And the covariance matrix, which we shall henceforth denote by $\mathbf{S}_{\mathcal{G}}$, is the submatrix of $\mathbf{S}$ defined by the rows/columns corresponding to variables with indices in $\mathcal{G}$. In other words, *the multiple correlation coefficient between the j-th PC and the* k *variables defined by subset* $\mathcal{G}$ is given by:

$$r_m = \sqrt{\lambda_j} \cdot \sqrt{\mathbf{a}_j{}^{\mathcal{G}\prime}\mathbf{S}_{\mathcal{G}}{}^{-1}\mathbf{a}_j{}^{\mathcal{G}}} \qquad \text{(II.4.1)}$$

As with equation (II.2.2), it is not necessary to know the data set in order to compute these multiple correlations. All that is needed is the covariance matrix.

The values of equation (II.4.1) for Kendall's soil data are given in Table II.7. They are to be compared with those in Table II.5. The corresponding values for single variable subsets are naturally the same as in Table II.4.

Variables	PC1	PC2	PC3	PC4
$\mathbf{x}_1,\mathbf{x}_2$	1.0000	0.9992	0.0398	0.0037
$\mathbf{x}_1,\mathbf{x}_3$	0.9971	0.5112	0.8583	0.0883
$\mathbf{x}_1,\mathbf{x}_4$	0.9963	0.0863	0.1842	0.9829
$\mathbf{x}_2,\mathbf{x}_3$	0.8036	0.7792	0.8598	0.0884
$\mathbf{x}_2,\mathbf{x}_4$	0.7355	0.6775	0.1855	0.9827
$\mathbf{x}_3,\mathbf{x}_4$	0.1740	0.5116	0.8459	0.9962
$\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3$	1.0000	1.0000	0.9948	0.1021
$\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_4$	1.0000	0.9992	0.1886	0.9829
$\mathbf{x}_1,\mathbf{x}_3,\mathbf{x}_4$	0.9971	0.5121	0.8666	0.9963
$\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4$	0.8037	0.7793	0.8672	0.9974

Table II.7: Multiple correlations between each PC of the soil data and the 2- and 3-variable subsets

Comparing the multiple correlation coefficients with the correlations in Table II.5 we see that, although there are instances of significant improvement, where values were already high there has tended to be little change, which in part is only natural since there is then little room for improvement. Specifically, for the previous choices of variables which should be associated with each PC, there has been no improvement in the correlations, to four decimal places.

It is worth comparing equations (II.4.1) and (II.2.2), which are at the heart of both Tables. In the notation of this Section, equation (II.2.2) for the correlation between the j-th PC and the 'truncated PC' based on the terms in the k variables of subset $\mathcal{G}$ can be written as:

$$r_t = \sqrt{\lambda_j} \cdot \frac{\mathbf{a}_j^{\mathcal{G}'} \mathbf{a}_j^{\mathcal{G}}}{\sqrt{\mathbf{a}_j^{\mathcal{G}'} \mathbf{S}_{\mathcal{G}} \mathbf{a}_j^{\mathcal{G}}}} \tag{II.4.2}$$

If we now consider the ratio of the multiple correlation and the correlation between the PC and the 'truncated PC', we obtain:

$$\frac{r_m}{r_t} = \sqrt{\frac{\mathbf{a}_j^{\mathcal{G}'} \mathbf{S}_{\mathcal{G}} \mathbf{a}_j^{\mathcal{G}}}{\mathbf{a}_j^{\mathcal{G}'} \mathbf{a}_j^{\mathcal{G}}} \cdot \frac{\mathbf{a}_j^{\mathcal{G}'} \mathbf{S}_{\mathcal{G}}^{-1} \mathbf{a}_j^{\mathcal{G}}}{\mathbf{a}_j^{\mathcal{G}'} \mathbf{a}_j^{\mathcal{G}}}} \tag{II.4.3}$$

This is the geometric mean of the values determined by the vector $\mathbf{a}_j^{\mathcal{G}}$ in the Rayleigh-Ritz ratios of $\mathbf{S}_{\mathcal{G}}$ and $\mathbf{S}_{\mathcal{G}}^{-1}$. It represents the factor by which we can improve on the correlation r_t if we use multiple regression, rather than the sum of terms with variables in $\mathcal{G}$, to approximate the PC.

Even if the gains in correlation are minimal, there are advantages in using the multiple correlation coefficient, rather than r_t, in choosing which subset of the k variables should be used to approximate a given PC. Among these advantages are:

i). the use of multiple correlations in choosing a subset of variables to provide a 'good' approximation for another variable is well-established (see, for example, Section 6.7.1 of Mardia, Kent and Bibby (1981) [42]);

ii). the undesirable feature of having additional variables decreasing the value of r_t cannot occur with multiple correlations. In regression, as is well known (see, for example, Draper and Smith (1981) [11, (pgs. 206/7)]), adding any new variables to the set of regressor variables can only increase the value of the multiple correlation;

iii). if the criterion to accept a given subset of k variables is that the correlation of the approximating vector it generates exceeds some threshold, using r_t may lead to rejecting a potentially adequate subset only because we are using a sub-optimal approximating vector;

iv). formula (II.4.1) for the multiple correlation is no more involved than formula (II.4.2). Therefore, even minimal gains are obtained at no extra cost.

But the overriding consideration that makes the use of multiple correlations more appropriate is that it is not always the case that projected PCs and 'truncated PCs' will be similarly correlated with the PCs themselves, as happens with Kendall's soil data. For other data sets, the gains in correlation from using the regression approach can be considerable.

Let us consider a second data set, this time with a larger number of variables. The data set is given in Table II.8 and has been reproduced from Lebart *et al* (1982) [40, (pg.283)]. It summarizes a study of yearly expenditures on foodstuffs, by French families. There are $p = 7$ variables, corresponding to groups of foodstuffs: bread, vegetables, fruit, red meat, poultry, milk and wine. The $n = 12$ 'individuals' are in reality the average expenditures of the sampled families for each of 12 categories – the combinations of each of three social groups (manual workers [M], non-manual workers

[NM], technical and managerial staff [S]) with each of four family sizes (parents with 2,3,4 or 5 children).

Categories		BREAD	VEG.	FRUIT	MEAT	POULTRY	MILK	WINE
Child.	Social	($\mathbf{y}_1$)	($\mathbf{y}_2$)	($\mathbf{y}_3$)	($\mathbf{y}_4$)	($\mathbf{y}_5$)	($\mathbf{y}_6$)	($\mathbf{y}_7$)
2	M	332	428	354	1437	526	247	427
	NM	293	559	388	1527	567	239	258
	S	372	767	562	1948	927	235	433
3	M	406	563	341	1507	544	324	407
	NM	386	608	396	1501	558	319	363
	S	438	843	689	2345	1148	243	341
4	M	534	660	367	1620	638	414	407
	NM	460	699	484	1856	762	400	416
	S	385	789	621	2366	1149	304	282
5	M	655	776	423	1848	759	495	486
	NM	584	995	548	2056	893	518	319
	S	515	1097	887	2630	1167	561	284

Table II.8: French families' foodstuffs expenditures (Lebart data set) [M-Manual workers; NM-Non-manual workers; S-technical and managerial staff]

Although Lebart *et al* choose to analyze the correlation matrix for this data set, we have a situation where a Covariance Matrix PCA is also appropriate. All variables are in the same units (French Francs) and separate re-scalings of each variable do not make much sense. The choice between Covariance and Correlation Matrix PCA is often depicted as a decision on whether the foodstuffs for which there is greater variability in consumption among the 12 groups will weigh more in the analysis, or whether all 7 food types will be given equal importance. We shall work with the covariance matrix, in what follows.

The covariance and correlation matrices for this data set are given in Table II.9.

Tables II.10 and II.11 give the loadings and corrected loadings, respectively, of each centered variable, for the (Covariance Matrix) PCs of this data set. Table II.10 also gives the eigenvalues and relative eigenvalues associated with each PC.

Variable	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$
$\mathbf{y}_1$	10 524	(0.593)	(0.196)	(0.321)	(0.248)	(0.856)	(0.304)
$\mathbf{y}_2$	11 021	32 807	(0.856)	(0.881)	(0.827)	(0.663)	(-0.357)
$\mathbf{y}_3$	3 180	24 514	24 984	(0.960)	(0.923)	(0.332)	(-0.486)
$\mathbf{y}_4$	12 488	60 468	57 464	143 567	(0.982)	(0.375)	(-0.437)
$\mathbf{y}_5$	6 079	35 781	34 955	88 884	57 090	(0.233)	(-0.400)
$\mathbf{y}_6$	9 843	13 463	5 888	15 916	6 240	12 576	(0.007)
$\mathbf{y}_7$	2 142	-4 437	-5 283	-11 386	-6 571	53	4 723

Table II.9: Covariance (and Correlation) Matrices for the foodstuffs expenditure data

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
$\mathbf{x}_1$	0.073	0.576	0.404	0.114	0.169	-0.674	0.068
$\mathbf{x}_2$	0.328	0.409	-0.292	0.608	-0.427	0.183	-0.235
$\mathbf{x}_3$	0.303	-0.100	-0.340	-0.397	-0.568	-0.432	0.341
$\mathbf{x}_4$	0.753	-0.108	0.068	-0.294	0.285	0.001	-0.499
$\mathbf{x}_5$	0.465	-0.244	0.381	0.330	0.065	0.208	0.650
$\mathbf{x}_6$	0.091	0.632	-0.225	-0.414	0.237	0.439	0.350
$\mathbf{x}_7$	-0.059	0.144	0.660	-0.307	-0.571	0.301	-0.174
Eigenvalues	251 928	24 215	5 733	2 108	1 916	310	60
Relative eigenvalues	0.8800	0.0846	0.0200	0.0074	0.0067	0.0011	0.0002

Table II.10: Eigenpairs for the PCs of the foodstuffs expenditures data

We shall focus on choosing subsets of the 7 variables with which to attempt interpretations of the first 3 PCs, which account for over 98% of the total variance.

For argument's sake, we shall begin by retaining, for each PC, those variables whose (conventional) loading's magnitude exceeds 0.30 (the precise threshold is not crucial for the discussion that follows). Then, we use the same criterion on the

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
x_1	0.024	0.424	0.300	0.061	0.103	-0.549	0.027
x_2	0.188	0.532	-0.383	0.569	-0.458	0.263	-0.165
x_3	0.151	-0.114	-0.390	-0.324	-0.532	-0.542	0.209
x_4	0.903	-0.294	0.187	-0.576	0.640	0.003	-0.734
x_5	0.352	-0.418	0.660	0.408	0.091	0.394	0.604
x_6	0.032	0.508	-0.183	-0.240	0.157	0.391	0.152
x_7	-0.013	0.071	0.329	-0.109	-0.232	0.164	-0.047

Table II.11: Corrected loadings for the PCs of the foodstuffs expenditures data

corrected loadings. The choices that result are given in Table II.12 (the variables being listed by decreasing magnitude of their loadings and with 'vs.' denoting a contrast).

PC	Loadings	Corrected loadings
1	x_4,x_5,x_2,x_3	x_4,x_5
2	x_6,x_1,x_2	x_2,x_6,x_1 vs. x_5
3	x_7,x_1,x_5 vs. x_3	x_5,x_7,x_1 vs. x_3,x_2

Table II.12: Variables retained to interpret each PC (with a loading threshold of 0.30) of the Lebart data set

To judge how well each of these choices performs, we now compute the multiple correlations and the r_t coefficients for the first three PCs and the above sets of variables, as well as for all single-variable subsets of $\{x_1, ..., x_7\}$. The results are summarized in Tables II.13 and II.14.

The first PC is very well approximated by the subsets given by either the loadings or the corrected loadings, with little difference between taking the 'truncated' or the projected PCs. This appears to be largely due to the fact that both subsets include

Variable	PC1	PC2	PC3
$\mathbf{x}_1$	0.3564	0.8735	0.2982
$\mathbf{x}_2$	0.9092	0.3516	0.1219
$\mathbf{x}_3$	0.9608	0.0986	0.1630
$\mathbf{x}_4$	0.9977	0.0444	0.0136
$\mathbf{x}_5$	0.9774	0.1588	0.1207
$\mathbf{x}_6$	0.4077	0.8764	0.1522
$\mathbf{x}_7$	0.4292	0.3269	0.7270

Table II.13: Absolute correlations of PCs with single variables for Lebart's data set

PC	Variables in subset	r_t	r_m
1	$\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5$	0.9996	1.0000
1	$\mathbf{x}_4,\mathbf{x}_5$	0.9957	0.9977
2	$\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_6$	0.7656	0.9641
2	$\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_5,\mathbf{x}_6$	0.9251	0.9986
3	$\mathbf{x}_1,\mathbf{x}_3,\mathbf{x}_5,\mathbf{x}_7$	0.7725	0.9790
3	$\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_5,\mathbf{x}_7$	0.9300	0.9968

Table II.14: r_t and r_m coefficients for the first three PCs and some subsets of variables in the Lebart data set

variable $\mathbf{x}_4$ which, by itself, is already an excellent 'approximate PC', forming an angle of under 4° with the PC in $\mathbb{R}^n$. The performance of $\mathbf{x}_4$ is not surprising given its large loading, combined with its large variance. The addition of $\mathbf{x}_5$ alone, whose corrected loading is 0.352, actually worsens the performance of the 'truncated PC' and does not improve, to four decimal places, that of the projection. Overall, the criteria used above have produced good results, although at a higher cost (more variables used) than is actually necessary.

But the results for the second and third PCs are much more interesting. Reasonably good approximations of both PCs can be obtained using the subsets selected by either kind of loadings, since the lowest of the four multiple correlations (that of $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_6\}$ with the second PC) is 0.9641, which corresponds to an angle of $15°24'$ in $\mathbb{R}^n$. But in all four cases, the approximations provided by the 'truncated PCs' are worse and even – in particular for the subsets which result from the conventional loadings – significantly worse than those given by regression.

The implications of this are considerable. Not only might we reject some potentially adequate subsets of variables, on the basis of the poor performance of their 'truncated PCs', but a more serious problem arises: that of *differing interpretations* for a given PC, based on which method of approximation is chosen.

To illustrate this problem, the second PC and the subset $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_6\}$ selected by the conventional loadings approach will be looked at in detail. The approximation to the PC provided by taking only those terms of $\mathbf{X}\mathbf{a}_2$ in the variables $\mathbf{x}_1$,$\mathbf{x}_2$ and $\mathbf{x}_6$ is:

$$\mathbf{v}_1 = 0.576\mathbf{x}_1 + 0.409\mathbf{x}_2 + 0.632\mathbf{x}_6$$

Thus, this second PC might be viewed as a weighted average of bread, vegetables and milk. But a linear multiple regression of the PC on the same set of variables would produce the approximate vector which, after re-scaling the coefficients to a comparable size (the same sum of squares as in $\mathbf{v}_1$) is given by:

$$\mathbf{v}_2 = 0.546\mathbf{x}_1 - 0.266\mathbf{x}_2 + 0.727\mathbf{x}_6$$

The latter approximation will lead us to view the PC as a *contrast* between bread and milk, on the one hand, and vegetables, on the other.

There should be little doubt that, between $\mathbf{v}_1$ and $\mathbf{v}_2$, it is the latter that should be chosen to approximate the PC. Its correlation with the PC is significantly better (0.9641 as compared with 0.7656) and so it is a more faithful reflection (scaling factors aside) of that PC. But a conventional approach to the approximation of PCs using their 'truncations' would lead us to take the weighted average view.

As with the other problems discussed in the previous Sections, the use of corrected loadings tends to improve the situation. Again, for the second PC of the Lebart data set, the 'truncated PC' is now:

$$\mathbf{v}_3 = 0.576\mathbf{x}_1 + 0.409\mathbf{x}_2 - 0.244\mathbf{x}_5 + 0.632\mathbf{x}_6$$

whereas the comparable projected PC is:

$$\mathbf{v}_4 = 0.745\mathbf{x}_1 + 0.312\mathbf{x}_2 - 0.408\mathbf{x}_5 + 0.373\mathbf{x}_6$$

Although the relative importance of each variable changes in a non-negligible way, there is at least qualitative agreement between $\mathbf{v}_3$ and $\mathbf{v}_4$. This is not surprising, given that the correlations of each of those vectors with the PC are both fairly high (0.9251 for $\mathbf{v}_3$ and 0.9986 for $\mathbf{v}_4$), effectively ruling out major differences between the vectors.

An identical situation exists in the case of the third PC. The 'truncated PC' using only $\mathbf{x}_1$,$\mathbf{x}_3$,$\mathbf{x}_5$ and $\mathbf{x}_7$ can be seen from Table II.10 to contrast $\mathbf{x}_3$ (fruit) with the other three variables. But the corresponding projected PC, whose correlation with the PC grows from 0.7725 to 0.9790, is:

$$\mathbf{v} = -0.057\mathbf{x}_1 - 0.481\mathbf{x}_3 + 0.408\mathbf{x}_5 + 0.675\mathbf{x}_7$$

Thus, bread ($\mathbf{x}_1$) has 'changed sides' in the contrast, although with a relatively small coefficient. This small coefficient, coupled with $\mathbf{x}_1$'s unimpressive variance suggests that maybe the third, fifth and seventh variables are sufficient for a good approximation. In fact, the multiple correlation of the third PC with $\mathbf{x}_3$, $\mathbf{x}_5$ and $\mathbf{x}_7$ turns out to be only marginally smaller: 0.9752. Incidentally, neither loadings nor corrected loadings could have produced this subset, since it leaves out the variable with the second largest magnitude loading ($\mathbf{x}_1$) and the third largest magnitude corrected loading ($\mathbf{x}_2$).

The general conclusion to be retained is that the use of the multiple correlation coefficient in assessing the performance of a subset of variables with regards to a given PC is not just a question of minor improvements. In the same way that loadings can be misleading when choosing variables which make significant contributions to a PC, so too *'truncating' a PC is a potentially misleading means of assessing how a given subset of variables combines to best approximate a PC.* The problem is all the more serious if a fairly high correlation between the PC and its 'truncation' is not achieved with a particular choice of variables. *Interpretation of PCs based on such 'truncated PCs' may then be flawed.*

II.5 Which subset of variables?

The discussion in the previous Sections concentrated on considering whether a given subset of variables could provide a good approximation to a given PC and, if so, how this approximation should be computed.

But two further aspects must be considered: (i) how can we be confident that all potentially 'good' subsets of the p variables are being considered, now that loadings are no longer entirely reliable? ; (ii) can an unambiguous choice of 'best' subset of the variables always be made?

When the number, p , of variables is very small, it is feasible to calculate the multiple correlations of each PC with *all* $2^p - 2$ proper subsets of the p variables. A complete table of such multiple correlations is of great assistance in considering the problem of which subset to associate with each PC, as was illustrated in the discussion of Kendall's soil data set. But for even moderate values of p , such a complete set of values of r_m rapidly becomes a heavy computational burden.

Working in a regression context enables us to rely on a vast literature and widely available algorithms and software in seeking shortcuts for this search, which eliminate unlikely subsets of variables. Draper and Smith (1981) [11] devote a whole Chapter (Chapter 6) to a detailed discussion of the problem and various algorithms that have been suggested to tackle it. Part of this discussion involves inferential considerations which we have chosen to ignore.

Some of these algorithms involve beginning the search with a given subset of the variables (possibly all the variables, or a single one), which is then improved by including and deleting individual variables until some criterion for acceptability is met. For our specific problem, it is only fair to point out that a sensible starting point for such a search can be provided by the corrected loadings for any given PC. Initially selecting the subset of variables whose corrected loadings exceed some threshold (like 0.20 or 0.30 in absolute values) should provide further economies of time and resources which cannot be obtained in a general regression problem.

The criterion for acceptability of a given subset of variables often involves the multiple correlation coefficient. Mardia, Kent and Bibby (1979) [42] suggest that for a subset of variables to be considered an adequate set of regressor variables, its squared multiple correlation with the 'dependent' variable must be (pg.175) "at least 90 or 95% of the squared multiple correlation involving all [p] variables". In our case, the latter value is always 1, and so we are talking about values of r_m exceeding 0.95 (angles in $\mathbb{R}^n$ of less than 18°30′) or 0.975 (angles of less than 13°). Draper and Smith (1981) [11], on the other hand, appear to take a more flexible approach and focus on a comparative study of values of r_m for several alternative subsets. This is, for example, the spirit of their case study in Section 6.1, where they also address the issue of several alternative subsets with similar values for the multiple correlation, stating [11, (pg.297)]:

> The examination of all possible regressions does not provide a clear-cut answer to the problem. Other information, such as knowledge of the characteristics of the product studied and the physical role of the X-variables, must as always be added to enable a decision to be made.

To illustrate the second problem mentioned above (that of possible ambiguities), we again turn our attention to Table II.13 where the correlations between PCs and variables in the Lebart data set are given. Although a single variable ($\mathbf{x}_4$) is clearly sufficient to approximate the first PC, as was seen before, it is nonetheless useful to know that other variables, such as $\mathbf{x}_5$,$\mathbf{x}_3$ and even $\mathbf{x}_2$ are strongly correlated with the first PC. Although including these variables can only marginally improve the quality of the approximation provided by $\mathbf{x}_4$, it can be argued that the first PC can also be associated with the subset of four variables $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$, rather than with the fourth variable alone (see Table II.14). In $\mathbb{R}^n$, these four variables are represented by vectors which, although of differing lengths, form relatively small angles between themselves. The PC is piercing this 'angle-wise cluster' of vectors, although in a way that almost makes it overlap $\mathbf{x}_4$. Thus, both views of the PC are useful for a researcher.

The situation is further complicated by the fact that the three variables $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5\}$ can, on their own, provide an approximation to the PC which is just marginally worse than that provided by the fourth variable alone. In fact, their multiple correlation with the PC is 0.9970. Thus, an alternative interpretation can be found for the PC which does not even involve $\mathbf{x}_4$.

Another example, involving a slightly different problem, results from considering the second PC for the same data set. There are four combinations of two variables which produce values of r_m greater than 0.95: $\{\mathbf{x}_1, \mathbf{x}_5\}$ (0.9556) ; $\{\mathbf{x}_3, \mathbf{x}_6\}$ (0.9689) ; $\{\mathbf{x}_4, \mathbf{x}_6\}$ (0.9642) ; $\{\mathbf{x}_5, \mathbf{x}_6\}$ (0.9525). Choosing one of these subsets on the basis of marginal differences in the multiple correlations may then be unwise. Any possible interpretation will have some degree of ambiguity and might not be robust, in the sense that measurement errors (or sampling variability, in an inferential context) might be determining the choice of the subset. No single-variable subset seems adequate either, both because the highest correlations are not altogether satisfactory (0.8764 corresponds to an angle of almost 29° in $\mathbb{R}^n$) and because the ambiguity would not be solved, since the top two correlations (for $\mathbf{x}_1$ and $\mathbf{x}_6$) are very similar.

As with the previous example, the situation is further complicated by the fact that all four 5-variable complements of the above subsets have larger multiple correlations with the PC (the lowest of these values is 0.9790) and can therefore provide better approximations.

It is becoming apparent that it is not sufficient to consider how good a fit can be provided by some subset of variables. It is also important to assess how well the PC can be approximated by the remaining variables, which do not belong to the subset.

If the variables in the complementary subset can provide an approximation which is at least 'almost as good', then it might be advisable to look for an alternative 'explanatory' subset.

Let us take a fresh look at the two examples considered above.

We had already seen that $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5\}$ provided a similar approximation to the first PC of the French families data set, as did $\mathbf{x}_4$ alone. Our new considerations suggest that the 'explanatory' subset should not consist only of $\mathbf{x}_4$. As was already seen the multiple correlation of a given PC with any subset of variables must be greater than its correlation with any single variable in that subset. This fact suggests that at least the four variables $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ should be included in the 'explanatory' subset, for otherwise the multiple correlation involving the complementary subsets would exceed 0.90. Since the multiple correlation of the PC with the three remaining variables ($\{\mathbf{x}_1, \mathbf{x}_6, \mathbf{x}_7\}$) is considerably lower (0.6755), the PC cannot be well approximated without the former four variables. A satisfactory subset appears to have been found.

As for the second PC for the same data set, we had concluded that none of the four 2-variable subsets which generated multiple correlations greater than 0.95 appeared to be satisfactory. The appropriate way to overcome the problem appeared to be by including more variables in the 'explanatory' subsets. One possible choice is to select variables whose loadings exceeded 0.3 in magnitude: $\mathbf{x}_1$,$\mathbf{x}_2$,$\mathbf{x}_6$. Their multiple correlation with the PC is reasonably high (0.9641). The remaining variables can only generate a multiple correlation of 0.8187, which corresponds to an angle of about 35° in $\mathbb{R}^n$. The variables whose *corrected* loadings exceed 0.3 ($\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_5, \mathbf{x}_6\}$) provide an even more convincing choice. Their multiple correlation with the PC is 0.9986 and that of the remaining variables a very modest 0.3626. Selecting between these two alternative subsets will therefore depend on what values of r_m are considered adequate, both for the 'explanatory' subset and for its complement. Should the choice go for the subset generated by the corrected loadings, an additional problem will emerge. An alternative four-variable subset ($\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_6\}$) can provide an even better fit ($r_m = 0.9995$) with only a slightly worse performance of its complement ($r_m = 0.4052$).

Thus, two problems remain. First, no precise threshold has been set for what values of r_m are considered adequate, both for the 'explanatory' subset and for its complement. Second, it cannot be guaranteed that all ambiguities of choice among alternative subsets can be eliminated.

One noteworthy point concerning any surviving ambiguities in the choice of subsets associated with a given PC is that, in itself, this provides some information concerning the PC and its interpretability. It is possible that, for a specific data set, one or other of such similarly appropriate subsets may, as Draper and Smith suggest, prove to be a 'natural' and easily interpretable choice. But this may not be the case. As a last resort, one way of resolving such ambiguities exists in our case: to take the whole set of p variables with which the PC is defined.

However, Draper and Smith's central conclusion for the discussion of this problem also seems to be the most adequate for our case [11, (pg.342)]:

> No technique will work well in all circumstances, no matter how good it may look on a particular example. No technique is always better than all the others. (...) The techniques discussed in this chapter can be useful tools. However, none of them can compensate for common sense and experience.

The last sentence in this quote is also the most appropriate conclusion for the whole of this Section.

II.6 Conclusions

Throughout this chapter we have attempted to show that more rigour and quantification are called for when PCs are associated with subsets of variables, as a prelude to interpretations.

The traditional practice of selecting such subsets from their corresponding loadings and of 'truncating' PCs as a means for obtaining approximate PCs may be seriously unreliable. Multiple correlation coefficients and the regression of PCs on subsets of the variables are more appropriate alternatives.

Regressing PCs on variables – which is a curious inversion of the relation which originally motivated Hotelling's approach to PCA – raises numerous technical difficulties, in particular when the number of variables is not very small. In many cases, previous work on multiple linear regression provides useful background material to tackle such difficulties. In other instances, it is the precise nature of our regression problem which provides useful tips. An example of the latter case is the use of corrected loadings to obtain a preliminary association of variables with a PC.

It may not be possible to determine a single, unambiguously 'best', subset of k $(k < p\ )$ variables which should be associated with a given PC. Furthermore, many details (such as precise values for thresholds) have been purposely left in rather vague terms. Much of this discussion requires flexibility, knowledge of the specific data set involved and experience.

But it is hoped that some of the points that have been made here will contribute towards setting more solid foundations for the second stage of the interpretation process.

On a more general level, the discussion in this Chapter stresses the importance of clearly distinguishing between PCs and their vectors of loadings. All too often the two sets of vectors are confused – to the point of using the term 'principal components' in both cases. Although there is a very strong relationship between both sets of vectors, obscuring their distinct nature can be misleading.

Chapter III

PCA and the Geometry of Matrix Spaces

In previous Chapters, an $n \times p$ data matrix has been viewed as a collection of points or vectors in either $\mathbb{R}^n$ or $\mathbb{R}^p$. But it is also possible to view it as a single element in the space of all $n \times p$ matrices. Covariance and correlation matrices can also be usefully viewed as points in the space of all $p \times p$ matrices, and the matrices of the form $\frac{1}{n}\mathbf{X}\mathbf{X}'$, whose eigenvectors are the PCs of the column-centered data matrix $\mathbf{X}$, as points in the space of all $n \times n$ matrices.

In this Chapter, we shall consider such matrix spaces. Many concepts of a geometric nature in these spaces are relevant to the study of PCA.

III.1 The spaces $\mathbb{M}_{n\times p}$

The set of all $n\times p$ matrices, with the usual operations of addition and scalar multiplication forms a **linear space** (see, for example, Horn and Johnson (1985) [25, (pg.5)]) which we shall denote by $\mathbb{M}_{n\times p}$. For our purposes, it can be viewed as the space of all possible raw data matrices. The set of $n\times p$ matrices $\{\mathbf{E}_{ij}\}_{i=1}^{n}{}_{j=1}^{p}$ defined as having a single non-zero element, of value 1, in the (i,j)-th position, forms a basis for the space $\mathbb{M}_{n\times p}$. Thus, this linear space of matrices is of dimension np. As such, it is isomorphic to any other np-dimensional linear space (see, for example, Shilov (1961) [61, (pg.48)]) like, for example, $\mathbb{R}^{np}$.

An isomorphism between $\mathbb{M}_{n\times p}$ and $\mathbb{R}^{np}$ is defined by the *vec* operator (see Searle (1982) [59, (pg.332)]). This operator 'vectorizes' an $n\times p$ matrix by stacking its columns on top of each other so that the (i,j)-th element of the matrix becomes the $[n(j-1)+i]$-th element of the np-dimensional vector.

III.1.1 An inner product in $\mathbb{M}_{n\times p}$

Geometric concepts in a linear space depend, to a large extent, on the existence of an **inner product** in that space. The most common inner product in $\mathbb{M}_{n\times p}$ is (see, for example, Ramsay *et al* (1984) [52]):

$$< \mathbf{A},\mathbf{B} >= \mathrm{tr}(\mathbf{A}'\mathbf{B}) \ , \qquad \forall \mathbf{A},\mathbf{B} \in \mathbb{M}_{n\times p} \tag{III.1.1}$$

The bilinearity and symmetry of the inner product are guaranteed by the properties of traces. Positiveness for $\mathbf{A} = \mathbf{B} \neq \mathbf{0}$ is guaranteed because, as was seen is Section I.3, $\mathbf{A}'\mathbf{A}$ is always a positive semi-definite matrix. Hence, its trace is the sum of its real and non-negative eigenvalues, not all of which can be zero if $\mathbf{A} \neq \mathbf{0}$.

It is useful to re-write this inner product in terms of the elements of matrices $\mathbf{A}$ and $\mathbf{B}$. We have:

$$< \mathbf{A}, \mathbf{B} >= \sum_{i=1}^{n} \sum_{j=1}^{p} a_{ij} b_{ij} \ , \qquad \forall \mathbf{A}, \mathbf{B} \in \mathbb{M}_{n \times p} \tag{III.1.2}$$

where a_{ij} and b_{ij} are the generic elements of matrices $\mathbf{A}$ and $\mathbf{B}$, respectively. It can be seen that this is also the sum of all elements in the Hadamard (entry-wise) product of $\mathbf{A}$ and $\mathbf{B}$.

Yet another possible expression for the inner product can be obtained by considering the Singular Value Decompositions of $\mathbf{A}$ and $\mathbf{B}$ (with notation as in (I.3.5)):

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i^A \boldsymbol{\alpha}_i \mathbf{a}_i{}'$$

$$\mathbf{B} = \sum_{j=1}^{q} \sigma_j^B \boldsymbol{\beta}_j \mathbf{b}_j{}'$$

We then have:

$$< \mathbf{A}, \mathbf{B} >= \mathrm{tr} \left(\sum_{i=1}^{r} \sum_{j=1}^{q} \sigma_i^A \sigma_j^B \mathbf{a}_i (\boldsymbol{\alpha}_i' \boldsymbol{\beta}_j) \mathbf{b}_j{}' \right)$$

$$\Longleftrightarrow \quad < \mathbf{A}, \mathbf{B} >= \sum_{i=1}^{r} \sum_{j=1}^{q} \sigma_i^A \sigma_j^B (\boldsymbol{\alpha}_i' \boldsymbol{\beta}_j)(\mathbf{b}_j{}' \mathbf{a}_i) \tag{III.1.3}$$

from the properties of the trace. Thus, the inner product of two matrices is a weighted sum of products of the cosines of angles between each possible pair of left singular vectors (one from each matrix) and the cosines of the angles between each corresponding pair of right singular vectors. The weights are the product of the singular values associated with each pair.

It is not hard to see from equation (III.1.2) that if $vec(\mathbf{A})$ and $vec(\mathbf{B})$ are the vectors which result from vectorizing matrices $\mathbf{A}$ and $\mathbf{B}$, then:

$$< \mathbf{A}, \mathbf{B} >_{\mathbb{M}_{n\times p}} = < vec(\mathbf{A}), vec(\mathbf{B}) >_{\mathbb{R}^{np}}$$

Thus, $\mathbb{M}_{n\times p}$ with inner product (III.1.1) and $\mathbb{R}^{np}$ with the usual Euclidean inner product are, *as inner product spaces*, also isomorphic. The implication is that geometric considerations based on inner products relative to the np-dimensional Euclidean space $\mathbb{R}^{np}$ are also valid for the matrix space $\mathbb{M}_{n\times p}$ (see also Shilov (1961) [61, (pg.144)] or Gelfand (1961) [16, (pg.32)]).

III.1.2 A norm in $\mathbb{M}_{n\times p}$

The concept of *length* in $\mathbb{M}_{n\times p}$ can be derived by using the inner product to define a **norm**, in the standard way. The norm defined by the inner product (III.1.1) is:

$$||\mathbf{A}|| = \sqrt{< \mathbf{A}, \mathbf{A} >} = \sqrt{\mathrm{tr}(\mathbf{A}'\mathbf{A})} \;, \quad \forall \mathbf{A} \in \mathbb{M}_{n\times p} \qquad \text{(III.1.4)}$$

This norm can be re-written in terms of the eigenvalues of $\mathbf{A}'\mathbf{A}$, $\{\lambda_i(\mathbf{A}'\mathbf{A})\}_{i=1}^{p}$, and therefore in terms of $\mathbf{A}$'s singular values, $\{\sigma_i(\mathbf{A})\}_{i=1}^{r}$, (where r is the rank of $\mathbf{A}$):

$$||\mathbf{A}|| = \sqrt{\sum_{i=1}^{p} \lambda_i(\mathbf{A}'\mathbf{A})} = \sqrt{\sum_{i=1}^{r} \sigma_i^2(\mathbf{A})} \qquad \text{(III.1.5)}$$

Alternatively, it is possible to write the norm in terms of the elements of $\mathbf{A}$, using (III.1.2):

$$||\mathbf{A}|| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}^2} \qquad \text{(III.1.6)}$$

The norm (III.1.4) – which has just been seen to be the usual vector norm of $vec(\mathbf{A})$ – is known in Matrix Theory (in particular for spaces of square matrices) as the **Frobenius norm**. Alternatively, it is called the Schur or Hilbert-Schur norm (see Horn and Johnson (1985) [25, (pg.291)]).

III.1.3 Angles in $\mathbb{M}_{n\times p}$

In keeping with standard practice for linear spaces (see Section 50.2 in Shilov (1961) [61]) we can now use the Cauchy-Schwarz inequality to define the **angle** between two matrices in $\mathbb{M}_{n \times p}$:

$$\sphericalangle(\mathbf{A}, \mathbf{B}) = \arccos\left(\frac{<\mathbf{A}, \mathbf{B}>}{||\mathbf{A}|| \cdot ||\mathbf{B}||}\right) \quad , \quad \forall \mathbf{A}, \mathbf{B} \in \mathbb{M}_{n \times p} \qquad \text{(III.1.7)}$$

Naturally, the cosine of this angle, which is also the cosine of the angle in $\mathbb{R}^{np}$ between $vec(\mathbf{A})$ and $vec(\mathbf{B})$, is:

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{<\mathbf{A}, \mathbf{B}>}{||\mathbf{A}|| \cdot ||\mathbf{B}||} \qquad \text{(III.1.8)}$$

This cosine is sometimes called a **matrix correlation** (see Ramsay *et al* (1984) [52]). As with correlations between variables, its value is unaffected if $\mathbf{A}$ and/or $\mathbf{B}$ are multiplied by positive scalars. Negative scalars will change signs (there being no change if both $\mathbf{A}$ and $\mathbf{B}$ are multiplied by negative scalars). From the general properties of the Cauchy-Schwarz inequality (see, for example, Shilov (1961) [61, (pg.139)]) we know that $\cos(\mathbf{A},\mathbf{B}) = \pm 1$ if and only if $\mathbf{B} = \alpha\mathbf{A}$ $(\alpha \neq 0)$, the sign of α determining the sign of the correlation.

The cosine of the angle between two matrices can also be written in terms of their SVDs. Using (III.1.3) and (III.1.5) we have:

$$\cos(\mathbf{A},\mathbf{B}) = \sum_{i=1}^{r}\sum_{j=1}^{q}\sqrt{\pi_i^A\pi_j^B}(\mathbf{a}_i{}'\mathbf{b}_j)(\boldsymbol{\alpha}_i'\boldsymbol{\beta}_j) \tag{III.1.9}$$

where $r, q, \mathbf{a}_i, \mathbf{b}_j, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_j$ are as before and π_i^A, π_j^B are the (see I.4.12) relative eigenvalues of the matrices $\mathbf{A}'\mathbf{A}$ and $\mathbf{B}'\mathbf{B}$, respectively. Thus, the cosine between two matrices is a weighted sum similar to the inner product, but with different weights. The new weights for each pair of vectors are the geometric means of the relative eigenvalues of $\mathbf{A}'\mathbf{A}$ and $\mathbf{B}'\mathbf{B}$ associated with that pair. Ramsay *et al* (1984) [52] also write the correlation (III.1.8) in terms of the SVDs of $\mathbf{A}$ and $\mathbf{B}$, but using the 'matrix version' of the SVD (equation (I.3.6)) which does not highlight the precise way in which singular vectors and values combine to give $\cos(\mathbf{A},\mathbf{B})$. As we shall see in the next Section, expression (III.1.9) will actually play a crucial role in the context of PCA.

III.1.4 Orthogonality

Having defined an inner product in $\mathbb{M}_{n\times p}$, it is possible to say that two matrices in $\mathbb{M}_{n\times p}$ are **mutually orthogonal** (not to be confused with the concept of an orthogonal matrix) when their inner product is zero. Concepts such as the orthogonal complement of a subspace of $\mathbb{M}_{n\times p}$ and the orthogonal projection of an element of $\mathbb{M}_{n\times p}$ onto some subspace thus become possible (see Shilov (1961) [61, (Chapter 8)]).

For example, the set of all possible column-centered data matrices can be characterized as an orthogonal complement of a subspace of $\mathbb{M}_{n\times p}$. In fact, any column-centered data matrix is orthogonal to all matrices of the form $\mathbf{C} = \sum_{i=1}^{p} a_i\mathbf{C}_i$, where $\{a_i\}_{i=1}^{p}$ are scalars and $\{\mathbf{C}_i\}_{i=1}^{p}$ are some set of $n\times p$ matrices whose only non-zero elements are all equal and form the i-th column of $\mathbf{C}_i$. Thus, *the column-centered data matrices form a subspace* $\mathbb{C}_{\mathrm{nxp}}$ *of* $\mathbb{M}_{\mathrm{nxp}}$ *, of dimension* $np - p = p(n-1)$.

As for any pair of orthogonal complements in an inner product space (see Halmos (1958) [22, (pg.129)]) this means that any raw data matrix $\mathbf{Y} \in \mathbb{M}_{n\times p}$ can be uniquely decomposed into a sum of a matrix of the form $\mathbf{C}$ above and a column-centered data matrix. It can easily be verified that if the matrices $\{\mathbf{C}_i\}_{i=1}^{p}$ above have all non-zero elements equal to 1, then the i-th scalar a_i is the mean of $\mathbf{Y}$'s i-th column. Thus, column-centering a data matrix corresponds to orthogonally projecting onto the subspace $\mathbb{C}_{n\times p}$.

III.1.5 Spaces of square matrices

A particular case of the matrix spaces $\mathbb{M}_{n\times p}$ occurs when we consider square matrices. The space of all $m\times m$ matrices will be denoted merely as $\mathbb{M}_m$.

Covariance and correlation matrices; the matrices of the form $\frac{1}{n}\mathbf{X}\mathbf{X}'$ whose eigenvectors are the PCs of the column-centered data matrix $\mathbf{X}$; the diagonal matrices of eigenvalues which appear in the spectral decomposition of covariance and correlation matrices; the orthogonal matrices of eigenvectors which also appear in the same spectral decompositions; all of these are examples of matrices belonging to a space of the type $\mathbb{M}_m$.

With the exception of the matrices of eigenvectors, the matrices in the examples above all share a further property: they are symmetric matrices. The $m\times m$ symmetric matrices form a subspace of $\mathbb{M}_m$, since they are closed under addition and scalar multiplication. This subspace will be denoted by $\mathbb{S}_m$. It is a linear space of dimension $m(m+1)/2$, which is the number of arbitrary elements in a symmetric matrix. The space $\mathbb{S}_m$ is therefore isomorphic to $\mathbb{R}^{m(m+1)/2}$.

An isomorphism between $\mathbb{S}_m$ and $\mathbb{R}^{m(m+1)/2}$ is given by the *vech* operator (Searle (1982) [59, (pg.332)]). This operator '*ve*ctorizes *h*alf' of a given $m\times m$ symmetric matrix by stacking the columns of its lower triangular portion on top of each other. Thus, the (i,j)-th element $(i \leq j)$ of an $m\times m$ symmetric matrix becomes the $\left\{\frac{m(m+1)}{2} - \frac{(m-j+1)(m-j+2)}{2} + i\right\}$-th element of an $m(m+1)/2$-dimensional vector.

However, for this isomorphism to extend to the inner product spaces, the usual inner product in $\mathbb{R}^{m(m+1)/2}$ has to be replaced by a weighted version giving weight 2 to elements of the vector associated with $j \neq i$.

The inner product, norm, angles and other concepts defined for the general matrix spaces $\mathbb{M}_{n\times p}$ will naturally apply to $\mathbb{M}_m$ and $\mathbb{S}_m$, as well. But some geometric concepts – which could also be defined in the general case – become particularly interesting in the case of spaces of square matrices. These will be considered in the next Subsections.

III.1.6 Flats and hyperplanes

The usual concept of 'displaced subspace' – lines and planes not containing the origin in $\mathbb{R}^3$ – can be generalized to linear spaces of any dimension. Using the terminology of Lay (1982) [39], for any given element $\alpha \in \mathcal{L}$, where $\mathcal{L}$ is an m-dimensional linear space, and for any r-dimensional subspace $\mathcal{S} \subset \mathcal{L}$, the set:

$$\mathcal{F} = \{\alpha + s : s \in \mathcal{S}\}$$

is called an **r-dimensional flat**. In the familiar case of $\mathbb{R}^3$, a point is a zero-dimensional flat, a line is a 1-dimensional flat and a plane is a 2-dimensional flat. Some authors (see Shilov (1961) [61, (Section 17)]) use the term *hyperplane* for the concept that has just been defined, but we shall restrict the usage of **hyperplane** to an (m -1)-dimensional flat, in keeping with Lay.

The general equation of a hyperplane can be expressed in terms of the inner product in the linear space $\mathcal{L}$. The set of elements $x \in \mathcal{L}$ such that:

$$< x, y >= \gamma \tag{III.1.10}$$

for a given scalar $\gamma \in \mathbb{R}$ and element $y \in \mathcal{L}$ is a hyperplane [39, (pg.31)]. The element $y \in \mathcal{L}$ is called the **normal** to the hyperplane. Two hyperplanes with the same normal are called **parallel hyperplanes.**

In $\mathbb{M}_m$, *all matrices with the same trace,* k , *define a hyperplane* whose normal is the identity matrix. Its equation is given by:

$$< \mathbf{A}, \mathbf{I}_m >= k \tag{III.1.11}$$

We shall refer to this hyperplane as the **trace k hyperplane.**

Thus, $\mathbb{M}_m$ (hence $\mathbb{S}_m$) is composed of 'layers' of parallel hyperplanes whose normal is $\mathbf{I}_m$, with all matrices in each layer sharing the same trace.

Any covariance or correlation matrix is in the same such hyperplane as the diagonal matrix of its eigenvalues, given by the spectral decomposition (I.3.2).

III.1.7 Hyperspheres

Generalizing from the usual definition in $\mathbb{R}^m$ (see, for example, Kendall (1961) [37, (pg.11)]) a hypersphere in $\mathbb{M}_m$ can be defined as the set of $m \times m$ matrices whose norm equals some fixed constant. The **norm-k hypersphere** is therefore given by the matrices $\mathbf{A} \in \mathbb{M}_m$ such that:

$$||\mathbf{A}|| = k \tag{III.1.12}$$

All orthogonal m×m *matrices* – as are the matrices of eigenvectors in a spectral decomposition (I.3.2) – *are on the norm-*$\sqrt{\mathrm{m}}$ *hypersphere in* $\mathbb{M}_{\mathrm{m}}$.

Also, *any covariance or correlation matrix is on the same hypersphere as its diagonal matrix of eigenvalues.*

III.2 Similarity indices for PCA

In this Section, we shall see how it is possible to obtain indicators of the degree to which two PCAs – or certain aspects of the PCAs – coincide, using some of the above geometric concepts.

III.2.1 A global similarity index

In Section I.4 we saw how for a column-centered data matrix, multiplied by the scalar $\frac{1}{\sqrt{n}}$, left singular vectors were the unit-norm PCs (Principal Axes in $\mathbb{R}^n$) of the original data; singular values were the 'natural' standard deviations of the PCs; and right singular vectors were the Principal Axes in $\mathbb{R}^p$.

The implication is that the *cosine between two* n×p *column-centered data matrices,* $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{C}_{\mathrm{nxp}}$ (the scalar $\frac{1}{\sqrt{n}}$ is not needed) *will measure the degree of similarity between the PCAs of* $\mathbf{X}_1$ *and* $\mathbf{X}_2$. In fact, from (III.1.9) we see that this cosine is *a weighted sum of the products of cosines of the angles between each possible pair of PCs (one from each data set) with cosines of the angles between the corresponding pair of PAs in $\mathbb{R}^p$. The weights are the geometric means of the proportion of total variance accounted for by each PC in the pair.*

Thus, the cosine between any two matrices in subspace$\mathbb{C}_{n\times p}$ is defined in terms of meaningful quantities of a PCA in both $\mathbb{R}^n$ and $\mathbb{R}^p$. It provides a global indicator of the degree of overall similarity of both sets of PAs in each space, taking into account their relative importance in each case. We shall call it the **global similarity index** of the PCAs of $\mathbf{X}_1$ and $\mathbf{X}_2$ and denote it by $s_g(\mathbf{X}_1, \mathbf{X}_2)$. It has the definite advantage of *not requiring that the PCAs be carried out.* We have:

$$s_g(\mathbf{X}_1, \mathbf{X}_2) = \cos(\mathbf{X}_1, \mathbf{X}_2) = \sum_{i=1}^{r} \sum_{j=1}^{q} \sqrt{\pi_i^{X_1} \pi_j^{X_2}} \cos(\mathbf{a}_i{}^{[1]}, \mathbf{a}_j{}^{[2]}) \cos(\boldsymbol{\alpha}_i^{[1]}, \boldsymbol{\alpha}_j^{[2]}) \qquad \text{(III.2.1)}$$

with notation similar to that used in (III.1.9).

A word of caution is advisable in any application of this index. Strictly speaking, the two data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ will only be comparable if the n individuals *and* the p variables in one data set are the same as, or directly related to, their corresponding counterparts in the other data set. An example of such data sets might be measurements of the same p variables on a given group of patients before and after some clinical trial. Another example are monthly averages of some meteorological variables, in a given location, for each of the $n = 12$ months in two different years. Examples of situations where we do *not* have such correspondences arise in the case of two independent samples of p given variables. In such cases, the comparisons involving $\mathbb{R}^n$ will not be very appropriate (see also Subsections III.2.2 and III.2.3). The index can also be a useful tool for comparing the degree of discrepancy between the PCAs of a single data set expressed in two different systems of units (say SI and Metric system units) or between the Covariance and Correlation Matrix PCAs of a given data set. This and other uses of the global similarity index will be considered in Chapter IV.

It should be stressed that $\mathbf{X}_1, \mathbf{X}_2$ or both matrices can be taken to be the raw (non-column-centered) data matrices or just row-centered matrices or both row- and column-centered matrices. The rationale for formula (III.2.1) remains intact in these comparisons involving non-centered, row-centered or doubly-centered PCs.

In a sense, the global similarity index is too powerful, in that it is measuring the degree to which results in both $\mathbb{R}^n$ and $\mathbb{R}^p$ match. One consequence of this is that if $\mathbf{X}_2$ is merely $\mathbf{X}_1$ with some of its columns permuted, then $\cos(\mathbf{X}_1, \mathbf{X}_2)$ will not be one. Although the PCs and their importance will be identical, the eigenvectors of their covariance matrices are not, since they will reflect the permutation in the order of the variables. The same logic applies to permutations of the rows of a data matrix. These considerations are linked to the above words of caution on the applicability of the index, but do not detract from its usefulness whenever applicable.

In another sense, the global similarity index is not powerful enough, in that it compares *individual* PAs in both spaces, but not the subspaces which they successively span. This problem, which has been discussed by Krzanowski (1988) [38, (Section 5.3)], will be further considered in Subsection III.2.4.

It should be pointed out that the possible ambiguities in the SVD of a matrix (such as sign-switching and the multiple possible choices when there are equal singular values) do not affect the value of s_g. This is best seen by noting that the definition (III.1.8) of the index does not depend on any particular SVDs of the matrices.

The global similarity index of the PCAs of $\mathbf{X}_1$ and $\mathbf{X}_2$ can also be written as:

$$\cos(\mathbf{X}_1, \mathbf{X}_2) = \frac{\mathrm{tr}(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_2)}{\sqrt{\mathrm{tr}(\mathbf{S}_1) \cdot \mathrm{tr}(\mathbf{S}_2)}}$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ are the covariance matrices defined by $\mathbf{X}_1$ and $\mathbf{X}_2$. The matrix $\frac{1}{n}\mathbf{X}_1'\mathbf{X}_2$ is the cross-covariance matrix whose (i,j)-th element is the covariance between the i-th variable in the first data set ($\mathbf{x}_i{}^{[1]}$) and the j-th variable in the second data set ($\mathbf{x}_j{}^{[2]}$). Thus, the global similarity index only depends on the variances of each set of variables and the covariances between *corresponding* variables in both sets:

$$\cos(\mathbf{X}_1, \mathbf{X}_2) = \frac{\sum_{i=1}^{p} \mathrm{cov}(\mathbf{x}_i{}^{[1]}, \mathbf{x}_i{}^{[2]})}{\sqrt{\left(\sum_{i=1}^{p} \mathrm{var}(\mathbf{x}_i{}^{[1]})\right) \left(\sum_{j=1}^{p} \mathrm{var}(\mathbf{x}_j{}^{[2]})\right)}} \qquad \text{(III.2.2)}$$

This alternative expression for the index – which gives another way of seeing the importance of a meaningful comparison of the p corresponding columns in both data matrices – provides further insight into how the results of a PCA are affected by changes in the data.

III.2.2 A similarity index for PCs

In Chapter I we saw how the unit-norm PCs of a (column-centered) $n \times p$ data matrix $\mathbf{X}$ are the orthonormal eigenvectors of matrix $\frac{1}{n}\mathbf{X}\mathbf{X}'$. We also saw how matrices of this form are always positive semi-definite, in which case their SVD and Spectral Decompositions coincide.

If we are given two matrices $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{C}_{n \times p}$, the cosine between the $n \times n$ matrices $\mathbf{X}_1\mathbf{X}_1'$ and $\mathbf{X}_2\mathbf{X}_2'$ (no need for the scalar) is given by:

$$\cos(\mathbf{X}_1\mathbf{X}_1', \mathbf{X}_2\mathbf{X}_2') = \sum_{i=1}^{r}\sum_{j=1}^{q} \sqrt{\pi_i^{S_1^2}\pi_j^{S_2^2}}(\boldsymbol{\alpha}_i'\boldsymbol{\beta}_j)^2 \qquad \text{(III.2.3)}$$

where $r, q, \boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_j$ are as before and $\mathbf{S}_k^2 = (\frac{1}{n}\mathbf{X}_k'\mathbf{X}_k)^2$, for $k = 1, 2$. A direct derivation of (III.2.3) produces weights using the relative eigenvalues of matrices $(\mathbf{X}_k\mathbf{X}_k')^2$. But we have used the matrices $\mathbf{S}_k^2$ in the notation. Their non-zero eigenvalues are the same, as was seen in Subsection I.4.1. Thus, we have a *weighted sum of the squares of cosines of the angles (correlations) between all possible pairs of PCs from* $\mathbf{X}_1$ *and* $\mathbf{X}_2$. *The weights are the geometric means of the relative eigenvalues of* $\mathbf{S}_k^2$ *associated with the vectors in the pair.* The Principal Axes in $\mathbb{R}^p$ play no role in (III.2.3).

The cosine between $\mathbf{X}_1\mathbf{X}_1'$ and $\mathbf{X}_2\mathbf{X}_2'$ is an indicator of similarity between the PCs of $\mathbf{X}_1$ and $\mathbf{X}_2$. We shall call it the **similarity index in $\mathbb{R}^n$** for the two PCAs and represent it by $s_n(\mathbf{X}_1, \mathbf{X}_2)$.

Unfortunately, the weights in (III.2.3) are not the same as in the global similarity index. They do not involve the relative eigenvalues of the covariance matrices, but rather those of the squares of these matrices. However, both sets of weights play similar roles. Denoting the vector of $\mathbf{S}_k$'s eigenvalues ($k = 1, 2$) by $\boldsymbol{\lambda}^{[k]}$, we have that the weights in (III.2.3) can be written as:

$$\sqrt{\pi_i^{S_1^2}\pi_j^{S_2^2}} = \frac{\lambda_i^{[1]}}{\|\boldsymbol{\lambda}^{[1]}\|_2} \cdot \frac{\lambda_j^{[2]}}{\|\boldsymbol{\lambda}^{[2]}\|_2}$$

where $\|\cdot\|_2$ is the usual (ℓ_2) Euclidean norm for vectors in $\mathbb{R}^p$. Thus, the weights for the similarity index in $\mathbb{R}^n$, although different from those in the global similarity index, also reward 'important' PCs and penalize PCs which account for little of the total variance. We shall call each eigenvalue divided by the ℓ_2-norm of the vector of all eigenvalues, the **$\boldsymbol{\ell}_2$-norm relative eigenvalues**. Division by the ℓ_1-norm, that is, the conventional standardization of eigenvalues in PCA, will produce the **$\boldsymbol{\ell}_1$-norm relative eigenvalues**, or just relative eigenvalues, introduced in Subsection I.4.1.

The similarity index for PCs has been used in other contexts. It is known as **Escoufier's RV-coefficient** for the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ (see Robert and Escoufier (1976) [57] or Ramsay *et al* (1984) [52], who point out its equivalence to $\cos(\mathbf{X}_1\mathbf{X}_1', \mathbf{X}_2\mathbf{X}_2')$). This coefficient was devised as a measure of global likeness of two configurations of n points in $\mathbb{R}^p$ and $\mathbb{R}^q$ (q does not have to be equal to p) given by the n rows of data matrices $\mathbf{X} \in \mathbb{M}_{n \times p}$ and $\mathbf{Y} \in \mathbb{M}_{n \times q}$. Escoufier sought to ensure that the coefficient was insensitive to translations of the origins in $\mathbb{R}^p$ and/or $\mathbb{R}^q$, global changes of scale and rotations in $\mathbb{R}^p$ and/or $\mathbb{R}^q$. It is not surprising to find that it can be re-interpreted in terms of the PCs of $\mathbf{X}$ and $\mathbf{Y}$, which have these properties, as was seen in Chapter I.

III.2.3 A similarity index for loadings

An analogous definition can provide a similarity index which focuses on the results in $\mathbb{R}^p$. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are the two column-centered data matrices whose PAs in $\mathbb{R}^p$ we wish to compare, then a **similarity index in $\mathbb{R}^p$** can be defined as:

$$s_p(\mathbf{X}_1, \mathbf{X}_2) = \cos(\mathbf{S}_1, \mathbf{S}_2) = \sum_{i=1}^{r} \sum_{j=1}^{q} \sqrt{\pi_i^{S_1^2} \pi_j^{S_2^2} (\mathbf{a}_i{}'\mathbf{b}_j)^2} \qquad \text{(III.2.4)}$$

where $\mathbf{S}_k, (k = 1, 2)$ are the covariance matrices of $\mathbf{X}_1$ and $\mathbf{X}_2$.

This is, of course, Escoufier's RV-coefficient for $\mathbf{X}_1'$ and $\mathbf{X}_2'$. We are merely reversing the roles of $\mathbb{R}^n$ and $\mathbb{R}^p$ in the previous Subsection. The weights in (III.2.3) and (III.2.4) are the same.

Data matrices whose similarity index in $\mathbb{R}^p$ is 1 have proportional covariance matrices, which Flury (1988) [15, (pg.60)] describes as the second level of similarity between covariance matrices (equality of all elements representing the first level).

As with both previous indices, it is not necessary to actually carry out the PCAs of the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. In fact, in this case it is not even necessary to know the data matrices. Knowledge of their covariance matrices is sufficient.

Furthermore, the index can be meaningfully used even when the individuals in both data sets are not directly comparable, as is the case with two independent samples of observations of the same p variables. Even the number of individuals in each sample can be different.

One important difference between the similarity indices in $\mathbb{R}^n$ and $\mathbb{R}^p$, on the one hand, and the global similarity index on the other, is that the former can only take positive values, as can be seen from the r.h.s of definitions (III.2.3) and (III.2.4), whereas the global similarity index can take both positive and negative values, as can be seen by considering $s_g(\mathbf{A}, -\mathbf{A})$ for any matrix $\mathbf{A}$. This difference can be viewed as the result of the presence or absence of *squares* of the cosines in the defining formulas. The different way in which the cosines are dealt with in both cases (which is also related to the differences in the weights which was discussed above), although somewhat unpleasant, can be seen as a reflection of the sign-switching ambiguities. We have seen how reflections about the origin are irrelevant in eigendecompositions, but not in SVDs — unless they are tied up with similar reflections of the corresponding singular vector in the other space. The global similarity index is picking up not just the similarity in each space, but also the connection between vectors in both spaces, as given by equation (I.4.17).

III.2.4 Similarity indices for subspaces

Common to the indicators studied so far is the fact that they focus on the degree of coincidence between *individual* PCs and/or PAs in $\mathbb{R}^p$ for two different data matrices.

However, as Krzanowski (1988) [38, (pg.155)] points out, it might be the case that the first k PCs, say, of two data matrices span the same subspace of $\mathbb{R}^n$, even if no pair of PCs actually coincide. The previous similarity indices do not convey the information that such 'Principal Subspaces' are equal.

The problem of comparing subspaces spanned by two different sets (with equal number) of arbitrary vectors in $\mathbb{R}^m$ has also been studied by Yanai, who suggested the **Generalized Coefficient of Determination (GCD)** for this purpose (see Ramsay *et al* (1984) [52]). Given two mxq matrices $\mathbf{A}$ and $\mathbf{B}$, whose columns are linearly independent, their GCD is defined as:

$$\mathrm{GCD}(\mathbf{A}, \mathbf{B}) = \frac{\mathrm{tr}(\mathbf{P}_A \mathbf{P}_B)}{q} \qquad \text{(III.2.5)}$$

where $\mathbf{P}_A = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ and $\mathbf{P}_B = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$.

The GCD compares the degree of similarity of the subspaces of $\mathbb{R}^m$ spanned by the columns of $\mathbf{A}$ and $\mathbf{B}$. When the columns of $\mathbf{A}$ and $\mathbf{B}$ are orthonormal sets, we have the case discussed in detail in Section 5.3 of Krzanowski (1988) [38].

Yanai's GCD is also a cosine between two matrices in $\mathbb{M}_m$, the matrices $\mathbf{P}_A$ and $\mathbf{P}_B$ of orthogonal projections onto the subspaces spanned by the columns of $\mathbf{A}$ and $\mathbf{B}$ ($\Re(\mathbf{A})$ and $\Re(\mathbf{B})$), respectively (see Proposition A.19). In fact, we have:

$$\mathrm{GCD}(\mathbf{A}, \mathbf{B}) = \cos(\mathbf{P}_A, \mathbf{P}_B) \qquad \text{(III.2.6)}$$

To address the problem of comparing Principal Subspaces, $\mathbf{A}$ and $\mathbf{B}$ should be either $n \times q$ (if comparing in $\mathbb{R}^n$) or $p \times q$ (in $\mathbb{R}^p$) matrices whose columns are the first

q PAs in either $\mathbb{R}^n$ or $\mathbb{R}^p$ for $\mathbf{X}_1$ and $\mathbf{X}_2$. It is also possible to consider different number of PAs for each data matrix, equation (III.2.6) not requiring that $\mathbf{P}_A$ and $\mathbf{P}_B$ be of the same rank.

Unlike previous indicators, it is actually necessary to compute the PCAs of $\mathbf{X}_1$ and $\mathbf{X}_2$ in order to calculate the GCD. In addition, it is necessary to calculate one GCD for each pair of subspaces we wish to compare. Thus, comparing two PCAs might require several values of the GCD.

III.2.5 An example

An example of how the similarity indices work in practice will now be considered.

Two different, but related data sets are given in Tables III.1 and III.2. The data sets consist of the same four meteorological variables, observed at the Lisbon Meteorological Station, for each day of January in two different years: 1948 and 1967. The four observed variables are: maximum temperature (in degrees Centigrade); minimum temperature (in degrees Centigrade); sunshine (hours of sunshine as a proportion of total daylight hours); and rainfall (in millimeters). The sources for these data sets are the 1948 and 1967 Climatological Yearbooks (Anuários Climatológicos) of the Portuguese National Institute for Meteorology and Geophysics (Instituto Nacional de Meteorologia e Geofísica).

The two data sets are obviously related, but a comparison involving $\mathbb{R}^n$ does not

Day	Max Temp	Min Temp	Sunshine	Rainfall
1	15.2	5.3	0.91	0.0
2	14.7	7.5	0.94	0.0
3	16.3	5.5	0.88	0.0
4	17.7	6.8	0.42	0.0
5	17.0	13.4	0.09	4.7
6	16.9	12.9	0.00	0.6
7	17.0	11.0	0.65	17.2
8	16.6	9.5	0.39	1.4
9	17.3	13.3	0.05	11.2
10	16.7	11.9	0.13	1.0
11	17.6	9.5	0.63	3.8
12	16.2	9.1	0.70	0.3
13	14.8	6.7	0.94	0.0
14	15.9	6.5	0.95	0.0
15	14.5	9.4	0.52	0.5
16	12.7	7.0	0.90	7.0
17	13.4	5.8	0.88	0.0
18	15.6	8.1	0.29	0.0
19	15.8	8.4	0.46	3.3
20	13.1	7.7	0.78	7.6
21	16.1	7.6	0.42	0.9
22	15.0	11.4	0.00	0.3
23	12.7	4.8	0.27	18.2
24	13.7	4.4	0.56	4.0
25	15.8	9.0	0.38	4.6
26	15.4	11.4	0.18	22.0
27	16.0	11.4	0.66	15.6
28	16.0	13.1	0.00	0.0
29	17.0	13.0	0.48	3.4
30	16.9	11.7	0.23	0.0
31	17.3	10.4	0.19	0.0
Means	16.2	9.5	0.48	4.1

Table III.1: Data recorded at the Lisbon Meteorological Station on the 31 days of January, 1948

Day	Max Temp	Min Temp	Sunshine	Rainfall
1	16.3	6.0	0.88	0.0
2	13.9	5.6	0.73	0.0
3	14.2	7.6	0.88	0.0
4	14.9	5.0	0.77	0.0
5	16.0	8.4	0.57	0.0
6	13.4	6.7	0.94	0.0
7	11.8	3.3	0.89	0.0
8	12.9	7.2	0.08	0.2
9	7.6	5.2	0.00	3.7
10	11.5	5.4	0.60	37.8
11	8.0	4.0	0.12	0.2
12	10.4	1.6	0.82	0.0
13	11.8	4.4	0.94	0.0
14	13.0	4.7	0.90	0.0
15	13.6	4.3	0.76	0.0
16	12.8	2.9	0.74	0.0
17	15.6	5.5	0.80	0.0
18	14.2	6.1	0.26	0.0
19	15.2	9.9	0.36	6.5
20	15.0	9.2	0.28	2.4
21	15.5	7.8	0.33	2.3
22	14.6	6.6	0.64	5.7
23	16.1	10.1	0.26	6.6
24	15.0	10.2	0.02	0.3
25	15.8	12.5	0.04	0.1
26	18.0	12.3	0.46	0.0
27	16.4	9.1	0.52	0.0
28	18.6	9.1	0.32	0.0
29	18.3	10.2	0.77	0.0
30	14.8	9.0	0.30	8.8
31	14.6	7.7	0.77	0.1
Means	14.2	7.0	0.54	2.4

Table III.2: Data recorded at the Lisbon Meteorological Station on the 31 days of January, 1967

seem advisable. In fact, there is no reason to suppose that any single day in January 1948 is specifically related to the same day in January 1967. For this reason, only the similarity index in $\mathbb{R}^p$ will be computed. In later Chapters, examples of applications of the global similarity index and the similarity index in $\mathbb{R}^n$ will also be given.

Since the data are in different units of measurement, a standardization of the data and a Correlation Matrix PCA are appropriate. In Table III.3 we find the correlation matrices determined by the two data sets.

Variables	Maximum Temp. (°C)	Minimum Temp. (°C)	Sunshine (proportion)	Rainfall (mm.)
Max.Temp.	1.00	(+0.71)	(+0.04)	(−0.18)
Min. Temp.	[+0.59]	1.00	(−0.48)	(−0.01)
Sunshine	[−0.40]	[−0.69]	1.00	(−0.09)
Rainfall	[−0.17]	[+0.15]	[−0.11]	1.00

Table III.3: Correlation matrices for the 1948 (in square brackets) and 1967 (in round brackets) Meteorological data sets

Table III.4 gives us the eigenvalues, ℓ_1-norm and ℓ_2-norm relative eigenvalues of both correlation matrices. These latter quantities are used in the similarity index in $\mathbb{R}^p$. The picture emerging from Table III.4 is one of overall affinity in the relative eigenvalues for both data sets.

In Table III.5 we have the unit-norm PAs in $\mathbb{R}^p$ for both data sets. The global picture that emerges is one of fairly similar first and last vectors, with differences as well as common traits in the second and third PAs.

The broad similarity between corresponding PAs from both sets is confirmed by

1948			1967		
Eigenvalue	ℓ_1-rel. eig.	ℓ_2-rel. eig.	Eigenvalue	ℓ_1-rel. eig.	ℓ_2-rel. eig.
2.1304	0.5326	0.8636	1.8493	0.4623	0.7838
1.1020	0.2755	0.4467	1.1912	0.2978	0.5049
0.5220	0.1305	0.2116	0.8456	0.2114	0.3584
0.2456	0.0614	0.0995	0.1140	0.0285	0.0483

Table III.4: Eigenvalues, ℓ_1-norm and ℓ_2-norm relative eigenvalues of the correlation matrices of the 1948 and 1967 Meteorological data

First PA		Second PA		Third PA		Fourth PA	
1948	1967	1948	1967	1948	1967	1948	1967
0.5229	0.5959	0.3773	0.3969	0.6508	0.3700	−0.4008	−0.5920
0.6295	0.7092	−0.0925	−0.0942	0.0227	0.0711	0.7711	0.6951
−0.5714	−0.3649	0.1464	0.6341	0.6604	0.5506	0.4646	0.4019
0.0609	−0.0939	−0.9098	−0.6569	0.3739	0.7449	−0.1699	−0.0694

Table III.5: Unit-norm PAs in $\mathbb{R}^p$ for the 1948 and 1967 Meteorological data sets

the large values in the diagonal of Table III.6, which gives the cosines between each of the four PAs for the 1948 data set and each of the four PAs for the 1967 data set.

		1967			
yr.	PA no.	PA1	PA2	PA3	PA4
1	PA1	0.96	−0.25	−0.03	−0.11
9	PA2	0.19	0.85	−0.46	−0.17
4	PA3	0.13	0.43	0.88	−0.13
8	PA4	0.15	0.17	0.04	0.97

Table III.6: Cosines of the angles between the PAs in $\mathbb{R}^p$ of the 1948 and 1967 Meteorological data sets

From the above Tables we would expect the similarity index in $\mathbb{R}^p$ to have a relatively high value, indicating a good degree of agreement between the eigendecompositions of the correlation matrices of both data sets (when ℓ_2-norm relative eigenvalues are considered). The value of s_p is in fact quite high: 0.9541. It will be recalled that the implication is that the two correlation matrices form an angle of just under $17°30'$ in the PSD cone of $\mathbb{S}_4$.

III.3 The PSD cone

With the exception of orthogonal matrices, the examples of square matrices given in Section III.1 all share an additional property: they are positive semi-definite (**p.s.d.**) matrices.

Positive semi-definite matrices are particularly rich in geometric relations which provide interesting visualizations of concepts relevant to PCA. In this Section we shall consider some geometric properties of **p.s.d.** matrices.

III.3.1 Definitions

If $\mathcal{L}$ is a linear space, a subset $\mathcal{C} \subset \mathcal{L}$ is called a **cone** if for any pair of its elements, $x, y \in \mathcal{C}$ and any non-negative scalars $\alpha, \beta \geq 0$ the linear combination $\alpha x + \beta y$ is also an element of the subset $\mathcal{C}$ (see Hill and Waters (1987) [23]). It can easily be verified that the $m \times m$ **p.s.d.** matrices form a cone in either the linear space $\mathbb{M}_m$ of all $m \times m$ square matrices or the linear space $\mathbb{S}_m$ of $m \times m$ symmetric matrices (see also the previous reference or Barker (1981) [3]). We shall work with the cone of **p.s.d.** matrices defined in $\mathbb{S}_m$ and shall denote it (using the notation of Hill and

Waters) as **PSD** .

Given any matrix $\mathbf{V} \in$ **PSD** , any non-negative scalar multiple of it will also be in the cone. The set of all such scalar multiples, $\{\alpha\mathbf{V}\}$ (with $\alpha \geq 0$) will be called a **ray**, defined by any matrix on it. The matrices in any given ray are the **p.s.d.** matrices with Flury's ([15]) second level of similarity. All rays in the **PSD** cone meet at the origin of $\mathbb{S}_m$, the $m \times m$ zero matrix, which is also an element of the **PSD** cone. All matrices in a given ray share the same eigenvectors and relative eigenvalues. The ray defined by the $m \times m$ identity matrix (which is also in the **PSD** cone) will be called the **central ray** for reasons which are to become clear shortly.

The angle between any two matrices in the **PSD** cone can never exceed 90°. In fact, Fejer's theorem (see Horn and Johnson (1986) [25, (pg.459)]) tells us that a matrix $\mathbf{V} \in \mathbb{M}_m$ is **p.s.d.** if and only if for all **p.s.d.** matrices $\mathbf{W}$ we have $\sum_{i=1}^{m}\sum_{j=1}^{m} v_{ij}w_{ij} \geq 0$. But this means, from equation (III.1.2) that the inner product of any two **p.s.d.** matrices is non-negative. Hence, $\mathbf{V} \in$ **PSD** if and only if:

$$\cos(\mathbf{V},\mathbf{W}) \geq 0 \ , \quad \forall \mathbf{W} \in \mathbf{PSD} \qquad \text{(III.3.1)}$$

Incidentally, this confirms that *the similarity indices in* $\mathbb{R}^n$ *and* $\mathbb{R}^p$*and Yanai's G.C.D. can only take values in the interval [0,1].*

III.3.2 Eigenvalues and the PSD cone

Tarazaga (1990) [65] has shown that the **PSD** cone has a special geometric structure with the central ray playing a pivotal role. In fact, many of Tarazaga's results arise from considering the angle between any **p.s.d.** matrix and the identity $\mathbf{I}_m$.

Working directly from the definitions, Tarazaga noted that for any non-zero $m \times m$ **p.s.d.** matrix $\mathbf{V}$, we have:

$$\cos(\mathbf{V}, \mathbf{I}_m) = \frac{\mathrm{tr}(\mathbf{V})}{\sqrt{\mathrm{tr}(\mathbf{V}'\mathbf{V}) \cdot m}} = \frac{||\boldsymbol{\lambda}||_1}{\sqrt{m} \cdot ||\boldsymbol{\lambda}||_2} \qquad \text{(III.3.2)}$$

where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_m)$ is the vector of $\mathbf{V}$'s m eigenvalues (all real and non-negative) and $||\cdot||_1$ and $||\cdot||_2$ are the usual ℓ_1- and ℓ_2-norms on vectors in $\mathbb{R}^m$. Thus, $\cos(\mathbf{V}, \mathbf{I}_m)$ is the *cosine of the angle in* $\mathbb{R}^m$ *between vectors* $\boldsymbol{\lambda}$ *and* $\mathbf{1}_m$, where $\mathbf{1}_m$ is the vector of ones.

Using the concept of relative eigenvalue defined in Chapter I for any **p.s.d.** matrix, we can improve on Tarazaga's expression (III.3.2) for $\cos(\mathbf{V}, \mathbf{I}_m)$. In fact, we have:

$$\cos(\mathbf{V}, \mathbf{I}_m) = \frac{1}{\sqrt{m} \cdot \sqrt{\sum_{i=1}^m \pi_i^2}} \qquad \text{(III.3.3)}$$

where

$$\pi_i = \frac{\lambda_i}{\sum_{i=1}^m \lambda_i}$$

is $\mathbf{V}$'s i-th relative eigenvalue. Thus, the angle between a **p.s.d.** matrix $\mathbf{V}$ and the identity depends only on the sum of squares of $\mathbf{V}$'s relative eigenvalues.

It might be helpful to re-interpret that quantity in terms of the *variance* of **V**'s relative eigenvalues. By definition, the mean of the relative eigenvalues must be $\frac{1}{m}$. Thus, the variance of the m relative eigenvalues is:

$$var_\pi = \frac{1}{m}\sum_{i=1}^{m}\pi_i^2 - \frac{1}{m^2}$$

Formula (III.3.3) now becomes:

$$\cos(\mathbf{V}, \mathbf{I}_m) = \frac{1}{\sqrt{1 + m^2 \cdot var_\pi}} \qquad \text{(III.3.4)}$$

If all relative eigenvalues of **V** are equal (which would imply that **V** is a scalar multiple of $\mathbf{I}_m$) the angle between **V** and $\mathbf{I}_m$ will be zero. This justifies defining $\cos(\mathbf{V}, \mathbf{I}_m) = 1$ if **V** is the $m \times m$ matrix of zeros, a case in which formula (III.3.2) does not apply. As the dispersion of **V**'s relative eigenvalues grows, so will its angle with $\mathbf{I}_m$.

Since **p.s.d.** matrices which are not of full rank have some eigenvalues fixed at zero, we will not be surprised to see that the rank of **V** bears some relation to its angle with the identity. Tarazaga proves the following result.

Theorem III.1 (Tarazaga) *If* **V** *is a rank* k m×m *positive semi-definite matrix, then* $\cos(\mathbf{V}, \mathbf{I}_m) \leq \sqrt{\frac{k}{m}}$.

Proof. Tarazaga proves this result by noting that, for any vector $\mathrm{x} \in \mathbb{R}^k$, $||\mathbf{x}||_1 \leq \sqrt{k} \cdot ||\mathbf{x}||_2$ (as can be seen by applying the Cauchy-Schwarz inequality to $||\mathbf{x}||_1 =$

$\sum_{i=1}^{k} |\mathrm{x}_i| \cdot |1|$). Thus, defining $\boldsymbol{\lambda}$ to be the vector of $\mathbf{V}$'s k non-zero eigenvalues, the Theorem is a direct consequence of the above inequality applied to (III.3.2). □

On the other hand, for any vector x, we have $\|\mathrm{x}\|_1 \geq \|\mathrm{x}\|_2$ (since for any $\mathbf{x}$ we have $\|\mathrm{x}\|_1^2 = (\sum_{i=1}^{k} |x_i|)^2 \geq \sum_{i=1}^{k} |x_i|^2 = \|\mathrm{x}\|_2^2$). Thus, it is also true that for any **p.s.d.** matrix $\mathbf{V}$, we have:

$$\sqrt{\tfrac{1}{m}} \leq \cos(\mathbf{V}, \mathbf{I}_m) \qquad \text{(III.3.5)}$$

One implication of these results is that *any* $m \times m$ **p.s.d.** matrix $\mathbf{V}$ forms an angle no greater than $\arccos(\frac{1}{\sqrt{m}})$ with $\mathbf{I}_m$. All rank one matrices must form precisely such an angle with the identity. Rank k matrices can be found in the section of the **PSD** cone where angles with $\mathbf{I}_m$ are between $\arccos(\frac{1}{\sqrt{m}})$ and $\arccos(\sqrt{\frac{k}{m}})$. Only full rank (positive definite) matrices can form angles with $\mathbf{I}_m$ smaller than $\arccos(\sqrt{\frac{m-1}{m}})$.

It is not hard to see that in the **PSD** cone the maximum angle with $\mathbf{I}_m$ is achieved *only* by rank one **p.s.d.** matrices.

Theorem III.2 *Any* m×m **p.s.d.** *matrix* $\mathbf{V}$ *forms the maximum angle* $\arccos(\frac{1}{\sqrt{m}})$ *with* $\mathbf{I}_m$ *if and only if it is a rank one* **p.s.d.** *matrix.*

Proof. We can exclude the case where $\mathbf{V} = \mathbf{0}_m$ since we have defined $\cos(\mathbf{0}_m, \mathbf{I}_m) = 1$. For non-zero matrices $\mathbf{V}$, we have already seen that if $\mathbf{V}$ is of rank one, then $\cos(\mathbf{V}, \mathbf{I}_m) = \frac{1}{\sqrt{m}}$. Conversely, if $\cos(\mathbf{V}, \mathbf{I}_m) = \frac{1}{\sqrt{m}}$, then from equation (III.3.3) we have $\sum_{i=1}^{m} \pi_i^2 = 1$. But, by the definition of relative eigenvalues, we also have

$\sum_{i=1}^{m} \pi_i = 1$. Hence:

$$\begin{aligned} \sum_{i=1}^{m} \pi_i^2 &= \sum_{i=1}^{m} \pi_i \\ \Longleftrightarrow \quad \sum_{i=1}^{m} \pi_i(1-\pi_i) &= 0 \end{aligned} \qquad \text{(III.3.6)}$$

Since the relative eigenvalues must all satisfy the inequalities

$$0 \leq \pi_i \leq 1 \ , \qquad \forall i = 1, ..., m$$

all terms in (III.3.6) are non-negative. Hence, they must all be zero. Then, all relative eigenvalues are equal to either zero or one. But for $\mathbf{V} \neq \mathbf{0}_m$ this is only possible if there is a single non-zero relative eigenvalue. Hence $\mathbf{V}$ is of rank one. □

For a general rank k **p.s.d.** matrix $\mathbf{V}$ it is not possible to improve on the bounds for $\cos(\mathbf{V}, \mathbf{I}_m)$ given by Theorem III.1 and inequality (III.3.5), except by making the latter a strict inequality when $k > 1$. In fact, for any $k > 1$, there will always be rank k matrices whose angles with $\mathbf{I}_m$ are arbitrarily close (though not equal to) the maximum angle $\arccos(\frac{1}{\sqrt{m}})$, since $k-1$ of the non-zero relative eigenvalues can be taken arbitrarily close to zero. On the other hand, a rank k matrix $\mathbf{V} \in \mathbf{PSD}$ with all k non-zero eigenvalues equal means that $\sum_{i=1}^{m} \pi_i^2 = k\left(\frac{1}{k^2}\right) = \frac{1}{k}$. Then, the equality in Theorem III.1 is attained. We shall now see that the matrices of this sort are the *only* rank k **p.s.d.** matrices for which the minimum angle is achieved.

Theorem III.3 *If* $\mathbf{V}$*is a rank* k m×m **p.s.d.** *matrix, then* $\cos(\mathbf{V}, \mathbf{I}_m) = \sqrt{\frac{\mathrm{k}}{\mathrm{m}}}$ *if and only if all of* $\mathbf{V}$*'s non-zero eigenvalues are equal.*

Proof. For any rank k matrix $\mathbf{V}$, we have, from equation (III.3.3):

$$\cos(\mathbf{V}, \mathbf{I}_m) = \sqrt{\tfrac{k}{m}} \iff \sum_{i=1}^{m} \pi_i^2 = \sum_{i=1}^{k} \pi_i^2 = \tfrac{1}{k}$$

We have already seen that if $\mathbf{V}$'s k non-zero eigenvalues are equal, this equation is satisfied. We shall now prove the converse. Let us assume that $\mathbf{V}$'s k non-zero relative eigenvalues are written in terms of their difference from their mean value $\frac{1}{k}$:

$$\pi_i = \frac{1}{k} + d_i \quad (i = 1, ..., k)$$

Since, by definition, $\sum_{i=1}^{m} \pi_i = \sum_{i=1}^{k} \pi_i = 1$ we must have $\sum_{i=1}^{k} d_i = 0$. At the same time, substituting the above expression for each π_i in $\sum_{i=1}^{k} \pi_i^2$, we have:

$$\sum_{i=1}^{k} \pi_i^2 = \frac{1}{k} + \frac{2}{k} \sum_{i=1}^{k} d_i + \sum_{i=1}^{k} d_i^2$$

Thus,

$$\begin{aligned} \cos(\mathbf{V}, \mathbf{I}_m) = \sqrt{\tfrac{k}{m}} & \iff \sum_{i=1}^{k} d_i^2 = 0 \\ & \iff d_i = 0 \end{aligned}$$

In other words, the minimum angle for rank k matrices is achieved if and only if $\mathbf{V}$ has all equal non-zero eigenvalues. □

We have not, of course, ruled out the possibility that matrices of rank greater than k will also form an angle of $\arccos(\sqrt{\frac{k}{m}})$ with the central ray. What is being said is that the rays of rank k **p.s.d.** matrices which form the smallest possible angle with the central ray are scalar multiples of matrices with all non-zero eigenvalues equal to one. It can easily be checked that the latter matrices are idempotent, hence rank k matrices of orthogonal projections (see Proposition A.19).

III.3.3 Projecting p.s.d matrices onto the central ray

The identity matrix $\mathbf{I}_m$ defines the central ray in the **PSD** cone. But it also defines a one-dimensional subspace in $\mathbb{S}_m$ if we consider *all* scalar products $\alpha\mathbf{I}_m$ (and not just those where $\alpha \geq 0$).

Some interesting results emerge if we consider the orthogonal projections of any **p.s.d.** matrix onto the subspace spanned by the identity $\mathbf{I}_m$. These results could be obtained by considering orthogonal projections onto the central ray (which is also a cone). Orthogonal projections on cones are described, among others, by Critchley (1980) [9]. But the presentation is easier and more familiar if we consider projections onto the whole subspace spanned by $\mathbf{I}_m$, which contains the central ray. The images of these projections will always belong to the central ray and will coincide with the alternative approach.

As is known (see, for example, Basilevski (1983) [4, (pg.58)]), for any inner product linear space $\mathcal{L}$ and any of its subspaces $\mathcal{S}$, the image x^* of the **orthogonal projection** of $x \in \mathcal{L}$ onto $\mathcal{S}$ is defined by the equation:

$$< x - x^*, s >= 0 \quad \forall s \in \mathcal{S} \tag{III.3.7}$$

The distance from x to its orthogonal projection x^* is, of course, given by $\|x - x^*\|$.

This is also the smallest distance between x and any point on the subspace $\mathcal{S}$ (Basilevski (1983) [4, (pg.58)]) and will be called the *distance from* x *to the subspace* $\mathcal{S}$.

We now have the following Theorem.

Theorem III.4 *Let* $\mathbf{V} \in$ **PSD** . *The image of the orthogonal projection of* $\mathbf{V}$ *onto the subspace* $span(\mathbf{I}_m)$ *is* $\mathbf{V}^* = \overline{\lambda}\mathbf{I}_m$, *where* $\overline{\lambda}$ *is the mean of the eigenvalues of* $\mathbf{V}$. *The distance from* $\mathbf{V}$ *to* $\mathbf{V}^*$ *is given by* $\sqrt{m \cdot var_\lambda}$, *where* var_λ *is the variance of* $\mathbf{V}$*'s* m *eigenvalues*, $\lambda_1, ..., \lambda_m$.

Proof. $\mathbf{V}^*$ must be a scalar multiple of $\mathbf{I}_m$. Condition (III.3.7) thus becomes:

$$
\begin{aligned}
& < \mathbf{V} - \alpha^* \mathbf{I}_m, \alpha \mathbf{I}_m >= 0 \\
\Longleftrightarrow \quad & \alpha \cdot \mathrm{tr}(\mathbf{V}) = \alpha^* \alpha \cdot \mathrm{tr}(\mathbf{I}_m) \\
\Longleftrightarrow \quad & \alpha^* = \frac{\mathrm{tr}(\mathbf{V})}{m} = \frac{1}{m} \sum_{i=1}^{m} \lambda_i = \overline{\lambda}
\end{aligned}
$$

On the other hand, orthogonality implies that:

$$
\begin{aligned}
\|\mathbf{V} - \alpha^* \mathbf{I}_m\|^2 &= \|\mathbf{V}\|^2 - \|\alpha^* \mathbf{I}_m\|^2 \\
&= \sum_{i=1}^{m} \lambda_i^2 - m\overline{\lambda}^2 \\
&= m \cdot var_\lambda
\end{aligned}
$$

□

As mentioned previously, we worked with projections on $span(\mathbf{I}_m)$. But $\overline{\lambda} \geq 0$ for any **p.s.d.** matrix, so that the projected matrix actually lies on the central ray of the **PSD** cone.

We had already seen how the loci of matrices in the **PSD** cone forming a common angle with the central ray were characterized by those matrices' variance of relative eigenvalues. We have now seen that the loci of matrices in the **PSD** cone which are equidistant from the central ray are characterized by the variance of the (absolute) eigenvalues of those matrices.

It has also just been seen that the trace of a matrix $\mathbf{V}$ is the coefficient of the unit-norm matrix on $span(\mathbf{I}_m)$ (that is, of $\frac{1}{m}\mathbf{I}_m$) which gives the scalar multiple of the identity that is closest to $\mathbf{V}$.

In Section III.1 we saw how all matrices in $\mathbb{S}_m$ whose trace was k formed a hyperplane with normal $\mathbf{I}_m$, characterized by the equation $< \mathbf{B}, \mathbf{I}_m >= k$, $(\mathbf{B} \in \mathbb{S}_m)$. The orthogonal projection of $\mathbf{V}$ onto the central ray gives the unique point of intersection of the central ray with the hyperplane of symmetric matrices whose trace is $\mathrm{tr}(\mathbf{V})$.

We now consider some results concerning the eigenvectors of **p.s.d.** matrices.

III.3.4 Spectral Decompositions and the PSD cone

Tarazaga (1990) [65] stresses the importance of rank one symmetric matrices. These matrices can always be written in the form $\mathbf{E} = \mathbf{x}\mathbf{x}'$, for some vector $\mathbf{x} \in \mathbb{R}^m$ (with $\mathbf{x} = \sqrt{\lambda_1}\mathbf{a}_1$ if $\mathbf{E}$'s spectral decomposition is given by (I.3.1)). In particular, he notes that the orthogonal projection of any **p.s.d.** matrix $\mathbf{V}$ onto the subspace of scalar products of a given rank one matrix $\mathbf{x}\mathbf{x}'$ will give an important result.

Theorem III.5 (Tarazaga) *Let* $\mathbf{V} \in \mathbf{PSD}$ *and* $\mathbf{E} = \mathbf{x}\mathbf{x}'$ *for some vector* $\mathbf{x} \in \mathbb{R}^m$*. The image of the orthogonal projection of* $\mathbf{V}$ *onto the direction defined by* $\mathbf{E}$ *can be written as* $v^* \frac{\mathbf{E}}{||\mathbf{E}||}$*, where* $v^* = \frac{\mathbf{x}'\mathbf{V}\mathbf{x}}{\mathbf{x}'\mathbf{x}}$ *is the value of* $\mathbf{V}$*'s Rayleigh-Ritz quotient for the vector* $\mathbf{x}$*. The cosine of the angle between* $\mathbf{V}$ *and* $\mathbf{E}$ *is* $\frac{v^*}{||\mathbf{V}||}$*.*

Proof. From (III.3.7) we have that:

$$< \mathbf{V} - v^* \frac{\mathbf{E}}{||\mathbf{E}||}, v\frac{\mathbf{E}}{||\mathbf{E}||} >= 0$$

$$\Longleftrightarrow \quad \tfrac{v}{||\mathbf{E}||}\mathrm{tr}(\mathbf{V}\mathbf{E}) = \tfrac{v^* v}{||\mathbf{E}||^2}\mathrm{tr}(\mathbf{E}^2)$$

Since $\mathbf{E} = \mathbf{x}\mathbf{x}'$, $\mathrm{tr}(\mathbf{E}^2) = ||\mathbf{x}||^4_{\mathbb{R}^m} = ||\mathbf{E}||^2_{\mathbb{S}_m}$. Hence,

$$v^* = \frac{\mathrm{tr}(\mathbf{V}\mathbf{E})}{||\mathbf{E}||} = \frac{\mathbf{x}'\mathbf{V}\mathbf{x}}{\mathbf{x}'\mathbf{x}}$$

Finally,

$$\cos(\mathbf{V}, \mathbf{E}) = \frac{< \mathbf{V}, \mathbf{x}\mathbf{x}' >}{||\mathbf{V}|| \cdot ||\mathbf{x}\mathbf{x}'||} = \left(\frac{\mathbf{x}'\mathbf{V}\mathbf{x}}{\mathbf{x}'\mathbf{x}}\right) \cdot \frac{1}{||\mathbf{V}||} = \frac{v^*}{||\mathbf{V}||}$$

□

Corollary *The first unit-norm eigenvector* $\mathbf{a}$ *of a* **p.s.d.** *matrix* $\mathbf{V}$ *is characterized by the fact that the matrix* $\mathbf{E_a} = \mathbf{a}\mathbf{a}'$ *is the unit-norm rank one matrix which forms the smallest angle with* $\mathbf{V}$ *in the* **PSD** *cone. The cosine of this angle is* $\frac{\lambda_{max}(\mathbf{V})}{||\mathbf{V}||}$*.*

Proof. Minimizing the angle is maximizing $\cos(\mathbf{V}, \mathbf{a}\mathbf{a}') = \frac{\mathbf{a}'\mathbf{V}\mathbf{a}}{\mathbf{a}'\mathbf{a}} \cdot \frac{1}{||\mathbf{V}||}$. □

This cosine is the first ℓ_2-norm relative eigenvalue of $\mathbf{V}$, introduced in Subsection III.2.2.

The Corollary can be easily generalized to successive eigenpairs, providing an analogue to the Rayleigh-Ritz quotient variational characterization of eigenpairs in terms of the **PSD** cone. The following result will be crucial in proving that generalization.

Theorem III.6 *Let* $\mathbf{x} \in \mathbb{R}^m$ *and* $\mathbf{E_x} = \mathbf{xx}'$ *be the rank one* **p.s.d.** *matrix it defines. Then:*

i). $\mathrm{tr}(\mathbf{E_x}) = ||\mathbf{E_x}||$

ii). $||\mathbf{E_x}||_{\mathbb{S}_m} = 1 \iff ||\mathbf{x}||_{\mathbb{R}^m} = 1$

iii). $\mathbf{x}_1 \perp_{\mathbb{R}^m} \mathbf{x}_2 \iff \mathbf{E}_{\mathbf{x}_1} \perp_{\mathbb{S}_m} \mathbf{E}_{\mathbf{x}_2}$

iv). $\mathbf{E_x}$ *is idempotent if and only if* $||\mathbf{x}||_{\mathbb{R}^m} = 1$

Proof.

i). Trivial.

ii). Trivial, from the definitions.

iii). $< \mathbf{E}_{\mathbf{x}_1}, \mathbf{E}_{\mathbf{x}_2} >_{\mathbb{S}_m} = \mathrm{tr}(\mathbf{x}_1\mathbf{x}_1{}'\mathbf{x}_2\mathbf{x}_2{}') = (\mathbf{x}_1{}'\mathbf{x}_2)^2$. Thus, $< \mathbf{E}_{\mathbf{x}_1}, \mathbf{E}_{\mathbf{x}_2} >_{\mathbb{S}_m} = 0$ if and only if $< \mathbf{x}_1, \mathbf{x}_2 >_{\mathbb{R}^m} = 0$.

iv). $\mathbf{E}_\mathbf{x}^2 = ||\mathbf{x}||^2\mathbf{xx}' = ||\mathbf{x}||^2\mathbf{E_x}$. Thus, $\mathbf{E_x}$ is idempotent if and only if $||\mathbf{x}||_{\mathbb{R}^m} = 1$.

□

The above Theorem implies that the rank one symmetric idempotent matrices (*i.e.*, matrices of orthogonal projections – see Proposition A.19) lie on the intersection of the **PSD** cone with the unit hypersphere and with the trace-1 hyperplane in $\mathbb{S}_m$. All rank one matrices on this intersection are symmetric idempotent matrices.

Theorem III.7 *Let* $\mathbf{V} \in$ **PSD** . *Consider the problem of finding the rank one idempotent matrices* $\{\mathbf{b}_i\mathbf{b}_i'\}_{i=1}^m$ *which successively minimize the angles formed with* $\mathbf{V}$, *subject to the constraint of orthogonality with any previously determined solutions. The vectors* $\{\mathbf{b}_i\}_{i=1}^m$ *which characterize a complete set of solutions to the problem are eigenvectors of* $\mathbf{V}$. *The cosines of the angles formed by these solutions with* $\mathbf{V}$ *are* $\mathbf{V}$*'s* ℓ_2*-norm relative eigenvalues,* $\left\{\frac{\lambda_i}{\|\boldsymbol{\lambda}\|_2}\right\}_{i=1}^m$. *The coefficients of the orthogonal projection of* $\mathbf{V}$ *onto the rays defined by the solution matrices are* $\mathbf{V}$*'s (absolute) eigenvalues,* $\{\lambda_i\}_{i=1}^m$.

Proof. A trivial consequence of Theorems III.5 and III.6 and the Rayleigh-Ritz variational characterization of the eigenpairs of a symmetric matrix. □

It is interesting to note that the indeterminacy of sign-switching in the Spectral Decomposition of a **p.s.d.** matrix is directly reflected in the fact that the rank one **p.s.d.** matrices determined by vectors $\mathbf{x}$ and (-1)$\mathbf{x}$ are identical. Theorem III.7 implies that the indeterminacies which result from equal eigenvalues correspond to the fact that the matrix $\mathbf{V}$ is positioned in the **PSD** cone at equal angles with an infinity of rank one unit-norm matrices which solve the minimization problem (one for each unit-norm vector in the subspace spanned by the eigenvectors associated with equal eigenvalues). It therefore comes as no surprise that the identity matrix $\mathbf{I}_m$ (or any scalar multiple of it) which, as we saw in Theorem III.2, forms the same angle with all rank one matrices, has any orthogonal set of vectors in $\mathbb{R}^m$ as a complete set of eigenvectors.

III.3.5 Common eigenvectors

Theorem III.7 highlights the geometrical significance of a Spectral Decomposition of a **p.s.d.** matrix, in terms of the **PSD** cone. This decomposition expresses the matrix as a non-negative linear combination of symmetric idempotent matrices (matrices of orthogonal projections) of rank one.

If we are given a set of m mutually orthogonal rank one idempotent matrices in the **PSD** cone, $\mathcal{S} = \{\mathbf{E}_{\mathbf{a}_i}\}_{i=1}^{m}$, then we can speak of the cone spanned by these matrices. This cone is by definition the set of all linear combinations of the form $\sum_{i=1}^{m} \mu_i \mathbf{E}_{\mathbf{a}_i}$ for some non-negative scalars $\{\mu_i\}_{i=1}^{m}$. The eigenpairs of any matrix in this subcone are $\{(\mu_i, \mathbf{a}_i)\}_{i=1}^{m}$. This cone is a *subcone* of the **PSD** cone since the original 'building blocks' (the set $\mathcal{S}$ of idempotent matrices) are **p.s.d.** matrices. The elements of this subcone are all the **p.s.d.** matrices which have $\{\mathbf{a}_i\}_{i=1}^{m}$ as a set of unit-norm eigenvectors, regardless of their eigenvalues. Subcones of this type will henceforth be called **common-eigenvector subcones** of the **PSD** cone. When working with $p \times p$ covariance matrices, the elements of any such subcone are what Flury (1988) [15, (pg.60)] calls covariance matrices of the third level of similarity.

Flury defined the *Common Principal Components model* as the assumption that a given set of covariance matrices share the same eigenvectors, that is, belong to a single common-eigenvector subcone of the **PSD** cone.

Since the Spectral Decomposition of a **p.s.d.** matrix with all different eigenvalues is unique (apart from sign-switching), any such **p.s.d.** matrix will belong to a single subcone of the type defined above. Only matrices with two or more equal eigenvalues may belong to different common-eigenvector subcones. This provides a characterization of the points of intersection of these subcones. In particular, the central ray is the intersection of *all* common-eigenvector subcones.

III.3.6 The subcone of diagonal matrices

Of particular interest is the subcone of the matrices whose common eigenvectors are $\{\mathrm{e}_i\}_{i=1}^m$, where e_i is the i-th vector in the canonical basis for $\mathbb{R}^m$, that is, the vector whose only non-zero element is a 1 in the i-th position.

The non-negative linear combinations of $\{\mathbf{E}_{\mathrm{e}_i}\}_{i=1}^m$ are the diagonal matrices with non-negative diagonal elements. We will therefore call this common-eigenvector subcone the **subcone of diagonal p.s.d. matrices**. The matrices of eigenvalues in the Spectral Decompositions of **p.s.d.** matrices belong to this subcone.

III.3.7 Rotations around the central ray

Two matrices $\mathbf{M}, \mathbf{N} \in \mathbb{M}_m$ are said to be **similar** if there exists a non-singular matrix $\mathbf{T} \in \mathbb{M}_m$ such that $\mathbf{M} = \mathbf{T}^{-1}\mathbf{N}\mathbf{T}$ (see, for example, Horn and Johnson (1986) [25, (pg.44)]). A case of particular interest for our purposes arises when we consider similarities defined by orthogonal matrices, that is, when $\mathbf{M} = \mathbf{A}'\mathbf{N}\mathbf{A}$ for some orthogonal matrix $\mathbf{A}$. These similarities are known as **orthogonal equivalences** ([25, (pg.73)]) or **orthogonal similarities**.

If $\mathbf{A}$ is a given $m \times m$ orthogonal matrix, we can define a mapping from the space $\mathbb{S}_m$ onto itself, given by:

$$\mathcal{M}_{\mathbf{A}}(\mathbf{V}) = \mathbf{A}'\mathbf{V}\mathbf{A} \qquad \forall\, \mathbf{V} \in \mathbb{S}_m \tag{III.3.8}$$

What is the effect of a mapping $\mathcal{M}_{\mathbf{A}}(\cdot)$ on $\mathbb{S}_m$ and the **PSD** cone, in particular? The next theorem deals with this issue. We recall that an **automorphism** on an inner product space is a one-to-one mapping of the linear space onto itself, which preserves inner products (see Halmos (1958) [22, (pg.143)]).

Theorem III.8 *Let $\mathbf{A} \in \mathbb{M}_{\mathrm{m}}$ be an orthogonal matrix and $\mathcal{M}_{\mathbf{A}}$ the similarity mapping (III.3.8) defined by $\mathbf{A}$. Then, the following statements hold:*

i). $\mathcal{M}_{\mathbf{A}}$ is an automorphism on the inner product space $\mathbb{S}_{\mathrm{m}}$.

ii). All matrices on the subspace spanned by the identity $\mathbf{I}_m$ are left unchanged by $\mathcal{M}_{\mathbf{A}}$.

iii). The common-eigenvector subcone spanned by the rank one idempotent matrices $\{\mathbf{E}_{\mathbf{b}_i}\}_{i=1}^{m}$ *is mapped onto the common-eigenvector subcone spanned by the rank one idempotent matrices* $\{\mathbf{E}_{\mathbf{A}'\mathbf{b}_i}\}_{i=1}^{m}$.

iv). $\mathcal{M}_{\mathbf{A}}$ *preserves eigenvalues.*

v). The **PSD** *cone is an invariant set under* $\mathcal{M}_{\mathbf{A}}$.

Proof.

i). We must prove that $\mathcal{M}_{\mathbf{A}}$ is an inner-product preserving, one-to-one mapping of $\mathbb{S}_m$ onto itself, for any orthogonal matrix $\mathbf{A}$. That $\mathcal{M}_{\mathbf{A}}$ is a bijection can be guaranteed by noting that, for any orthogonal matrix $\mathbf{A}$, the mapping $\mathcal{M}_{\mathbf{A}'}$ is a well-defined inverse mapping in $\mathbb{S}_m$. In fact, for any symmetric matrices $\mathbf{V}$ and $\mathbf{W}$, we have $\mathcal{M}_{\mathbf{A}}(\mathbf{V}) = \mathbf{W} \iff \mathbf{V} = \mathcal{M}_{\mathbf{A}'}(\mathbf{W})$. That the inner product is preserved is also easily proved. For any $\mathbf{V}, \mathbf{W} \in \mathbb{S}_m$ and any orthogonal matrix $\mathbf{A}$, we have:

$$< \mathcal{M}_{\mathbf{A}}(\mathbf{V}), \mathcal{M}_{\mathbf{A}}(\mathbf{W}) >= \mathrm{tr}(\mathbf{A}'\mathbf{V}\mathbf{A} \cdot \mathbf{A}'\mathbf{W}\mathbf{A}) = \mathrm{tr}(\mathbf{V}\mathbf{I}_m\mathbf{W}\mathbf{I}_m) =< \mathbf{V}, \mathbf{W} >$$

ii). Any matrix on $span(\mathbf{I}_m)$ is of the form $\alpha\mathbf{I}_m$, for some $\alpha \in \mathbb{R}$. Thus, $\mathcal{M}_{\mathbf{A}}(\alpha\mathbf{I}_m) = \alpha\mathbf{A}'\mathbf{I}_m\mathbf{A} = \alpha\mathbf{I}_m$.

iii). Any matrix on the subcone spanned by the matrices $\{\mathbf{E}_{\mathbf{b}_i}\}_{i=1}^{m}$ can be written as $\mathbf{V} = \sum_{i=1}^{m} \mu_i \mathbf{E}_{\mathbf{b}_i} = \sum_{i=1}^{m} \mu_i \mathbf{b}_i\mathbf{b}_i{}'$, for some set of scalars $\{\mu_i\}_{i=1}^{m}$. The image of

$\mathbf{V}$ under $\mathcal{M}_{\mathbf{A}}$ can therefore be written as $\mathbf{A}'\mathbf{V}\mathbf{A} = \sum_{i=1}^{m} \mu_i(\mathbf{A}'\mathbf{b}_i)(\mathbf{A}'\mathbf{b}_i)'$. The matrices $\mathbf{E}_{\mathbf{A}'\mathbf{b}_i} = (\mathbf{A}'\mathbf{b}_i)(\mathbf{A}'\mathbf{b}_i)'$ also define a subcone because they are a set of mutually orthogonal rank one idempotent matrices. In fact, given Theorem III.6, all that has to be shown is that $\{\mathbf{A}'\mathbf{b}_i\}_{i=1}^{m}$ is, like $\{\mathbf{b}_i\}_{i=1}^{m}$, an orthonormal set of vectors in $\mathbb{R}^m$. This can be easily checked, and is a well-known property of the ℓ_2-norm called *orthogonal invariance* (see also Shilov (1961) [61, (pg.148)]).

iv). $\mathcal{M}_{\mathbf{A}}$ is a particular case of a similarity relation and it is well-known that similarities are eigenvalue-preserving relations (see, for example, Horn and Johnson (1986) [25, (pg.45)]).

v). If $\mathbf{V} \in \mathbf{PSD}$, then $\mathbf{A}'\mathbf{V}\mathbf{A} \in \mathbf{PSD}$ because it is still a symmetric matrix with the same, non-negative, eigenvalues as $\mathbf{V}$.

□

Point i) of Theorem III.8 tells us that the angles and distances between any two matrices in $\mathbb{S}_m$ are the angles and distances between their images under $\mathcal{M}_{\mathbf{A}}$, for any orthogonal matrix $\mathbf{A}$. Coupled with the invariance property of point ii), this suggests a geometric description of the mappings $\mathcal{M}_{\mathbf{A}}$ as rigid rotations of $\mathbb{S}_m$ around the subspace $span(\mathbf{I}_m)$. Focusing on the $\mathbf{PSD}$ cone and considering also point v) of the Theorem, we can say that *the mappings $\mathcal{M}_{\mathbf{A}}$ induce rigid rotations of the* $\mathbf{PSD}$ *cone around its central ray.* Naturally, some specific orthogonal matrices $\mathbf{A}$ may define mappings which also leave other rays invariant.

III.3.8 Common eigenvalues

We can now give a geometric characterization of the sets of **p.s.d.** matrices with the same eigenvalues.

Point iv) of Theorem III.8 tells us that $\mathbf{V}, \mathbf{W} \in \mathbf{PSD}$ have the same eigenvalues if we can write $\mathbf{W} = \mathcal{M}_{\mathbf{A}}(\mathbf{V})$ for some orthogonal matrix $\mathbf{A}$. It can also be shown by a standard matrix argument that two **p.s.d.** matrices (or, more generally, symmetric matrices) share the same set of eigenvalues only if one of them is an image of the other under some appropriate orthogonal equivalence. In fact, if $\mathbf{V}, \mathbf{W} \in \mathbf{PSD}$ with spectral decompositions $\mathbf{V} = \mathbf{B}_1\mathbf{\Lambda}\mathbf{B}_1'$ and $\mathbf{W} = \mathbf{B}_2\mathbf{\Lambda}\mathbf{B}_2'$, then $\mathbf{W} = \mathcal{M}_{\mathbf{A}}(\mathbf{V})$ with $\mathbf{A} = \mathbf{B}_1\mathbf{B}_2'$. Matrix $\mathbf{A}$ is necessarily orthogonal since it is the product of two orthogonal matrices (see Horn and Johnson (1986) [25, (pg.68)]).

Our previous discussion therefore tells us that two **p.s.d.** matrices share the same set of eigenvalues if and only if there is some rigid rotation of the **PSD** cone around its central ray which can take one of the matrices to the position where the other originally stood.

III.3.9 The Spectral Decomposition and rigid rotations

What has been said so far also provides a geometric interpretation for the 'matrix version' of the Spectral Decomposition of a **p.s.d.** matrix. This is an alternative interpretation to the one given in Theorem III.7.

At the core of the 'matrix version' (I.3.2) of the Spectral Decomposition of a symmetric matrix **V** lies the notion that **V** and its matrix of eigenvalues $\mathbf{\Lambda}$ are orthogonally equivalent. The geometric implication of this, in the **PSD** cone, is that rigid rotations around the central ray, as defined by mappings of the type $\mathcal{M}_{\mathbf{A}}$, *map* **V** *onto the subcone of diagonal matrices* if and only if the columns of **A** form an orthonormal set of eigenvectors of **V** and the diagonal elements of the image $\mathcal{M}_{\mathbf{A}}(\mathbf{V})$ are the eigenvalues of **V**. It should be recalled that if **V**'s eigenvalues are all different, there will be m ! such rotations and images of **V** on the subcone of the diagonals (corresponding to the different orderings of the eigenpairs). For matrices with equal eigenvalues, the number of images on the subcone of diagonals decreases – to a single image (the matrix itself) when all eigenvalues are equal, since for any orthogonal matrix **A**, $\mathbf{A}(\alpha\mathbf{I}_m)\mathbf{A}' = \alpha\mathbf{A}\mathbf{A}' = \alpha\mathbf{I}_m$. In general, if an $m \times m$ **p.s.d.** matrix has k different eigenvalues with multiplicities $m_1, ..., m_k$, the number of its possible images on the subcone of diagonals is given by the multinomial coefficient (see Meyer (1970) [44, (pg.177)]):

$$\frac{m!}{m_1!m_2!\cdots m_k!}$$

However, the number of rotations producing these fewer images *increases* as there are now an infinity of possible eigenvectors from which to choose the columns of **A**.

III.3.10 PCA and the PSD cone

The discussion in this Section is directly relevant to PCA, since the fundamental concepts of PCA can be characterized in terms of the **PSD** cones in $\mathbb{S}_n$ and $\mathbb{S}_p$.

In particular, Theorem III.7 provides the basis for a geometric 'visualization' of the relation between the position of the matrix $\frac{1}{n}\mathbf{XX}'$ on $\mathbb{S}_n$'s **PSD** cone and the PCs of $\mathbf{X}$. Likewise, the Principal Axes in $\mathbb{R}^p$ of $\mathbf{X}$ can be characterized by considering the position of the covariance matrix in the **PSD** cone of $\mathbb{S}_p$.

Curiously, the ℓ_2-norm relative eigenvalues which are not normally considered in PCA but which appeared in the weights for the similarity indices in $\mathbb{R}^n$ and $\mathbb{R}^p$, also have an interesting geometric characterization, as was seen in Theorem III.7.

It should be pointed out that the results in this Section do not tell us anything fundamentally new about PCA. Rather, they are re-phrasing well-known results in a new context. In so doing, however, they lay the groundwork for new insights into PCA.

Future research along these lines looks promising. Among possible directions for such future work are the consideration of how certain types of linear transformations of the data – in particular the common standardization – affect the position of covariance matrices and the matrices $\frac{1}{n}\mathbf{XX}'$ in their respective **PSD** cones and whether it is possible to extend some of the above results to the space $\mathbb{M}_{n \times p}$.

Chapter IV

PCA and linear transformations of the data

Data sets are often pre-processed in some way prior to a statistical analysis. Column-centering a data matrix (equation (I.1.1)) and standardizing it (I.1.3) are examples of such transformations in the context of PCA.

Linear transformations of the data are commonly used in Principal Component Analysis although non-linear transformations are also in use, either as a generalization of PCA (see Gnanadesikan (1977) [17, (Subsection 2.4.2)]) or due to the nature of the problem being tackled (see, for example the discussion on the removal of isometric size from morphometric data in Chapter V).

Processing the data prior to a PCA will, in general, change the PCs, their 'natural lengths' and the Principal Axes in $\mathbb{R}^p$. In this Chapter we explore these effects.

The first Section provides background material on generalized inner products and generalized SVDs. Section IV.2 discusses how a PCA of a transformed data matrix can also be viewed as a change in the inner product with which the PCA is defined. Section IV.3 gives some results on how linear transformations of a data matrix affect its PCA. Finally, Sections IV.4 and IV.5 provide simple methods to compare the PCAs of different transformations of a given data matrix, both pairwise and for groups of t (t >2) transformations.

IV.1 Generalized inner products

IV.1.1 Basic definitions

Every possible inner product in the vector space $\mathbb{R}^m$ is defined by some positive definite $m \times m$ matrix $\mathbf{M}$ in the following way (Shilov (1961) [61, (pg.137)]):

$$< \mathbf{x}, \mathbf{y} >_M = \mathbf{x}'\mathbf{M}\mathbf{y} \ , \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^m \qquad \text{(IV.1.1)}$$

The inner product $< \cdot, \cdot >_M$ shall be called the **M-inner product** in $\mathbb{R}^m$.

We can thus speak of **M-orthogonal vectors** $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ when $< \mathbf{x}, \mathbf{y} >_M = 0$. Likewise, the **M-orthogonal complement** of a subspace $\mathcal{M}$ of $\mathbb{R}^m$ is the subspace of all vectors of $\mathbb{R}^m$ which are M-orthogonal to every vector in $\mathcal{M}$. The inner product (IV.1.1) induces a norm and a distance (or metric) in $\mathbb{R}^m$, in the standard way. The **M-norm** is defined as:

$$\|\mathbf{x}\|_M = \sqrt{< \mathbf{x}, \mathbf{x} >_M} \qquad \forall \mathbf{x} \in \mathbb{R}^m$$

The **M-metric** is defined as:

$$d_M(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_M \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^m$$

M-angles in $\mathbb{R}^m$ can also be defined using Definition A.5:

$$\begin{aligned} \sphericalangle_M(\mathbf{x}, \mathbf{y}) &= \arccos\left(\frac{< \mathbf{x}, \mathbf{y} >_M}{\|\mathbf{x}\|_M \cdot \|\mathbf{y}\|_M}\right) & \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^m - \{0\} \\ \sphericalangle_M(\mathbf{x}, \mathbf{0}) &= 0 & \forall \mathbf{x} \in \mathbb{R}^m \end{aligned}$$

The Orthogonal Projection Theorem (Proposition A.16) is valid for any inner product linear space, hence for $\mathbb{R}^m$ with the **M**-inner product. However, the results in Proposition A.19, which assume that the inner product in $\mathbb{R}^m$ is the $\mathbf{I}_m$-inner product, must be adapted accordingly. The **M-orthogonal projection** of any vector $\mathbf{x} \in \mathcal{L}$ onto a subspace $\mathcal{M}$ must be the vector of $\mathcal{M}$ which is closest *in the* **M**-*metric* to $\mathbf{x}$.

The general form of the **matrices of M-orthogonal projection** can be found in Basilevski (1983) [4, (pg.165)] or Rao (1980) [56, (pg.12)]. It is freely reproduced in the following theorem.

Theorem IV.1 (M-orthogonal projection matrices) *Let* **M** *be an* m×m *positive definite matrix. Let* $\mathcal{M}$ *be a* k-dimensional *subspace of* $\mathbb{R}^m$. *Let* **A** *be an* m×k *matrix whose columns form a basis for the subspace $\mathcal{M}$. Then* $\mathbf{P}$ *is an* $\mathbf{M}$*-orthogonal projector onto $\mathcal{M}$ if and only if it can be written as:*

$$\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{M}\mathbf{A})^{-1}\mathbf{A}'\mathbf{M} \qquad \text{(IV.1.2)}$$

An intuitive feeling for the decomposition (IV.1.2) is obtained by noting that any vector of $\mathcal{M}$ is a linear combination of the columns of $\mathbf{A}$, *i.e.*, can be written as $\mathbf{A}\mathbf{a}$ for some vector $\mathbf{a} \in \mathbb{R}^k$. We then have $\mathbf{P}\mathbf{A}\mathbf{a} = \mathbf{A}\mathbf{a}$, $\forall \mathbf{a} \in \mathbb{R}^k$, so that such vectors of $\mathcal{M}$ are mapped onto themselves by $\mathbf{P}$. At the same time, any vector $\mathbf{z}$ in the $\mathbf{M}$-orthogonal complement of $\Re(\mathbf{A})$ is annihilated by $\mathbf{P}$, since it must be $\mathbf{M}$-orthogonal to all columns of $\mathbf{A}$, thus implying that $\mathbf{A}'\mathbf{M}z = 0$.

As with Proposition A.19, the above result implicitly assumes that $\mathbf{P}$ is *pre-multiplying* the vectors of $\mathbb{R}^m$ which are to be projected.

It is important to note that the matrix $\mathbf{P}$ in (IV.1.2) is of the form $\mathbf{A}(\mathbf{B}'\mathbf{A})^{-1}\mathbf{B}'$, with $\mathbf{B} = \mathbf{M}\mathbf{A}$. Proposition A.19 thus tells us that we can always view $\mathbf{M}$-orthogonal projectors in $\mathbb{R}^m$ as non-orthogonal (*oblique*) projectors with the usual $\mathbf{I}_m$-inner product. The linear independence of $\mathbf{B}$'s columns (required in Proposition A.19) is guaranteed because $\mathbf{B}\alpha = \mathbf{M}\mathbf{A}\alpha = 0$ implies $\mathbf{A}\alpha = 0$ (since $\mathbf{M}$ is invertible) which, in turn, implies $\alpha = 0$ since $\mathbf{A}$'s columns are linearly independent. Hence, *an* $\mathbf{M}$*-orthogonal projection onto* $\mathcal{A} = \Re(\mathbf{A})$ *is a projection onto* $\mathcal{A}$ *along* $\mathcal{B} = \Re(\mathbf{M}\mathbf{A})^{\perp}$ *with the usual inner product.*

A generalization of the Frobenius inner product between matrices in $\mathbb{M}_{m\times q}$ may also be defined. If $\mathbf{N}$ is the inner product matrix in $\mathbb{R}^m$ and $\mathbf{M}$ the inner product matrix in $\mathbb{R}^q$, the natural generalization of (III.1.1) is (see Ramsay *et al.* (1984) [52, (pg.405)]) the **(N,M)-inner product in** $\mathbb{M}_{m\times q}$:

$$< \mathbf{A}, \mathbf{B} >_{(N,M)} = \mathrm{tr}(\mathbf{A}'\mathbf{N}\mathbf{B}\mathbf{M}) \quad \forall \mathbf{A}, \mathbf{B} \in \mathbb{M}_{m\times q} \qquad \text{(IV.1.3)}$$

This generalization is a natural one because $\mathbf{A}'\mathbf{N}\mathbf{B}$ gives the $\mathbf{N}$-inner products of the columns of $\mathbf{A}$ and $\mathbf{B}$ (which are vectors of $\mathbb{R}^m$) and $\mathbf{B}\mathbf{M}\mathbf{A}'$ gives the $\mathbf{M}$-inner products of the rows of $\mathbf{A}$ and $\mathbf{B}$ (which are vectors in $\mathbb{R}^q$).

In general, the use of non-standard inner products and their consequences are somewhat difficult in that our intuitive notions of distance and angle are changed. But the advantages, both conceptual and notational, can sometimes override this difficulty. The 'statistically natural' inner product in $\mathbb{R}^n$ (defined by the matrix $\frac{1}{n}\mathbf{I}_n$) is a simple example of this. In this Chapter we shall consider other such cases.

IV.1.2 The generalized SVD

In Subsection I.3.2, the Singular Value Decomposition of a generic matrix $\mathbf{Y} \in \mathbb{M}_{n\times p}$ was introduced. In its matrix version (equation (I.3.6)), it tells us that any rank r matrix $\mathbf{Y}$ can be written as $\mathbf{Y} = \boldsymbol{\alpha}\boldsymbol{\Sigma}\mathbf{A}'$ where $\boldsymbol{\alpha}$ is an $n\times r$ matrix with $\mathbf{I}_n$-orthonormal columns, $\mathbf{A}$ is a $p\times r$ matrix with $\mathbf{I}_p$-orthonormal columns and $\boldsymbol{\Sigma}$ is an $r\times r$ diagonal matrix with positive diagonal elements.

It is possible to generalize the SVD so as to replace the $\mathbf{I}_m$-orthonormality requirements (with $m = n, p$) with similar requirements for alternative inner products. The next Theorem is stated and proved by, among others, Ramsay *et al.* (1984) [52] and Greenacre (1984) [21], but its important role (and simple proof!) justify a closer look at it.

Theorem IV.2 (Generalized SVD) *Let* $\mathbf{Y} \in \mathbb{M}_{\mathrm{nxp}}$ *be a generic matrix of rank* r. *Let* $\mathbf{M}$ *be a positive definite matrix in* $\mathbb{M}_{\mathrm{p}}$ *and* $\mathbf{N}$ *a positive definite matrix in* $\mathbb{M}_{\mathrm{n}}$. *Then, it is possible to write:*

$$\mathbf{Y} = \boldsymbol{\beta}\mathcal{T}\mathbf{B}' \qquad \text{(IV.1.4)}$$

where $\mathbf{B} \in \mathbb{M}_{\mathrm{pxr}}$ *and* $\boldsymbol{\beta} \in \mathbb{M}_{\mathrm{nxr}}$, *such that* $\mathbf{B}'\mathbf{M}\mathbf{B} = \boldsymbol{\beta}'\mathbf{N}\boldsymbol{\beta} = \mathbf{I}_r$ *and* $\mathcal{T} \in \mathbb{M}_{\mathrm{r}}$ *is a diagonal matrix with positive diagonal elements.*

Proof. It is always possible to decompose the positive definite matrices $\mathbf{M}$,$\mathbf{N}$ into products of the form $\mathbf{M} = \mathbf{U}'\mathbf{U}$ and $\mathbf{N} = \mathbf{T}'\mathbf{T}$, for some non-singular matrices $\mathbf{U} \in \mathbb{M}_p$, $\mathbf{T} \in \mathbb{M}_n$. In fact, taking $\mathbf{U}$ and $\mathbf{T}$ to be the positive definite square roots of $\mathbf{M}$ and $\mathbf{N}$ (as defined by (I.3.4)) provides one such factorization. Then, if a conventional SVD of the matrix $\mathbf{TYU}'$ is given by $\mathbf{TYU}' = \boldsymbol{\alpha}\Sigma\mathbf{A}'$, a generalized SVD of $\mathbf{Y}$ is obtained by taking $\mathcal{T} = \Sigma$, $\boldsymbol{\beta} = \mathbf{T}^{-1}\boldsymbol{\alpha}$ and $\mathbf{B} = \mathbf{U}^{-1}\mathbf{A}$. □

Equation (IV.1.4) can also be written as:

$$\mathbf{Y} = \sum_{i=1}^{r} \tau_i \boldsymbol{\beta}_i \mathbf{b}_i{}' \qquad \text{(IV.1.5)}$$

where $\boldsymbol{\beta}_i'\mathbf{N}\boldsymbol{\beta}_j = \mathbf{b}_i'\mathbf{M}\mathbf{b}_j = \delta_{ij}$.

Theorem IV.2 is an *existence* theorem. The discussion on the *uniqueness* of decomposition (IV.1.4) is analogous to the discussion on the uniqueness of the conventional SVD of matrix $\mathbf{TYU}'$, with the extra possibility of non-uniqueness arising from the factorizations $\mathbf{M} = \mathbf{U}'\mathbf{U}$ and $\mathbf{N} = \mathbf{T}'\mathbf{T}$. However, it can be shown that the latter is *not* an additional source of non-uniqueness. Once again, the short but important proof of that fact deserves some attention. The key role in this respect is played by the following Theorem (see, for example, Horn and Johnson (1985) [25, (pg.406)]):

Theorem IV.3 *If* $\mathbf{M}$*is an* m×m *positive definite matrix and if* $\mathbf{M} = \mathbf{U}_1'\mathbf{U}_1 = \mathbf{U}_2'\mathbf{U}_2$, *with* $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{M}_{\mathrm{m}}$, *then* $\mathbf{U}_2 = \mathbf{R}\mathbf{U}_1$ *for some orthogonal matrix* $\mathbf{R}$.

Proof. Matrices $\mathbf{U}_1, \mathbf{U}_2$ must be invertible since $m = rank(\mathbf{U}_k'\mathbf{U}_k) = rank(\mathbf{U}_k)$, $(k = 1, 2)$ (see point (iv) in the characterization of PCs in Subsection I.4.1). Thus, $\mathbf{M} = \mathbf{U}_1'\mathbf{U}_1 = \mathbf{U}_2'\mathbf{U}_2$ implies that $(\mathbf{U}_2\mathbf{U}_1^{-1})'(\mathbf{U}_2\mathbf{U}_1^{-1}) = \mathbf{I}_m$. In other words, $\mathbf{R} = \mathbf{U}_2\mathbf{U}_1^{-1}$ is an orthogonal matrix and $\mathbf{U}_2 = \mathbf{R}\mathbf{U}_1$. □

How will such alternative decompositions of $\mathbf{M}$ and $\mathbf{N}$ affect Theorem IV.2 and its proof? Say $\mathbf{M} = \mathbf{U}_1'\mathbf{U}_1 = \mathbf{U}_2'\mathbf{U}_2$ and $\mathbf{N} = \mathbf{T}_1'\mathbf{T}_1 = \mathbf{T}_2'\mathbf{T}_2$, and that an SVD of $\mathbf{T}_1\mathbf{Y}\mathbf{U}_1{}'$ is given by $\mathbf{T}_1\mathbf{Y}\mathbf{U}_1{}' = \boldsymbol{\alpha}_1\boldsymbol{\Sigma}_1\mathbf{A}_1'$. Then, assuming $\mathbf{U}_2 = \mathbf{R}\mathbf{U}_1$ and $\mathbf{T}_2 = \mathbf{Q}\mathbf{T}_1$ for orthogonal matrices $\mathbf{R}$,$\mathbf{Q}$, we have $\mathbf{T}_2\mathbf{Y}\mathbf{U}_2{}' = (\mathbf{Q}\boldsymbol{\alpha}_1)\boldsymbol{\Sigma}_1(\mathbf{R}\mathbf{A}_1)'$. But since $\mathbf{Q}\boldsymbol{\alpha}_1$ and $\mathbf{R}\mathbf{A}_1$ still have orthonormal columns and since $\boldsymbol{\Sigma}_1$ is a diagonal matrix with positive diagonal elements, then the above equation is an SVD of $\mathbf{T}_2\mathbf{Y}\mathbf{U}_2{}'$. Now, the matrices $\boldsymbol{\beta}$ and $\mathbf{B}$ that would result from using this SVD of $\mathbf{T}_2\mathbf{Y}\mathbf{U}_2'$ are:

$$\begin{aligned}
\boldsymbol{\beta} = \mathbf{T}_2{}^{-1}(\mathbf{Q}\boldsymbol{\alpha}_1) &= \mathbf{T}_1{}^{-1}(\mathbf{Q}'\mathbf{Q})\boldsymbol{\alpha}_1 = \mathbf{T}_1{}^{-1}\boldsymbol{\alpha}_1 \\
\mathbf{B} = \mathbf{U}_2{}^{-1}(\mathbf{R}\mathbf{A}_1) &= \mathbf{U}_1{}^{-1}(\mathbf{R}'\mathbf{R})\mathbf{A}_1 = \mathbf{U}_1{}^{-1}\mathbf{A}_1
\end{aligned}$$

In other words, the generalized Singular Value Decomposition (IV.1.4) is determined by $\mathbf{M}$ and $\mathbf{N}$, but not by the precise factorizations of these matrices which are used in the proof.

The decomposition (IV.1.4) is therefore unique, apart from:

i). possible re-orderings of the columns of $\mathbf{B}, \boldsymbol{\mathcal{T}}$ and $\boldsymbol{\beta}$ (not relevant for (IV.1.5)).

ii). simultaneous sign-switching in corresponding columns of $\mathbf{B}$ and $\boldsymbol{\beta}$.

iii). equal diagonal elements in $\boldsymbol{\Sigma}$.

Ramsay *et al.* (1984) [52] speak of (IV.1.4) as an **(N,M)-orthogonal SVD of Y**. The columns of matrices $\boldsymbol{\beta}$, $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are known, respectively, as the **generalized left singular vectors**, the **generalized right singular vectors** and the **generalized singular values** of $\mathbf{Y}$ **for the inner products** $\mathbf{N}$ **and** $\mathbf{M}$.

This rather lengthy discussion of the (**N**,**M**)-orthogonal SVD of $\mathbf{Y}$ will prove useful shortly. For the moment, we shall merely note that in Subsection I.4.3 we had seen how a PCA of a column-centered data matrix $\mathbf{X}$ could be described in terms of the conventional SVD of the matrix $\frac{1}{\sqrt{n}}\mathbf{X}$. From the proof of Theorem IV.2 it should now be apparent that we could also speak of an $(\frac{1}{n}\mathbf{I}_n, \mathbf{I}_p)$-orthogonal SVD of $\mathbf{X}$. Once again, the 'statistically natural' inner product in $\mathbb{R}^n$ — corresponding to the matrix $\frac{1}{n}\mathbf{I}_n$ — proves useful.

IV.2 Linear transformations and PCA

Every linear transformation in the vector space $\mathbb{R}^m$ is represented (for a given basis of $\mathbb{R}^m$) by a matrix in $\mathbb{M}_m$.

We have seen how the p columns of a data matrix $\mathbf{Y} \in \mathbb{M}_{n \times p}$ can be viewed as points/vectors in $\mathbb{R}^n$. A linear transformation in $\mathbb{R}^n$, given by the matrix $\mathbf{T} \in \mathbb{M}_n$, will transform those p vectors into the columns of matrix $\mathbf{TY}$. Alternatively, (see Halmos (1958) [22, (pg.83)]), if $\mathbf{T}$ is invertible, we can think of the columns of $\mathbf{TY}$ as the new coordinates for the same p vectors after a transformation of the coordinate system. Similarly, the rows of $\mathbf{YU}'$ reflect the effects on the n points/vectors in $\mathbb{R}^p$ of the linear transformation given by the matrix $\mathbf{U} \in \mathbb{M}_p$ (using the general convention that the linear transformation described by $\mathbf{U}$ is the effect of it *pre*-multiplying the vectors of $\mathbb{R}^p$).

We have encountered examples of such transformations in Section I.1. Column-centering a data matrix $\mathbf{Y}$ was achieved in (I.1.1) by taking $\mathbf{T}$ to be the orthogonal projection matrix $\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n{}'$. Dividing all elements of a matrix by $\frac{1}{\sqrt{n}}$ (as in (I.1.2)) can be achieved by either taking $\mathbf{T} = \frac{1}{\sqrt{n}}\mathbf{I}_n$ or $\mathbf{U} = \frac{1}{\sqrt{n}}\mathbf{I}_p$. Standardization (I.1.3) is the result of taking $\mathbf{U}$ to be the diagonal matrix of the reciprocal standard deviations.

It should be noted that $\mathbf{TY}$ also affects the n points/vectors in $\mathbb{R}^p$ (rows of $\mathbf{Y}$) and, conversely, $\mathbf{YU}'$ also affects the p points/vectors in $\mathbb{R}^n$. In some cases, the effects of a linear transformation in one space are best discussed in terms of its effects on the representation of the data in the other space. For example, any diagonal matrix $\mathbf{U} \in \mathbb{R}^p$ has the effect of multiplying each *column* of $\mathbf{Y}$ by the corresponding diagonal element of $\mathbf{U}$, as we saw when standardizing the data. Diagonal matrices $\mathbf{U} \in \mathbb{R}^p$ are of particular importance for our purposes, since they can be viewed as describing the effects of separately changing the scales of measurement for each of the p variables.

Whilst the columns of matrix $\mathbf{YU}'$ will always lie in the subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{Y}$, for any matrix $\mathbf{U} \in \mathbb{R}^p$, the same need not be the case for the columns of the matrix $\mathbf{TY}$. If it *is* the case that the columns of $\mathbf{TY}$ lie in $\Re(\mathbf{Y})$, then we can always write $\mathbf{TY} = \mathbf{YW}'$, where $\mathbf{W} \in \mathbb{R}^p$ is given by $\mathbf{W} = (\mathbf{Y}^-\mathbf{TY})'$ (see I.3.9).

We now consider the problem of the relations between a PCA of a data matrix $\mathbf{Y}$ and that of a transformation of the kind:

$$\mathbf{Y} \longrightarrow \mathbf{YU}' \tag{IV.2.1}$$

for some non-singular $\mathbf{U} \in \mathbb{M}_p$.

In Subsection I.4.3 we saw that the essential entities in the PCA of matrix $\mathbf{Y}$ were given in a (conventional) SVD of the matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ with $\mathbf{X} = \mathbf{P}_{\mathbf{1}_n}\mathbf{Y}$. The PCA of

$\mathbf{YU}'$ will be summarized by the SVD of $\frac{1}{\sqrt{n}}\mathbf{XU}'$. Let us assume that these SVDs are given by:

$$\begin{aligned} \tfrac{1}{\sqrt{n}}\mathbf{X} &= \boldsymbol{\alpha}\boldsymbol{\Sigma}\mathbf{A}' \\ \tfrac{1}{\sqrt{n}}\mathbf{XU}' &= \boldsymbol{\alpha}_U\boldsymbol{\Sigma}_U\mathbf{A}_U' \end{aligned} \qquad \text{(IV.2.2)}$$

In general, the corresponding matrices in both decompositions are different.

The columns of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_U$ are two orthonormal bases for the subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}$ (which is also the subspace spanned by the columns of $\mathbf{XU}'$). The columns of $\mathbf{A}$ and $\mathbf{A}_U$ are two orthonormal bases for $\mathbb{R}^p$. In that sense, both sets of Principal Axes in each space can be directly compared, to assess the effects of the transformation $\mathbf{U}$.

But, in $\mathbb{R}^p$, another comparison can also be envisaged. The Principal Axes in $\mathbb{R}^p$ of $\mathbf{X}$ can be compared with those of $\mathbf{XU}'$ *when re-expressed in terms of the original coordinate system*, that is, 'undoing' the effect of the transformation $\mathbf{U}$.

This implies that, rather than using matrix $\mathbf{A}_U$ of (IV.2.2), we use matrix $\mathbf{U}^{-1}\mathbf{A}_U$ when comparing with the original Principal Axes (PAs) in $\mathbb{R}^p$, given by the columns of $\mathbf{A}$. In other words, we use the $\mathbf{M}$-orthonormal right singular vectors of $\mathbf{X}$, with $\mathbf{M}$ given by $\mathbf{M} = \mathbf{U}'\mathbf{U}$.

What has just been said is that *for invertible matrices* $\mathbf{U} \in \mathbb{M}_\mathrm{p}$, *the results of a conventional PCA of* $\mathbf{XU}'$ *with the PAs in* $\mathbb{R}^p$ *expressed in the original coordinate system for* $\mathbb{R}^p$ *are given by the* $(\frac{1}{n}\mathbf{I}_n, \mathbf{U}'\mathbf{U})$*-orthogonal SVD of* $\mathbf{X}$ (Theorem IV.2).

In this case, we speak of an **M-inner product PCA** of data matrix $\mathbf{X}$. In much of the literature on PCA, in particular in the French-language literature, the term **M-metric PCA** is also used (Cailliez and Pagés (1976) [7, (Chapter VIII)]).

The vectors $\mathbf{A}_M = \mathbf{U}^{-1}\mathbf{A}_U$ will be called the **M-orthogonal Principal Axes in** $\mathbb{R}^p$ of the matrix $\mathbf{X}$. They are *eigenvectors of the matrix* $\mathbf{SM}$, where $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ is the covariance matrix determined by $\mathbf{X}$. At the same time, the eigenvalues of $\mathbf{USU}'$, *i.e.*, the diagonal elements of $\mathbf{\Sigma}_U^2$ in (IV.2.2) are the *eigenvalues of* $\mathbf{SM}$ (see also Escoufier (1987) [13]).

It should be stressed that the PAs in $\mathbb{R}^p$ of $\mathbf{XU}'$ *without* reversing the linear transformation, will depend not just on $\mathbf{M}$ but also on the specific factorization $\mathbf{M} = \mathbf{U}'\mathbf{U}$ which is being considered. We are better off speaking about a (conventional) PCA of $\mathbf{XU}'$ if our main concern is with a set of $\mathbf{I}_p$-orthonormal axes in $\mathbb{R}^p$.

A simple example which highlights this distinction arises from considering any orthogonal transformation matrix $\mathbf{U}$. The Principal Axes in $\mathbb{R}^n$ (PCs) of $\mathbf{X}$ and $\mathbf{XU}'$ are identical, as are their 'natural lengths'. This can be seen by noting that if $\boldsymbol{\alpha}\mathbf{\Sigma}\mathbf{A}'$ is an SVD of $\frac{1}{\sqrt{n}}\mathbf{X}$, then $\boldsymbol{\alpha}\mathbf{\Sigma}(\mathbf{UA})'$ is an SVD of $\frac{1}{\sqrt{n}}\mathbf{XU}'$. The PAs in $\mathbb{R}^p$ of $\mathbf{XU}'$ are a rigid rotation of those of $\mathbf{X}$, determined by $\mathbf{U}$. But when re-expressed in the original coordinates, the two sets of PAs in $\mathbb{R}^p$ coincide, since $\mathbf{U}^{-1}(\mathbf{UA}) = \mathbf{A}$.

The above discussion assumed the invertibility of $\mathbf{U}$. Singular matrices $\mathbf{U} \in \mathbb{M}_p$ cannot produce a positive definite matrix $\mathbf{M} = \mathbf{U}'\mathbf{U}$ (again because rank$(\mathbf{U}'\mathbf{U}) =$

rank($\mathbf{U}$)). Then $\mathbf{U}'\mathbf{U}$ will not define a proper inner product. A further discussion of such cases is deferred to the next Chapter.

The point that has just been raised is also behind our previous restriction in considering only linear transformations of the type $\mathbf{Y} \longrightarrow \mathbf{YU}'$.

Whatever the linear transformation $\mathbf{T}$ in $\mathbb{R}^n$ which maps $\mathbf{Y}$ to $\mathbf{TY}$, the column-centering requirement of PCA implies that we should consider the (conventional) SVD of $\frac{1}{\sqrt{n}}\mathbf{P}_{\mathbf{1}_n}\mathbf{TY}$. If the columns of $\mathbf{TY}$ are in $\Re(\mathbf{Y})$ we can, as we saw above, write $\mathbf{TY} = \mathbf{YW}'$ for some matrix $\mathbf{W} \in \mathbb{R}^p$ and fit such cases in the above discussion. But for more general transformations $\mathbf{T}$, the $n \times n$ matrix $\mathbf{P}_{\mathbf{1}_n}\mathbf{T}$ is always singular, since the rank of $\mathbf{P}_{\mathbf{1}_n}$ is n-1, and then the above discussion will not apply (unless improper inner products were brought in). If, on the other hand, the invertible transformation $\mathbf{T}$ is applied to the matrix $\mathbf{Y}$ *after* column-centering, that is, if we consider the SVD of matrix $\frac{1}{\sqrt{n}}\mathbf{TP}_{\mathbf{1}_n}\mathbf{Y}$, then we could once again frame our discussion in terms of alternative inner products. But then $\mathbf{TX}$ is not a column-centered matrix (except when $\mathbf{T}$ and $\mathbf{P}_{\mathbf{1}_n}$ commute, *i.e.*, when all row and column sums of $\mathbf{T}$ coincide, as happens when $\mathbf{T}$ is a scalar multiple of $\mathbf{I}_n$) and so the requirements for a conventional PCA are not met.

For the remainder of this Chapter we shall focus only on linear transformations of the form $\mathbf{Y} \longrightarrow \mathbf{YU}'$.

IV.3 Effects of linear transformations of the data

Linear transformations of the data can produce some quite dramatic changes in PCA. In general, all key components of a data set's PCA (PCs, their 'natural lengths', PAs in $\mathbb{R}^p$) undergo changes. No general explicit relation between the 'old' and the 'new' vectors and values is known. This is rather unfortunate, given the widespread use of linear transformations of the data, as in the case of Correlation Matrix PCA.

In this Section we explore some partial results relating the PCAs of a data matrix and some linear transformation of it, which may assist in understanding how transforming the data affects its PCA. The results involving the PAs in $\mathbb{R}^p$ will be presented in either of their two forms (with or without a reversal of the linear transformation), whichever proves most convenient. This is not a severe limitation since we can always move from one to the other, at will.

IV.3.1 (Almost) anything can happen

The knowledge that linear transformations of a data matrix $\mathbf{X}$ will, in general, affect all key aspects of its PCA inevitably raises the question of whether, by a suitable choice of transformation, we can reach *every* possible set of PCs, PAs in $\mathbb{R}^p$ and variances accounted for by each PC, from the same starting point $\mathbf{X}$.

An obvious limitation results from the fact that, for any invertible matrix $\mathbf{U} \in \mathbb{M}_p$, the columns of $\mathbf{X}$ and $\mathbf{XU}'$ will span the same subspace of $\mathbb{R}^n$. Since the left singular vectors in both matrices' SVDs must be orthonormal bases for this common subspace, at least this restriction exists on the possible results of a PCA of $\mathbf{XU}'$. Implicit in this restriction is also the requirement that the rank of $\mathbf{X}$ and of $\mathbf{XU}'$ be the same (*i.e.*, the dimension of the subspace spanned by their columns). In fact, if $\mathbf{U}$ is invertible, the rank of $\mathbf{X}$ and of $\mathbf{XU}'$ must be equal (see Horn and Johnson (1985) [25, (pg.13)]). But apart from this (two-fold) restriction, any prescribed SVD for $\frac{1}{\sqrt{n}}\mathbf{XU}'$ can be met with a suitable choice of transformation matrix $\mathbf{U}$. This result is formally proved below, where for simplicity the rank of $\mathbf{X}$ is assumed to be p.

Theorem IV.4 *Let* $\mathbf{W}$ *be an* $\mathrm{n}\times\mathrm{p}$ *matrix of rank* p . *Let an SVD of* $\mathbf{W}$ *be given by* $\boldsymbol{\alpha}\Sigma\mathbf{A}'$. *Let* $\mathbf{B}$ *be any* $\mathrm{p}\times\mathrm{p}$ *orthogonal matrix,* $\boldsymbol{\tau}$ *be any* $\mathrm{p}\times\mathrm{p}$ *diagonal matrix with positive diagonal elements and* $\boldsymbol{\beta}$ *an* $\mathrm{n}\times\mathrm{p}$ *matrix whose orthonormal columns span the same subspace of* $\mathbb{R}^n$ *as do the columns of* $\boldsymbol{\alpha}$ *(i.e., of* $\mathbf{W}$*). Then the matrix* $\mathbf{WU}'$ *has an SVD* $\boldsymbol{\beta}\boldsymbol{\tau}\mathbf{B}'$ *if and only if* $\mathbf{U}$ *is an invertible matrix which can be written as:*

$$\mathbf{U} = \mathbf{B}\boldsymbol{\tau}\boldsymbol{\beta}'\boldsymbol{\alpha}\Sigma^{-1}\mathbf{A}' \qquad \text{(IV.3.1)}$$

Proof. If $\mathbf{U}$ is given by (IV.3.1) and $\mathbf{W} = \boldsymbol{\alpha}\Sigma\mathbf{A}'$, we have:

$$\mathbf{WU}' = \boldsymbol{\alpha}\Sigma\mathbf{A}'\mathbf{A}\Sigma^{-1}\boldsymbol{\alpha}'\boldsymbol{\beta}\boldsymbol{\tau}\mathbf{B}' = \boldsymbol{\alpha}\boldsymbol{\alpha}'\boldsymbol{\beta}\boldsymbol{\tau}\mathbf{B}'$$

But the $n\times n$ matrix $\boldsymbol{\alpha}\boldsymbol{\alpha}'$ is the matrix of orthogonal projections on the p-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{\alpha}$ (Proposition A.19). Since the columns of $\boldsymbol{\beta}$ are assumed to lie in that subspace, they are unaltered by that projection.

Hence, $\boldsymbol{\alpha}\boldsymbol{\alpha}'\boldsymbol{\beta} = \boldsymbol{\beta}$ and $\mathbf{W}\mathbf{U}' = \boldsymbol{\beta}\boldsymbol{\mathcal{T}}\mathbf{B}'$ as required. On the other hand, if $\mathbf{W} = \boldsymbol{\alpha}\boldsymbol{\Sigma}\mathbf{A}'$, $\mathbf{W}^{-}$ is the Moore-Penrose generalized inverse of $\mathbf{W}$ and $\mathbf{W}\mathbf{U}' = \boldsymbol{\beta}\boldsymbol{\mathcal{T}}\mathbf{B}'$, we have $\mathbf{U}' = \mathbf{W}^{-}\mathbf{W}\mathbf{U}'$, since $\mathbf{W}^{-}\mathbf{W} = \mathbf{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}'\boldsymbol{\alpha}\boldsymbol{\Sigma}\mathbf{A}' = \mathbf{A}\mathbf{A}' = \mathbf{I}_p$ (because $\mathbf{A}$ is of full rank – see also the comments following (I.3.8)) and so $\mathbf{U}' = \mathbf{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}'\boldsymbol{\beta}\boldsymbol{\mathcal{T}}\mathbf{B}'$. $\mathbf{U}$ is invertible because the $p{\times}p$ matrices $\mathbf{A}, \mathbf{B}, \boldsymbol{\mathcal{T}}$ and $\boldsymbol{\Sigma}^{-1}$ are invertible, as is the $p{\times}p$ matrix $\boldsymbol{\beta}'\boldsymbol{\alpha}$. The latter's inverse is $\boldsymbol{\alpha}'\boldsymbol{\beta}$ — as follows from the comments above on $\boldsymbol{\alpha}\boldsymbol{\alpha}'\boldsymbol{\beta}$. Thus, $\boldsymbol{\beta}'\boldsymbol{\alpha}$, which is the matrix of cosines of the angles between the PCs of $\mathbf{X}$ and $\mathbf{X}\mathbf{U}'$ is actually an orthogonal matrix. □

This Theorem is merely stating the fact that $\mathbf{W}$ can be transformed into any other $n{\times}p$ column-centered data matrix $\mathbf{Z}$ whose columns span the same subspace of $\mathbb{R}^n$ as do the columns of $\mathbf{W}$. The theorem was worded to focus attention on the SVDs of $\mathbf{W}$ and $\mathbf{Z}$. The following Corollary expresses the result in terms of the matrices $\mathbf{W}$ and $\mathbf{Z}$ themselves.

Corollary *Let* $\mathbf{W}$ *be an* $\mathrm{n}{\times}\mathrm{p}$ *matrix of rank* p . *Let* $\mathbf{Z}$ *be any other* $\mathrm{n}{\times}\mathrm{p}$ *matrix whose* p *columns span the same subspace of* $\mathbb{R}^n$ *as do those of* $\mathbf{W}$. *Then, we can write* $\mathbf{Z} = \mathbf{W}\mathbf{U}'$. *The matrix* $\mathbf{U}$ *is given by* $\mathbf{U} = (\mathbf{W}^{-}\mathbf{Z})'$, *where* $\mathbf{W}^{-}$ *is the Moore-Penrose generalized inverse of* $\mathbf{W}$.

Proof. Definition I.1 defines the (unique) Moore-Penrose generalized inverse of *any* matrix with SVD $\boldsymbol{\alpha}\boldsymbol{\Sigma}\mathbf{A}'$ as the matrix $\mathbf{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}'$. The result follows directly from the Theorem. □

By taking $\mathbf{W} = \frac{1}{\sqrt{n}}\mathbf{X}$, where $\mathbf{X}$ is any column-centered data matrix, the above results become directly meaningful for the PCA of $\mathbf{X}$. The crucial point being made by the above Theorem and which has not been sufficiently stressed in the literature on PCA, is that *the PCA of* $\mathbf{X}$ *can be transformed in any way we wish* (with the restriction on the subspace spanned by the PCs) for an appropriate choice of invertible transformation matrix $\mathbf{U} \in \mathbb{M}_p$. *Any* new set of PAs in $\mathbb{R}^p$ can coexist with *any* new set of PAs in $\mathbb{R}^n$ (unit-norm PCs) and with *any* new set of proportions of variance accounted for.

Clearly, the use of transformation matrices must be governed by considerations other than obtaining a 'convenient' result.

IV.3.2 Invariant transformations

Some types of transformation matrices leave some aspects of a PCA invariant. A well-known example was mentioned in Section IV.2: if $\mathbf{U}$ is orthogonal, the PCs and their 'natural lengths' are left untouched.

The transformations which leave any, or any combination of, the three key concepts in a PCA invariant will be looked at in this Subsection. During the discussion, the constraint on the singular values being ordered by decreasing magnitudes will be temporarily relaxed, to accommodate cases like the same set of unit-norm PCs in different rankings of importance. It will be noted that this constraint played no role in Theorem IV.4, which will be at the centre of our discussion.

i). *Invariance of unit-norm PCs.* Unit-norm PCs are preserved by taking $\boldsymbol{\beta} = \boldsymbol{\alpha}$ in (IV.3.1). The transformation matrices $\mathbf{U}$ with this property have an SVD given by $\mathbf{U} = \mathbf{B}\boldsymbol{\Delta}\mathbf{A}'$ with $\mathbf{B}$ and $\boldsymbol{\Delta}$ arbitrary orthogonal and non-negative diagonal matrices, respectively. Thus, these matrices $\mathbf{U}$ are only characterized by the fact that their right singular vectors are the PAs in $\mathbb{R}^p$ of $\mathbf{X}$, that is, the eigenvectors of the covariance matrix determined by $\mathbf{X}$ (although possibly in a different order of importance).

ii). *Invariance of PCs and their natural lengths.* This was the case discussed above. We now require $\boldsymbol{\beta} = \boldsymbol{\alpha}$ and $\mathcal{T} = \Sigma$ in (IV.3.1). The transformation matrix $\mathbf{U}$ becomes $\mathbf{B}\mathbf{A}'$ with $\mathbf{B}$ arbitrary, *i.e.*, any orthogonal matrix.

iii). *Invariance of unit-norm PAs in* $\mathbb{R}^p$. The requirement is that $\mathbf{A} = \mathbf{B}$ in (IV.3.1), with $\boldsymbol{\beta}$ and $\mathcal{T}$ arbitrary. We have already seen that $\boldsymbol{\beta}'\boldsymbol{\alpha}$ is an orthogonal matrix (in the proof of Theorem IV.4). Since $\boldsymbol{\beta}$ is arbitrary, we can write the transformation matrices for this case as $\mathbf{U} = \mathbf{A}\mathcal{T}\mathbf{R}\Sigma^{-1}\mathbf{A}'$, where Σ is the matrix of singular values of $\frac{1}{\sqrt{n}}\mathbf{X}$, $\mathcal{T}$ is an arbitrary diagonal matrix of positive diagonal elements and $\mathbf{R}$ is an arbitrary orthogonal matrix.

iv). *Invariance of PAs in* $\mathbb{R}^p$ *and the variances accounted for.* To the requirement of the preceding case we add the requirement that $\mathcal{T} = \Sigma$. The resulting transformation matrices are of the form $\mathbf{U} = \mathbf{A}\Sigma\mathbf{R}\Sigma^{-1}\mathbf{A}'$ with $\mathbf{R}$ an arbitrary orthogonal matrix, *i.e.*, they are similar to some orthogonal matrix with $\mathbf{A}\Sigma$ defining the similarity. This effectively rules out diagonal transformation matrices, other than those whose diagonal elements are ±1, which are the only possible real eigenvalues of orthogonal matrices (see [25, (pg.71)]). Likewise, $\mathbf{U}$ cannot be a **p.s.d.** matrix (other than the identity).

v). *Invariance of all pairs of (unit-norm) PAs.* The Principal Axes in both $\mathbb{R}^p$ and $\mathbb{R}^n$ remain invariant when $\boldsymbol{\beta} = \boldsymbol{\alpha}$ and $\mathbf{B} = \mathbf{A}$ in (IV.3.1). The resulting transformation matrices are of the form $\mathbf{U} = \mathbf{A}\boldsymbol{\Delta}\mathbf{A}'$ for an arbitrary positive diagonal matrix $\boldsymbol{\Delta}$. Thus, such transformation matrices are positive definite matrices whose eigenvectors are the eigenvectors of the covariance matrix determined by $\mathbf{X}$. This result assumes that both sets of PAs are the same *and* remain coupled in the same way. It can be generalized to lift the coupling requirement by taking $\mathbf{B}$ to be a matrix whose columns are some permutation of those of $\mathbf{A}$.

vi). *Invariance of the variances accounted for.* The singular values of $\frac{1}{\sqrt{n}}\mathbf{X}$ and $\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{U}'$ are the same if $\mathcal{T} = \Sigma$ in (IV.3.1). The matrices $\mathbf{U}$ are of the form $\mathbf{B}\Sigma\mathbf{R}\Sigma^{-1}\mathbf{A}'$ where $\mathbf{B}$ and $\mathbf{R}$ are arbitrary orthogonal matrices.

vii). *Invariance of the whole PCA.* All aspects of the PCA remain invariant if and only if $\mathbf{U} = \mathbf{I}_p$, as can be seen by taking $\boldsymbol{\beta} = \boldsymbol{\alpha}$, $\mathcal{T} = \Sigma$ and $\mathbf{B} = \mathbf{A}$ in (IV.3.1).

If the invariance of the *proportion* of the total variance explained by each PC (rather than the *absolute* variances) is required, an arbitrary constant may be multiplied in, whenever appropriate.

IV.3.3 The SVDs of X and XU′

So far, the emphasis has been on what kinds of transformation matrices $\mathbf{U}$ will ensure PCAs of $\mathbf{XU}'$ with certain characteristics. But the most common aspect of interest is how a *given* transformation matrix $\mathbf{U}$ will affect the PCA of $\mathbf{X}$. In other words, we assume that the SVDs of $\mathbf{X}$ and $\mathbf{U}$ are known, whilst that of $\mathbf{XU}'$ is not. For simplicity, we assume that $\mathbf{X}$ is of rank p .

A crucial role in this Subsection is played by the $p \times p$ matrix $\mathbf{P}$ whose (i, j)-th entry is the (conventional) *inner product between the i-th eigenvector of the covariance matrix* $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ *and the j-th eigenvector of the metric matrix* $\mathbf{M} = \mathbf{U}'\mathbf{U}$, where both eigenvectors are *standardized to have (conventional) norms squared equal to their respective eigenvalues.* We shall call this matrix $\mathbf{P}$ the **matrix of interactions** between the data matrix $\mathbf{X}$ and the transformation matrix $\mathbf{U}$.

Theorem IV.5 *Let* $\mathbf{W}$ *be an* n×p *matrix of rank* p, *whose SVD is given by* $\mathbf{W} = \boldsymbol{\alpha}\boldsymbol{\Sigma}\mathbf{A}' = \sum_{i=1}^{p} \sigma_i \boldsymbol{\alpha}_i \mathbf{a}_i{}'$. *Let* $\mathbf{U}$ *be a* p×p *invertible matrix with SVD* $\mathbf{U} = \mathbf{L}\boldsymbol{\Omega}\mathbf{K}' = \sum_{j=1}^{p} \omega_j \mathbf{l}_j \mathbf{k}_j{}'$. *Let* $\mathbf{P}$ *be the* p×p *matrix of interactions between* $\mathbf{W}$ *and* $\mathbf{U}$, *whose* (i,j)*-th element is* $< \sigma_i \mathbf{a}_i, \omega_j \mathbf{k}_j >_{I_p}$. *Then:*

i). The singular values of $\mathbf{WU}'$ *are those of the matrix of interactions* $\mathbf{P}$.

ii). The left singular vectors of $\mathbf{WU}'$ *are the columns of the product of the matrix of left singular vectors of* $\mathbf{W}$ *with the matrix of left singular vectors of* $\mathbf{P}$.

iii). The right singular vectors of $\mathbf{WU}'$ *are the columns of the product of the matrix of left singular vectors of* $\mathbf{U}$ *with the matrix of right singular vectors of* $\mathbf{P}$.

Proof. From the SVDs of $\mathbf{W}$ and $\mathbf{U}$, we have $\mathbf{WU}' = \boldsymbol{\alpha}(\boldsymbol{\Sigma}\mathbf{A}'\mathbf{K}\boldsymbol{\Omega})\mathbf{L}' = \boldsymbol{\alpha}\mathbf{P}\mathbf{L}'$. $\mathbf{P}$ must be a rank p matrix , since $\mathbf{A}$ is of rank p ($\mathbf{W}$ is of rank p) and pre- and post-multiplying any matrix by invertible matrices (as are $\boldsymbol{\Sigma}$, $\mathbf{K}$ and $\boldsymbol{\Omega}$) leaves the rank of the product unchanged (see Horn and Johnson (1985) [25, (pg.13)]). Thus, an SVD of $\mathbf{P}$ is of the form $\mathbf{P} = \mathbf{Q}\boldsymbol{\Delta}\mathbf{R}'$ with $\mathbf{Q}$,$\mathbf{R}$ orthogonal $p \times p$ matrices and $\boldsymbol{\Delta}$ a $p \times p$ diagonal matrix. Substituting above, we have:

$$\mathbf{WU}' = (\boldsymbol{\alpha}\mathbf{Q})\boldsymbol{\Delta}(\mathbf{LR})' \tag{IV.3.2}$$

But this is an SVD of $\mathbf{WU}'$, since $\boldsymbol{\Delta}$ is diagonal (with diagonal elements in descending order of magnitude) and the columns of $\boldsymbol{\alpha}\mathbf{Q}$ and $\mathbf{LR}$ are orthonormal. In fact:

$$(\boldsymbol{\alpha}\mathbf{Q})'(\boldsymbol{\alpha}\mathbf{Q}) \quad = \quad \mathbf{Q}'\mathbf{I}_p\mathbf{Q} = \mathbf{I}_p$$

$$(\mathbf{LR})'(\mathbf{LR}) \quad = \quad \mathbf{R}'\mathbf{I}_p\mathbf{R} = \mathbf{I}_p$$

The singular values of $\mathbf{WU}'$ are those of $\mathbf{P}$, the left singular vectors of $\mathbf{WU}'$ are the columns of $\alpha\mathbf{Q}$ and the right singular vectors of $\mathbf{WU}'$ are the columns of $\mathbf{LR}$. □

Again, the relevance of this Theorem for PCA comes from considering the matrix $\mathbf{W}$ to be a matrix of the form $\frac{1}{\sqrt{n}}\mathbf{X}$ where $\mathbf{X}$ is a column-centered data matrix. Its importance is not computational. Obviously, determining the SVD of $\mathbf{U}$ to obtain matrix $\mathbf{P}$, and then determining the SVD of $\mathbf{P}$ is more complicated a task then to compute $\mathbf{WU}'$ and its SVD directly. What the Theorem does is to highlight the *way* in which the SVD of $\mathbf{WU}'$ is determined by $\mathbf{W}$ and $\mathbf{U}$. In particular, it highlights the crucial role of the matrix of interactions, $\mathbf{P}$, whose elements are the inner products of the (weighted) eigenvectors of the covariance and metric matrices.

One special case which is immediately apparent from the Theorem arises when $\mathbf{P}$ is itself a diagonal matrix. This occurs when the eigenvectors of the covariance matrix $\mathbf{S}$ and the metric matrix $\mathbf{M}$ coincide (including in their order). Then, the singular values of $\mathbf{XU}'$ are merely the product of the singular values of $\mathbf{X}$ and those of $\mathbf{U}$. This case can be extended to include some permutation in the order of the common eigenvectors of $\mathbf{S}$ and $\mathbf{M}$. In the case, $\mathbf{\Sigma A'K\Omega}$ will not be a diagonal matrix, but a 'permuted diagonal matrix', that is, a matrix with a single non-zero element in each row and each column. These non-zero elements (which are the products of the σ_i's with some permutation of the ω_j's) are the singular values of $\mathbf{P}$. The matrices of right and left singular vectors of $\mathbf{P}$ will be **permutation matrices** (that is, matrices with a single non-zero entry of value 1 in each row and each column; such matrices are orthogonal – Horn and Johnson (1985) [25, (pg.25)]) suitably designed to 'unscramble' the ordered singular values and place them in their original positions in matrix $\mathbf{P}$. Thus, the PCs and PAs in $\mathbb{R}^p$ of $\mathbf{XU}'$ are still equal to those of $\mathbf{X}$, only re-ordered according to the effects of the permutation matrices of singular vectors of $\mathbf{P}$.

Another special case, of greater potential interest, occurs when the transformation matrix $\mathbf{U}$ is diagonal. Then, the matrices of its left and right singular vectors are also permutation matrices (the identity if the diagonal elements of $\mathbf{U}$ are already in decreasing order of magnitude). Now the *columns* of $\mathbf{R}$ in (IV.3.2) are the right singular vectors of $\mathbf{P}$. Pre-multiplication by the permutation matrix $\mathbf{L}$ permutes the *rows* of $\mathbf{R}$. This implies that the PAs in $\mathbb{R}^p$ of $\mathbf{XU}'$ are the right singular vectors of $\mathbf{P}$ with their entries suitably permuted.

IV.3.4 Linear transformations and the PSD cone

We shall now turn our attention to the effects of linear transformations of the data on the covariance matrices which the data determine.

We begin by considering a mapping from the space of $p \times p$ symmetric matrices, $\mathbb{S}_p$, onto itself, defined as:

$$\mathcal{N}_U(\mathbf{V}) = \mathbf{UVU}' \qquad \forall \mathbf{V} \in \mathbb{S}_p \tag{IV.3.3}$$

where $\mathbf{U}$ is any invertible $p \times p$ matrix.

The mapping $\mathcal{N}_U$ is one possible generalization of the orthogonal similarity mapping $\mathcal{M}_A$ given by equation (III.3.8). But $\mathcal{N}_U$ is no longer a similarity since $\mathbf{U}'$ is not $\mathbf{U}^{-1}$ unless $\mathbf{U}$ is orthogonal.

The mapping (IV.3.3) has been studied in Matrix Theory, where matrices $\mathbf{V}$ and $\mathcal{N}_U(\mathbf{V})$ are known as **congruent** matrices (Horn and Johnson (1985) [25, (pg.220)]). We shall call mappings $\mathcal{N}_U$ **congruence mappings**. The most powerful result concerning congruent matrices is *Sylvester's law of inertia* which is formally stated below as given by Horn and Johnson ([25, (pg.223)]).

Theorem IV.6 (Sylvester's Law of Inertia) *Let* $\mathbf{V}, \mathbf{W} \in \mathbb{S}_{\mathrm{p}}$ *be any two symmetric* $\mathrm{p} \times \mathrm{p}$ *matrices. There is a non-singular matrix* $\mathbf{U} \in \mathbb{M}_{\mathrm{p}}$ *such that* $\mathbf{V} = \mathbf{U}\mathbf{W}\mathbf{U}'$ *if and only if* $\mathbf{V}$ *and* $\mathbf{W}$ *have the same* **inertia**, *that is, the same number of positive, negative and zero eigenvalues.*

The main implication of Sylvester's law of inertia, for our purposes, is that $\mathcal{N}_U$ *maps the* **PSD** *cone onto itself* and, in particular, the set of all positive definite matrices in $\mathbb{S}_p$ onto itself. This is merely reflecting the fact that $\mathcal{N}_U(\mathbf{S})$ is the covariance matrix of $\mathbf{X}\mathbf{U}'$ if $\mathbf{S}$ is the covariance matrix of $\mathbf{X}$, together with the results of Subsection IV.3.1.

Unfortunately, there are no known simple and comprehensive results expressing the eigenvalues and eigenvectors of $\mathbf{U}\mathbf{S}\mathbf{U}'$ in terms of those of $\mathbf{S}$. But some results can be obtained which, although of little computational interest, provide further insight into how linear transformations are affecting the PCA of a data matrix. We shall therefore look into how the **PSD** cone in $\mathbb{S}_p$ is affected by the mapping $\mathcal{N}_U$.

In Section IV.1 we saw how the class of transformation matrices $\mathbf{U}$ which gave rise to the same inner product matrix $\mathbf{M} = \mathbf{U}'\mathbf{U}$ produced transformed data matrices $\mathbf{XU}'$ whose PCAs shared many common aspects: same unit-norm PCs, same relative importance (variance accounted for) of each PC, same $\mathbf{M}$-orthonormal PAs in $\mathbb{R}^p$. One implication of this is that the covariance matrices $\mathbf{USU}'$ for transformation matrices $\mathbf{U}$ in this class will all have common eigenvalues.

That this is so can also be seen from Theorem IV.3. This Theorem implies that the covariance matrices of $\mathbf{XU}_1{}'$ and of $\mathbf{XU}_2{}'$, where $\mathbf{U}_1{}'\mathbf{U}_1 = \mathbf{U}_2{}'\mathbf{U}_2 = \mathbf{M}$ are orthogonally similar. In fact, the two covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ determined by $\mathbf{XU}_1{}'$ and $\mathbf{XU}_2{}'$ must be related by a mapping of the kind (III.3.8), for some orthogonal matrix $\mathbf{A}$. Theorem III.8 then applies, confirming that $\mathbf{S}_1$ and $\mathbf{S}_2$ have common eigenvalues.

This implies that the class of covariance matrices $\mathbf{USU}'$ for which $\mathbf{U}'\mathbf{U}$ gives a common inner product matrix $\mathbf{M}$ must share certain geometric characterizations of their locations on the **PSD** cone, which depend only on the eigenvalues of the matrix. In particular, these covariance matrices must all lie on a common norm-k hypersphere around the origin of $\mathbb{S}_p$ (III.1.12); they will all form a common angle with the central ray of the **PSD** cone (III.3.2); they are all at a common distance from the central ray of the **PSD** cone (Theorem III.4); the angles which solve the successive minimization problem described in Theorem III.7 are the same (although the rank one idempotent matrices with which these angles are formed will differ). All of this can be visualized by recalling that the orthogonal similarity mapping $\mathcal{M}_A$ mentioned in the previous paragraph induced what we called a rigid rotation of the **PSD** cone around its central ray (Theorem III.8). The set of covariance matrices $\mathbf{USU'}$ with $\mathbf{U'U} = \mathbf{M}$ are all images of each other under appropriate orthogonal similarity mappings $\mathcal{M}_A$.

But we can take one step further than the results from Chapter III, and quantify some of the above geometric characteristics in terms only of the eigenvalues and the **M**-inner products of the eigenvectors of the original covariance matrix **S**.

Theorem IV.7 *Let* **X** *be a rank* p n×p *column-centered data matrix, whose covariance matrix* $\mathbf{S} = \frac{1}{n}\mathbf{X'X}$ *has spectral decomposition* $\mathbf{S} = \mathbf{A\Lambda A'} = \sum_{i=1}^{p} \lambda_i \mathbf{a}_i \mathbf{a}_i'$. *Let* **M** *be a positive definite matrix. For any factorization* $\mathbf{M} = \mathbf{U'U}$, *the following results involving the covariance matrix* $\mathbf{USU'}$ *of the transformed data matrix* $\mathbf{XU'}$ *hold:*

i). The matrix **Q** *whose (i, j)-th element is* $<\sqrt{\lambda_i}\mathbf{a}_i, \sqrt{\lambda_j}\mathbf{a}_j>_M$ *has the same eigenvalues (counting multiplicities) as* $\mathbf{USU'}$.

ii).

$$\mathrm{tr}(\mathbf{USU'}) = \sum_{i=1}^{p} \lambda_i \|\mathbf{a}_i\|_M^2 \qquad \text{(IV.3.4)}$$

iii).

$$||\mathbf{USU'}|| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} \lambda_i\lambda_j < \mathbf{a}_i, \mathbf{a}_j >_M^2} \qquad \text{(IV.3.5)}$$

iv).

$$\cos(\mathbf{USU'}, \mathbf{I}_p) = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\omega_{ij}}{p\cdot\sum_{i=1}^{p}\sum_{j=1}^{p}\omega_{ij}\cos_M^2(\mathbf{a}_i, \mathbf{a}_j)}} \qquad \text{(IV.3.6)}$$

where $\omega_{ij} = \lambda_i\lambda_j||\mathbf{a}_i||_M^2||\mathbf{a}_j||_M^2$ *and* $\cos_M(\mathbf{a}_i, \mathbf{a}_j)$ *is the cosine of the* **M***-angle between* $\mathbf{a}_i$ *and* $\mathbf{a}_j$*.*

v).

$$Var_\eta = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\omega_{ij}}{\sum_{k=1}^{p}\sum_{l=1}^{p}\omega_{kl}}\left[\cos_M^2(\mathbf{a}_i, \mathbf{a}_j) - \frac{1}{p}\right] \qquad \text{(IV.3.7)}$$

where Var_η *denotes the variance of the relative eigenvalues of* **USU'***.*

vi).

$$Var_\mu = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\omega_{ij}\left[\cos_M^2(\mathbf{a}_i, \mathbf{a}_j) - \frac{1}{p}\right] \qquad \text{(IV.3.8)}$$

where Var_μ *denotes the variance of the (absolute) eigenvalues of* **USU'***.*

Proof.

i). The matrix **Q** is given by $\mathbf{Q} = \mathbf{\Lambda}^{1/2}\mathbf{A'MA\Lambda}^{1/2}$. If the spectral decomposition of **M** is given by $\mathbf{M} = \mathbf{K}\boldsymbol{\mu}\mathbf{K'}$, then **Q** is given by

$$\mathbf{Q} = (\mathbf{\Lambda}^{1/2}\mathbf{A'K}\boldsymbol{\mu}^{1/2})(\boldsymbol{\mu}^{1/2}\mathbf{K'A\Lambda}^{1/2}) = \mathbf{PP'}$$

where **P** is the matrix of interactions between the data matrix **X** and the transformation matrix **U**, as defined in Theorem IV.5. In that case, the eigenvalues of **Q** are the squares of the singular values of **P'** (by definition), that is, of $\mathbf{P}$. But then the eigenvalues of $\mathbf{Q}$ are also the eigenvalues of $\mathbf{USU}'$ for all factorizations $\mathbf{M} = \mathbf{U}'\mathbf{U}$ (Theorem IV.5).

ii). From point i), $\mathrm{tr}(\mathbf{USU}') = \mathrm{tr}(\mathbf{Q}) = \sum_{i=1}^{p} \lambda_i \|\mathbf{a}_i\|_M^2$.

iii). In (III.1.6) we saw how the square of the Frobenius norm of a matrix is the sum of the squares of all its elements. In (III.1.5) we saw how it is also the sum of squares of its singular values. In Subsection I.3.2 we saw how, for positive definite matrices, the concepts of singular and eigen values coincide. Thus, denoting the eigenvalues of $\mathbf{USU}'$ by $\{\mu_i\}_{i=1}^{p}$, we have – from point i) above:

$$\sum_{i=1}^{p} \mu_i^2 = \|\mathbf{USU}'\|^2 = \|\mathbf{Q}\|^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_i \lambda_j {<}\, \mathbf{a}_i, \mathbf{a}_j {>}_M^2$$

iv). Direct from the definition of $\cos(\mathbf{USU}', \mathbf{I}_p)$ and the two previous points. The more compact notation $\omega_{ij} = \lambda_i \lambda_j \|\mathbf{a}_i\|_M^2 \|\mathbf{a}_j\|_M^2$ should not disguise the fact that $\sum_{i=1}^{p} \sum_{j=1}^{p} \omega_{ij}$ can be factorized as $\left(\sum_{i=1}^{p} \|\sqrt{\lambda_i}\mathbf{a}_i\|_M^2\right)^2$.

v). Directly from point iv) and from equation (III.3.4). It should be noted that $p \cdot Var_\eta$ is a weighted average of the differences $\cos_M^2(\mathbf{a}_i, \mathbf{a}_j) - \frac{1}{p}$.

vi). Directly from points ii) and v). It can also be obtained from Theorem III.4.

□

Theorem IV.7 tells us that the common trace-k hyperplane containing all covariance matrices of the type $\mathbf{USU}'$ with $\mathbf{U}'\mathbf{U} = \mathbf{M}$ will have $k = \mathrm{tr}(\mathbf{SM})$. In addition, it tells us that this trace is the sum of squares of the $\mathbf{M}$-norms of the eigenvectors of $\mathbf{S}$, when standardized to have (conventional) lengths squared equal to their corresponding eigenvalues. Similar expressions are given for the norm-k hypersphere on which they lie and the angle and distance from the central ray, in terms only of the matrix $\mathbf{M}$ and the eigenpairs of $\mathbf{S}$.

The matrix $\mathbf{Q}$ defined in point i) of the Theorem also provides some notion of how the eigenvalues of the original covariance matrix can be affected. If the matrix $\mathbf{Q}$ is diagonal, that is, if $< \mathbf{a}_i, \mathbf{a}_j >_M = 0$, for $i \neq j$, then the eigenvalues of $\mathbf{USU}'$ will be $\{\lambda_i ||\mathbf{a}_i||^2_M\}^p_{i=1}$. This happens not just in the trivial case when $\mathbf{M} = \mathbf{I}_p$ (that is, when the transformation matrix $\mathbf{U}$ is an orthogonal matrix), but also when $\mathbf{M}$ has the same eigenvectors as $\mathbf{S}$, in which case $||\mathbf{a}_i||^2_M$ will be the eigenvalue of $\mathbf{M}$ associated with that eigenvector. This result (which had also been derived from Theorem IV.5) can be extended to matrices $\mathbf{Q}$ which are 'approximately diagonal' (that is, with $< \mathbf{a}_i, \mathbf{a}_j >_M \simeq 0, \ \forall i, j$). A simple eigenvalue location theorem (like the Geršgorin disc theorem – see Horn and Johnson (1985) [25, (pg.344)]) will tell us that in such cases the eigenvalues of $\mathbf{Q}$ are still 'approximately' $\{\lambda_i ||\mathbf{a}_i||^2_M\}^p_{i=1}$. More involved eigenvalue location results may provide useful tips concerning the effects of using certain linear transformations on any given data set. A detailed discussion of such results may be found in Chapter VI of Horn and Johnson (1985) [25].

The fact that $\mathbf{Q}$ can be written in terms of the matrix $\mathbf{P}$ of interactions between $\mathbf{X}$ and $\mathbf{U}$ — noted in the proof to the first point of the previous theorem — implies that equivalent expressions to (IV.3.4) and (IV.3.8) can be obtained involving the usual inner products between the eigenvectors of **S** and **M**, weighted by the square roots of their respective eigenvalues. For example, (IV.3.4) can also be written as:

$$\mathrm{tr}(\mathbf{USU}') = \mathrm{tr}(\mathbf{Q}) = \mathrm{tr}(\mathbf{PP}') = \|\mathbf{P}\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} \lambda_i \mu_j \cos^2(\mathbf{a}_i, \mathbf{k}_j) \qquad \text{(IV.3.9)}$$

Unfortunately, the other expressions for the quantities in Theorem IV.7 are too involved to be of any great interpretive value.

A complete analogue of Theorem IV.7 can also be obtained if instead of considering the matrix $\mathbf{Q} = \mathbf{PP}'$, we use the matrix $\mathbf{P}'\mathbf{P}$. The generic element of this matrix is the **S**-inner product of the i-th and j-th weighted eigenvectors of the matrix **M**:

$$(\mathbf{P}'\mathbf{P})_{(i,j)} = < \sqrt{\mu_i}\mathbf{k}_i, \sqrt{\mu_j}\mathbf{k}_j >_S$$

This matrix is orthogonally similar to the matrix **Q**, since any pair of matrices of the form $\mathbf{C}'\mathbf{C}$ and $\mathbf{CC}'$ (for any square matrix **C**) are orthogonally similar (Horn and Johnson (1985) [25, (pg.414)]). Hence, $\mathbf{P}'\mathbf{P}$ also shares the same eigenvalues as all matrices of the form $\mathbf{USU}'$. The quantities (IV.3.4) through (IV.3.8) can therefore be re-written replacing **M** with **S** and the eigenpairs of **S** with those of **M**.

Alternative inner products in the PSD cone

The discussion of the effects of the transformation $\mathcal{N}_U$ on the **PSD** cone becomes much tidier — if less readily interpretable — when an alternative inner product is defined in the matrix space $\mathbb{S}_p$. This new inner product is a natural extension of the '**M**-metric PCA' approach described in Section IV.2.

In Section IV.1 we saw how the Frobenius inner product between matrices of a generic matrix space $\mathbb{M}_{m\times q}$ could be generalized (equation (IV.1.3)). This generalization corresponded to replacing the standard inner products in $\mathbb{R}^m$ and $\mathbb{R}^q$ with more general inner products.

When considering matrices in $\mathbb{M}_p$, with the inner product in $\mathbb{R}^p$ defined by a positive definite matrix **M**, the corresponding inner product in $\mathbb{M}_p$ can be the **(M,M)-inner product** defined as in (IV.1.3):

$$< \mathbf{A}, \mathbf{B} >_{(M,M)} = \mathrm{tr}(\mathbf{A}'\mathbf{M}\mathbf{B}\mathbf{M}) \qquad \forall \mathbf{A}, \mathbf{B} \in \mathbb{M}_p \qquad \text{(IV.3.10)}$$

The implications of the linear transformation $\mathcal{N}_U$ on the **PSD** cone can now be viewed in terms of the original matrices, with direct analogues to the results in Section III.3, but with the new inner product defined in the **PSD** cone.

It should be stressed that this approach is particularly well suited if we are concerned with the **M**-orthogonal PAs in $\mathbb{R}^p$ of a data matrix **X**, rather than with the usual PAs in $\mathbb{R}^p$ of the transformed data matrix $\mathbf{XU}'$ (with $\mathbf{M} = \mathbf{U}'\mathbf{U}$).

Theorem IV.8 *Let* $\mathbf{X} \in \mathbb{C}_{\mathrm{n\times p}}$ *be an* n×p *column-centered data matrix and let* $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ *be the covariance matrix it determines. Let* **M** *be a* p×p *positive definite matrix. Let* $\{(\lambda_i^{[M]}, \mathsf{b}_i^{[M]})\}_{i=1}^{p}$ *be a set of eigenpairs of the matrix* **SM**. *Then, the following results hold:*

i).

$$\cos_{(M,M)}(\mathbf{S},\mathbf{M}^{-1}) = \frac{\|\boldsymbol{\lambda}^{[M]}\|_1}{\sqrt{p}\|\boldsymbol{\lambda}^{[M]}\|_2} = \frac{1}{\sqrt{p}\sqrt{\sum_{i=1}^{p}\pi_i^2}} = \frac{1}{\sqrt{1+p^2\cdot Var_\pi}}$$

where $\cos_{(M,M)}(\cdot,\cdot)$ *denotes the cosine of the angle defined in* $\mathbb{M}_\mathrm{p}$ *by the (***M***,***M***)-inner product (IV.3.10);* $\boldsymbol{\lambda}^{[M]}$ *is the* p *-dimensional vector of eigenvalues of* **SM***:* $\pi_i = \frac{\lambda_i^{[M]}}{\sum_{i=1}^{p}\lambda_i^{[M]}}$ *is the i-th relative eigenvalue of* **SM** *and* Var_π *is the variance of the* π_i*'s.*

ii). If **S** *is of rank* k *, then:*

$$\sqrt{\tfrac{1}{m}} \leq \cos_{(M,M)}(\mathbf{S},\mathbf{M}^{-1}) \leq \sqrt{\tfrac{k}{m}}$$

Furthermore, $\cos_{(M,M)}(\mathbf{S},\mathbf{M}^{-1}) = \sqrt{\frac{k}{m}}$ *if and only if all the non-zero eigenvalues of* **SM** *are equal.*

iii). $\cos_{(M,M)}(\mathbf{S},\mathbf{M}^{-1}) = \sqrt{\frac{1}{m}}$ *if and only if* **S** *is of rank one.*

*iv). The image of the orthogonal projection (with the (***M***,***M***)-inner product in* $\mathbb{M}_\mathrm{p}$*) of* **S** *onto the subspace* $\mathrm{span}(\mathbf{M}^{-1}) \subset \mathbb{S}_\mathrm{p}$ *is* $\mathbf{S}^* = \overline{\lambda^{[M]}}\mathbf{M}^{-1}$*, where* $\overline{\lambda^{[M]}}$ *is the mean value of the eigenvalues* $\{\lambda_i^{[M]}\}_{i=1}^{p}$ *of* **SM***. The distance from* **S** *to* $\mathbf{S}^*$ *is given by* $\sqrt{p\cdot Var_{\lambda^{[M]}}}$*, where* $Var_{\lambda^{[M]}}$ *is the variance of* **SM***'s eigenvalues.*

v). The eigenvectors $\{\mathbf{b}_i^{[M]}\}_{i=1}^{p}$ *of* **SM** *solve the successive optimization problem of finding the rank one, (***M***,***M***)-norm one matrices* $\{\mathbf{b}_i^{[M]}\mathbf{b}_i^{[M]'}\}_{i=1}^{p}$ *which minimize the (***M***,***M***)-angles with* **S***, subject to their own (***M***,***M***)-orthogonality. The* *cosines of these angles are the ℓ_2-norm relative eigenvalues of* $\mathbf{SM}$, $\left\{\frac{\lambda_i^{[M]}}{\|\boldsymbol{\lambda}^{[M]}\|_2}\right\}_{i=1}^{p}$ *corresponding to each* $\mathbf{b}_i^{[M]}$. *The coefficients of the* $(\mathbf{M},\mathbf{M})$*-orthogonal projection of* $\mathbf{S}$ *onto the rays defined by the matrices* $\{\mathbf{b}_i^{[M]}\mathbf{b}_i^{[M]'}\}_{i=1}^{p}$ *are the (absolute) eigenvalues* $\{\lambda_i^{[M]}\}_{i=1}^{p}$ *of* $\mathbf{SM}$.

Proof. The Theorem is a direct consequence of the results in Section III.3 and of the fact that, *for any factorization* $\mathbf{M} = \mathbf{U}'\mathbf{U}$, the transformation $\mathcal{N}_U$, given by (IV.3.3), satisfies the following relation:

$$\langle \mathcal{N}_U(\mathbf{A}), \mathcal{N}_U(\mathbf{B})\rangle_{(\mathbf{I}_p,\mathbf{I}_p)} = \langle \mathbf{A}, \mathbf{B}\rangle_{(M,M)} \qquad \forall \mathbf{A}, \mathbf{B} \in \mathbb{S}_p \qquad \text{(IV.3.11)}$$

In fact,

$$\langle \mathcal{N}_U(\mathbf{A}), \mathcal{N}_U(\mathbf{B})\rangle_{(\mathbf{I}_p,\mathbf{I}_p)} = \langle \mathbf{UAU}', \mathbf{UBU}'\rangle_{(\mathbf{I}_p,\mathbf{I}_p)} = \mathrm{tr}(\mathbf{A}'\mathbf{MBM}) = \langle \mathbf{A}, \mathbf{B}\rangle_{(M,M)}$$

If we also note that $\mathcal{N}_U$ defines a linear transformation on the matrices of $\mathbb{S}_p$ (that is, $\mathcal{N}_U(\alpha\mathbf{A} + \beta\mathbf{B}) = \alpha\mathcal{N}_U(\mathbf{A}) + \beta\mathcal{N}_U(\mathbf{B})$ for all positive definite $p\times p$ matrices $\mathbf{A}$,$\mathbf{B}$ and all scalars α, β), then we see that *the mapping* $\mathcal{N}_U$ *defines an inner-product preserving isomorphism between the inner product space* $\mathbb{S}_\mathrm{p}$ *with the* $(\mathbf{M},\mathbf{M})$*-inner product and* $\mathbb{S}_\mathrm{p}$ *with the usual Frobenius* $(\mathbf{I}_p,\mathbf{I}_p)$*-inner product.* In particular, we have:

$$\left\langle \mathbf{S}, \mathbf{M}^{-1}\right\rangle_{(M,M)} = \left\langle \mathcal{N}_U(\mathbf{S}), \mathcal{N}_U(\mathbf{M}^{-1})\right\rangle_{(\mathbf{I}_p,\mathbf{I}_p)} = \langle \mathbf{USU}', \mathbf{I}_p\rangle_{(\mathbf{I}_p,\mathbf{I}_p)}$$

and

$$\begin{aligned}
\cos_{(M,M)}(\mathbf{S},\mathbf{M}^{-1}) &= \frac{<\mathbf{S},\mathbf{M}^{-1}>_{(M,M)}}{||\mathbf{S}||_{(M,M)}\cdot||\mathbf{M}^{-1}||_{(M,M)}} \\
&= \frac{<\mathbf{USU}',\mathbf{I}_p>_{(\mathbf{I}_p,\mathbf{I}_p)}}{||\mathbf{USU}'||_{(\mathbf{I}_p,\mathbf{I}_p)}\cdot||\mathbf{I}_p||_{(\mathbf{I}_p,\mathbf{I}_p)}} \qquad (IV.3.12) \\
&= \cos_{(\mathbf{I}_p,\mathbf{I}_p)}(\mathbf{USU}',\mathbf{I}_p)
\end{aligned}$$

Now, the eigenvalues of $\mathbf{SM}$ are those of $\mathbf{USU}'$. Hence:

i). Equations (III.3.2), (III.3.3) and (III.3.4), together with equation (IV.3.12) give the desired result.

ii). A direct consequence of Theorems III.1 and III.3, with equation (IV.3.12). For $\mathbf{M}$ positive definite, the ranks of $\mathbf{S}$ and $\mathbf{SM}$ are equal, as we saw in Subsection IV.3.1.

iii). A direct consequence of Theorem III.2 and (IV.3.12), together with the fact that $rank(\mathbf{S}) = rank(\mathbf{SM})$.

iv). This point involves ($\mathbf{M}$,$\mathbf{M}$)-orthogonal projections in $\mathbb{S}_p$. From equation (III.3.7) we have that the image, $\mathbf{S}^*$, of the ($\mathbf{M}$,$\mathbf{M}$)-orthogonal projection of $\mathbf{S}$ onto $span(\mathbf{M}^{-1})$ is the solution to the equation:

$$< \mathbf{S}-\mathbf{S}^*,\mathbf{A} >_{(M,M)} = 0 \qquad \forall \mathbf{A} \in span(\mathbf{M}^{-1})$$

or, writing $\mathbf{S}^* = \alpha^*\mathbf{M}^{-1}$ and $\mathbf{A} = \alpha\mathbf{M}^{-1}$, the solution of:

$$\left\langle \mathbf{S}-\alpha^*\mathbf{M}^{-1},\alpha\mathbf{M}^{-1}\right\rangle_{(M,M)} = 0$$

From (IV.3.11) this is in turn equivalent to:

$$\left\langle \mathcal{N}_U(\mathbf{S} - \alpha^*\mathbf{M}^{-1}), \mathcal{N}_U(\alpha\mathbf{M}^{-1}) \right\rangle_{(\mathbf{I}_p,\mathbf{I}_p)} = 0$$

$$\iff \quad \langle \mathbf{USU}' - \alpha^*\mathbf{I}_p, \alpha\mathbf{I}_p \rangle_{(\mathbf{I}_p,\mathbf{I}_p)} = 0$$

Thus, we are $(\mathbf{I}_p.\mathbf{I}_p)$-orthogonally projecting $\mathcal{N}_U(\mathbf{S}) = \mathbf{USU}'$ onto $span(\mathbf{I}_p)$. Theorem III.4 now applies and we have $\alpha^* = \frac{\operatorname{tr}(\mathbf{USU}')}{p} = \overline{\boldsymbol{\lambda}^{[M]}}$ and

$$||\mathbf{S} - \mathbf{S}^*||_{(M,M)} = ||\mathbf{USU}' - \alpha^*\mathbf{I}_p||_{(\mathbf{I}_p,\mathbf{I}_p)} = \sqrt{p \cdot Var_{\boldsymbol{\lambda}^{[M]}}}$$

v). The rank one, $(\mathbf{M},\mathbf{M})$-orthonormal matrices $\mathbf{E}^{[M]}_{\mathbf{b}_i} = \mathbf{b}_i^{[M]}\mathbf{b}_i^{[M]'}$ satisfy the requirement $\left\langle \mathbf{E}^{[M]}_{\mathbf{b}_i}, \mathbf{E}^{[M]}_{\mathbf{b}_j} \right\rangle_{(M,M)} = \operatorname{tr}(\mathbf{E}^{[M]'}_{\mathbf{b}_i}\mathbf{M}\mathbf{E}^{[M]}_{\mathbf{b}_j}\mathbf{M}) = \left[(\mathbf{b}_i^{[\mathrm{M}]})'\mathbf{M}\mathbf{b}_j^{[\mathrm{M}]}\right]^2 = \delta_{\mathrm{ij}}$. This implies that the images of these matrices under $\mathcal{N}_U$ are orthonormal in the usual sense, since $\left\langle \mathcal{N}_U(\mathbf{E}^{[M]}_{\mathbf{b}_i}), \mathcal{N}_U(\mathbf{E}^{[M]}_{\mathbf{b}_j}) \right\rangle_{(\mathbf{I}_p,\mathbf{I}_p)} = \left\langle \mathbf{E}^{[M]}_{\mathbf{b}_i}, \mathbf{E}^{[M]}_{\mathbf{b}_j} \right\rangle_{(M,M)} = \delta_{ij}$. Thus, we have

$$\cos_{(M,M)}(\mathbf{S}, \mathbf{E}^{[M]}_{\mathbf{b}_i}) = \cos_{(\mathbf{I}_p,\mathbf{I}_p)}(\mathbf{USU}', \mathcal{N}_U(\mathbf{E}^{[M]}_{\mathbf{b}_i}))$$

where the matrices $\mathcal{N}_U(\mathbf{E}^{[M]}_{\mathbf{b}_i}) = \mathbf{c}_i\mathbf{c}_i' = (\mathbf{Ub}_i^{[M]})(\mathbf{Ub}_i^{[M]})'$ are mutually orthogonal rank one idempotent matrices. Theorem III.7 now applies. The vectors $\{\mathbf{c}_i\}_{i=1}^p$ which solve the optimization problems in the usual inner product are the eigenvectors of $\mathbf{USU}'$. Therefore, the vectors $\mathbf{b}_i^{[M]} = \mathbf{U}^{-1}\mathbf{c}_i$ which solve the optimization problems in the $(\mathbf{M},\mathbf{M})$-inner product are the eigenvectors of $\mathbf{SM}$. The results concerning the eigenvalues also flow directly from Theorem III.7.

□

As stated before, Theorem IV.8 provides a geometric characterization of the eigenvalues and eigenvectors of the matrix $\mathbf{SM}$ in terms of the location of $\mathbf{S}$ on the $\mathbf{PSD}$ cone of $\mathbb{S}_p$, when the inner product in $\mathbb{S}_p$ is given by (IV.3.10). The eigenpairs of $\mathbf{SM}$ are, as was seen before, the eigenvalues of $\mathbf{USU}'$ and the eigenvectors of $\mathbf{USU}'$ when transformed back to the original coordinate system, by $\mathbf{U}^{-1}$.

Thus, a geometric characterization of the $\mathbf{M}$-orthonormal Principal Axes in $\mathbb{R}^p$ of $\mathbf{X}$ and of the variance accounted for by the PCs which they define has been given. The key role played by the ray defined by the identity $\mathbf{I}_p$ has now been taken on by the ray defined by matrix $\mathbf{M}^{-1}$.

Unfortunately, similar characterizations of the PCs of $\mathbf{XU}'$ do not follow directly. In Chapter III we noted how the results on the location of **p.s.d.** matrices could be applied both to the covariance matrices (whose eigenvectors are the PAs in $\mathbb{R}^p$ of $\mathbf{X}$) and to the matrices $\frac{1}{n}\mathbf{XX}' \in \mathbb{S}_n$ (whose eigenvectors are the unit-norm PCs of $\mathbf{X}$). But the transformation $\mathbf{U}$ will affect $\mathbf{X'X}$ and $\mathbf{XX}'$ in different ways. The approach described above will not enable us to relate the position of $\mathbf{XMX}'$ (whose eigenvectors are the unit-norm PCs of $\mathbf{XU}'$) and that of $\mathbf{XX}'$ with a different inner product.

The fundamental problem with the results of Theorem IV.8, however, is again similar to that of an $\mathbf{M}$-inner product PCA: the change in inner products makes the already difficult 'visualization' of the geometric concepts in the $\mathbf{PSD}$ cone almost prohibitive.

Thus, Theorem IV.8 is more appealing from a theoretical point of view than from the point of view of any practical consequences. The more limited characterizations in Theorem IV.7 are more useful from the latter point of view.

The feeling remains that the geometry of the **PSD** cone has much which is yet to be explored. It is unlikely that the results of Section III.3 and of this Subsection are exhausting the full potential of this geometric approach to provide results which are useful for a fuller understanding of PCA and linear transformations.

IV.3.5 Effects on PCs

In this Subsection we turn our attention to the PCs of transformed data matrices.

The general context is that of a given column-centered data matrix $\mathbf{X}$ and two different invertible linear transformations $\mathbf{U}_1$ and $\mathbf{U}_2$. We wish to compare a PC of $\mathbf{XU}_1$ with a PC of $\mathbf{XU}_2$. If $\mathbf{U}_1 = \mathbf{I}_p$, the comparison involves a PC of $\mathbf{X}$ and a PC of some transformation of $\mathbf{X}$.

The comparison will be in terms of covariances ($\frac{1}{n}\mathbf{I}_n$-inner products in $\mathbb{R}^n$) and correlations (cosines of the $\frac{1}{n}\mathbf{I}_n$-angles in $\mathbb{R}^n$) between the pair of 'natural length' PCs. Results are slightly tidier if we write the PCs using the $\mathbf{M}_k$-orthogonal Principal Axes in $\mathbb{R}^p$ of $\mathbf{X}$, where $\mathbf{M}_k = \mathbf{U}_k'\mathbf{U}_k$, $(k = 1, 2)$, although analogous results using the eigenvectors of $\mathbf{U}_k\mathbf{S}\mathbf{U}_k'$ are also easily obtained.

Theorem IV.9 *Let* $\mathbf{X} \in \mathbb{C}_{\mathrm{nxp}}$ *be an* $\mathrm{n}\times\mathrm{p}$ *column-centered data matrix and* $\mathbf{S}$ *its associated covariance matrix. Let* $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *be positive definite* $\mathrm{p}\times\mathrm{p}$ *matrices. Let* $\{\mathbf{a}_i\}_{i=1}^{p}$ *be* $\mathbf{M}_1$*-orthonormal eigenvectors of* $\mathbf{SM}_1$ *and* $\{\mathbf{b}_j\}_{j=1}^{p}$ $\mathbf{M}_2$*-orthonormal eigenvectors of* $\mathbf{SM}_2$*. Let* $\{\lambda_i\}_{i=1}^{p}$ *be the eigenvalues of* $\mathbf{SM}_1$ *and* $\{\mu_j\}_{j=1}^{p}$ *the eigenvalues of* $\mathbf{SM}_2$*. Then, we have:*

i). The covariance between the i-th PC of $\mathbf{X}$ *in an* $\mathbf{M}_1$*-metric PCA and the j-th PC of* $\mathbf{X}$ *in an* $\mathbf{M}_2$*-metric PCA is given by:*

$$\mathrm{cov}(\mathbf{XM}_1\mathbf{a}_i, \mathbf{XM}_2\mathbf{b}_j) = \mu_j < \mathbf{a}_i, \mathbf{b}_j >_{\mathbf{M}_1} = \lambda_i < \mathbf{a}_i, \mathbf{b}_j >_{\mathbf{M}_2} \qquad \text{(IV.3.13)}$$

ii). The correlation between the i-th PC of $\mathbf{X}$ *in an* $\mathbf{M}_1$*-metric PCA and the j-th PC of* $\mathbf{X}$ *in an* $\mathbf{M}_2$*-metric PCA is given by:*

$$\mathrm{corr}(\mathbf{XM}_1\mathbf{a}_i, \mathbf{XM}_2\mathbf{b}_j) = \sqrt{\tfrac{\mu_j}{\lambda_i}} < \mathbf{a}_i, \mathbf{b}_j >_{\mathbf{M}_1} = \sqrt{\tfrac{\lambda_i}{\mu_j}} < \mathbf{a}_i, \mathbf{b}_j >_{\mathbf{M}_2} \qquad \text{(IV.3.14)}$$

Proof.

i). We have: $\mathrm{cov}(\mathbf{XM}_1\mathbf{a}_i, \mathbf{XM}_2\mathbf{b}_j) = \mathbf{a}_i{}'\mathbf{M}_1\mathbf{SM}_2\mathbf{b}_j = \mu_j\mathbf{a}_i{}'\mathbf{M}_1\mathbf{b}_j$

or, alternatively, $= \lambda_i\mathbf{a}_i{}'\mathbf{M}_2\mathbf{b}_j$

ii). We have: $\mathrm{cov}(\mathbf{XM}_1\mathbf{a}_i, \mathbf{XM}_1\mathbf{a}_i) = \lambda_i\mathbf{a}_i{}'\mathbf{M}_1\mathbf{a}_i = \lambda_i$

and $\mathrm{cov}(\mathbf{XM}_2\mathbf{b}_j, \mathbf{XM}_2\mathbf{b}_j) = \mu_j$

Hence: $\mathrm{corr}(\mathbf{XM}_1\mathbf{a}_i, \mathbf{XM}_2\mathbf{b}_j) = \frac{\mu_j(\mathbf{a}_i{}'\mathbf{M}_1\mathbf{b}_j)}{\sqrt{\lambda_i\mu_j}} = \sqrt{\frac{\mu_j}{\lambda_j}}\mathbf{a}_i{}'\mathbf{M}_1\mathbf{b}_j$

or, alternatively, $= \frac{\lambda_i\mathbf{a}_i{}'\mathbf{M}_2\mathbf{b}_j}{\sqrt{\mu_j\lambda_i}} = \sqrt{\frac{\lambda_i}{\mu_j}}\mathbf{a}_i{}'\mathbf{M}_2\mathbf{b}_j$

□

Corollary *In the conditions of the Theorem, the following bounds apply:*

i). $|\mathrm{cov}(\mathbf{XM}_1\mathbf{a}_i, \mathbf{XM}_2\mathbf{b}_j)| \leq \min\left\{\mu_j\|\mathbf{b}_j\|_{\mathbf{M}_1}, \lambda_i\|\mathbf{a}_i\|_{\mathbf{M}_2}\right\}$

ii). $|\mathrm{corr}(\mathbf{XM}_1\mathbf{a}_i, \mathbf{XM}_2\mathbf{b}_j)| \leq \min\left\{\sqrt{\frac{\mu_j}{\lambda_i}}\|\mathbf{b}_j\|_{\mathbf{M}_1}, \sqrt{\frac{\lambda_i}{\mu_j}}\|\mathbf{a}_i\|_{\mathbf{M}_2}\right\}$

Proof. The direct consequence of applying the Cauchy-Schwarz inequality (valid for any inner product) to the expressions in the Theorem. □

The covariances and correlations between PCs in Theorem IV.9 are given in terms of the eigendecompositions of matrices $\mathbf{SM}_1$ and $\mathbf{SM}_2$ and either the $\mathbf{M}_1$-inner product or the $\mathbf{M}_2$-inner product. Again, knowledge of the original data matrix is not required, only of the covariance matrix $\mathbf{S}$ it generates.

Although not, strictly speaking, a result involving PCs, Theorem IV.9 incidentally gives an interesting relation between the individual eigenvalues of $\mathbf{SM}_1$ and $\mathbf{SM}_2$ (*i.e.*, of $\mathbf{U}_1\mathbf{S}\mathbf{U}_1'$ and $\mathbf{U}_2\mathbf{S}\mathbf{U}_2'$ for any factorizations $\mathbf{M}_k = \mathbf{U}_k'\mathbf{U}_k$ with $k = 1, 2$).

Corollary *Let* $\mathbf{S}$ *be a* p×p *covariance matrix and* $\mathbf{M}_1$,$\mathbf{M}_2$ *two* p×p *positive definite matrices. If* $\{(\lambda_i, \mathbf{a}_i)\}_{i=1}^{p}$ *and* $\{(\mu_j, \mathbf{b}_j)\}_{j=1}^{p}$ *are the eigenpairs of* $\mathbf{SM}_1$ *and* $\mathbf{SM}_2$, *then:*

$$\lambda_i < \mathbf{a}_i, \mathbf{b}_j >_{\mathbf{M}_2} = \mu_j < \mathbf{a}_i, \mathbf{b}_j >_{\mathbf{M}_1} \qquad \forall i, j = 1, ..., p \tag{IV.3.15}$$

Thus, the ratio of any two (non-zero) eigenvalues from the 'transformed covariance matrices' $\mathbf{U}_1\mathbf{S}\mathbf{U}_1'$, $\mathbf{U}_2\mathbf{S}\mathbf{U}_2'$ (for any factorizations $\mathbf{M}_1 = \mathbf{U}_1{}'\mathbf{U}_1$ and $\mathbf{M}_2 = \mathbf{U}_2{}'\mathbf{U}_2$) can be written in terms of the ratio of the (non-zero) $\mathbf{M}_1$- and $\mathbf{M}_2$-inner products of the corresponding eigenvectors, when re-expressed in the original coordinates. In particular, the two covariance matrices can only share a common eigenvalue if the $\mathbf{M}_1$-angle and the $\mathbf{M}_2$-angle between their corresponding eigenvectors (in the original coordinates) are equal. This can happen even when $\mathbf{M}_1 \neq \mathbf{M}_2$.

As with the main Theorem, choosing one of the metric matrices to be the identity will enable us to compare the individual eigenvalues of $\mathbf{S}$ and $\mathbf{USU}'$. This seems a stronger result than the one obtained in Theorem IV.7 since it gives an explicit formula for individual eigenvalues, but it also relies on stronger requirements. Here we assume that the eigenvectors of both $\mathbf{S}$ and $\mathbf{USU}'$ are known, whereas in the previous Subsection only knowledge of the eigenpairs of $\mathbf{S}$ was required.

IV.4 Comparing the PCAs of two different data transformations

The problem of relating the PCA of a data matrix $\mathbf{X}$ with that of some linear transformation of $\mathbf{X}$ will now be addressed from a different point of view. Rather than seeking to understand *how* a linear transformation affects the analysis, we shall use simple indicators to *measure the scale of the resulting changes* to the PCA.

The discussion will be carried out in a more general framework, where the PCAs of two different linear transformations of the same data matrix are compared. Taking one of the transformation matrices to be the identity will provide the comparison described above. This more general framework will, for example, enable us to globally assess how using one of the alternative methods of obtaining dimensionless data suggested by Gower (1966) [19] (see also Subsection I.5.3) will compare with the traditional approach of a Correlation Matrix PCA.

It is necessary to clarify from the outset what precisely is to be compared. In particular:

i). Are we interested in a global comparison of all aspects of both PCAs, including all corresponding PAs in $\mathbb{R}^n$ and $\mathbb{R}^p$, as well as their relative importance ('variance accounted for')? Or are we interested only in what happens in $\mathbb{R}^n$ or what happens in $\mathbb{R}^p$?

ii). If the comparisons involve the vectors in $\mathbb{R}^p$, do we want to revert to the original coordinate system (*i.e.*, apply the inverse transformation) before comparing?

iii). If we are interested only in a subset of either or both sets of axes, is our main concern with the vectors themselves (as when we 'interpret' PCs) or in the subspaces which they span (as in low-dimensional representations of the data)?

The precise nature of the comparisons will obviously depend on the answers to the above questions.

IV.4.1 Comparing without inverting the transformations

When we are *not* concerned with reversing the linear transformation in $\mathbb{R}^p$ prior to the comparison, the similarity indices given in Section III.2 can be used to compare the PAs in either or both spaces. This corresponds to viewing the transformation matrices $\mathbf{U}$ as affecting vectors rather than changing the coordinate system.

In particular, if $\mathbf{X}$ is a column-centered $n \times p$ data matrix with covariance matrix $\mathbf{S}$ and if $\mathbf{U}_1$ and $\mathbf{U}_2$ are two $p \times p$ matrices of transformation of the data, we have:

i). **global similarity index for the PCAs of $\mathbf{XU}_1{}'$ and $\mathbf{XU}_2{}'$**

$$s_g(\mathbf{XU}_1{}', \mathbf{XU}_2{}') = \cos(\mathbf{XU}_1{}', \mathbf{XU}_2{}') = \frac{\mathrm{tr}(\mathbf{U}_1\mathbf{SU}_2{}')}{\sqrt{\mathrm{tr}(\mathbf{U}_1\mathbf{SU}_1{}') \cdot \mathrm{tr}(\mathbf{U}_2\mathbf{SU}_2{}')}} \qquad \text{(IV.4.1)}$$

which, as noted in (III.2.2), can be interpreted in terms of the variances of the columns of $\mathbf{XU}_1{}'$ and $\mathbf{XU}_2{}'$ and the cross-covariances between their corresponding columns.

ii). **similarity index for the PCs of $\mathbf{XU}_1{}'$ and $\mathbf{XU}_2{}'$**

$$\begin{aligned} s_n(\mathbf{XU}_1{}', \mathbf{XU}_2{}') &= \cos(\mathbf{XU}_1{}'\mathbf{U}_1\mathbf{X}', \mathbf{XU}_2{}'\mathbf{U}_2\mathbf{X}') \\ &= \frac{\mathrm{tr}(\mathbf{U}_1\mathbf{SU}_2{}' \cdot \mathbf{U}_2\mathbf{SU}_1{}')}{\sqrt{\mathrm{tr}\left[(\mathbf{U}_1\mathbf{SU}_1{}')^2\right] \cdot \mathrm{tr}\left[(\mathbf{U}_2\mathbf{SU}_2{}')^2\right]}} \end{aligned} \qquad \text{(IV.4.2)}$$

iii). **similarity index for the PAs in $\mathbb{R}^p$ of $\mathbf{XU}_1{}'$ and$\mathbf{XU}_2{}'$**

$$\begin{aligned} s_p(\mathbf{XU}_1{}', \mathbf{XU}_2{}') &= \cos(\mathbf{U}_1\mathbf{SU}_1{}', \mathbf{U}_2\mathbf{SU}_2{}') \\ &= \frac{\mathrm{tr}(\mathbf{U}_1\mathbf{SU}_1{}' \cdot \mathbf{U}_2\mathbf{SU}_2{}')}{\sqrt{\mathrm{tr}\left[(\mathbf{U}_1\mathbf{SU}_1{}')^2\right] \cdot \mathrm{tr}\left[(\mathbf{U}_2\mathbf{SU}_2{}')^2\right]}} \end{aligned} \qquad \text{(IV.4.3)}$$

For some special types of transformation matrices $\mathbf{U}_1$,$\mathbf{U}_2$, the global similarity index can be discussed further.

IV.4.2 Diagonal transformation matrices

The class of diagonal transformation matrices is of particular interest since they arise when (separate) changes of the scales of measurement are introduced for each of the p variables. They are therefore natural transformations to consider whenever these scales are to some extent arbitrary. In such cases, the behaviour of the global similarity index is noteworthy.

Denoting any two diagonal transformation matrices by $\mathbf{D}_1$, $\mathbf{D}_2$, their diagonal elements by $d_{ii}^{[k]}$ $(k = 1, 2; i = 1, ...p)$ and the diagonal elements of the original covariance matrix $\mathbf{S}$ by s_{ii} $(i = 1, ..., p)$, equation (IV.4.1) becomes:

$$s_g(\mathbf{X}\mathbf{D}_1, \mathbf{X}\mathbf{D}_2) = \frac{\sum_{i=1}^{p} s_{ii} d_{ii}^{[1]} d_{ii}^{[2]}}{\sqrt{\left[\sum_{i=1}^{p} s_{ii} \left(d_{ii}^{[1]}\right)^2\right] \cdot \left[\sum_{i=1}^{p} s_{ii} \left(d_{ii}^{[2]}\right)^2\right]}}$$

$$\Longleftrightarrow s_g(\mathbf{X}\mathbf{D}_1, \mathbf{X}\mathbf{D}_2) = \cos(\mathbf{D}_1\mathbf{s}, \mathbf{D}_2\mathbf{s}) \qquad \text{(IV.4.4)}$$

where $\mathbf{s}$ is the p -dimensional vector of standard deviations $\{\sqrt{s_{ii}}\}_{i=1}^{p}$. The r.h.s of (IV.4.4) is the cosine of an angle between the two vectors in $\mathbb{R}^p$.

Any given $n \times p$ data matrix $\mathbf{X}$ can be mapped onto the vector $\mathbf{s} \in \mathbb{R}^p$ of the standard deviations of its p columns. Equation (IV.4.4) tells us that when $\mathbf{X}$ is transformed by two *diagonal* matrices $\mathbf{D}_1$, $\mathbf{D}_2$, the global similarity index for the

PCAs of $\mathbf{XD}_1$ and $\mathbf{XD}_2$ boils down to the cosine of the angle between the images of s under $\mathbf{D}_1$ and $\mathbf{D}_2$.

Interestingly, the off-diagonal elements of $\mathbf{S}$ — the covariances between the variables — play no role. Equation (III.2.2) for the global similarity index had already given us a foretaste of this, by using only the covariances of the *corresponding* columns of the two data matrices.

It can be argued that the absence of the covariances between the p original variables in (IV.4.4) implies that the global similarity index is 'filtering out' the *common underlying pattern of correlations* which characterizes the two transformed covariance matrices $\mathbf{D}_1\mathbf{SD}_1$ and $\mathbf{D}_2\mathbf{SD}_2$. In fact, diagonal transformation matrices only affect the *size* of the p vectors, not their relative positions. Thus, equation (IV.4.4) is focusing only on the crucial difference between such matrices: their diagonal elements. In so doing, however, the global similarity index is also displaying an interesting invariance property between *different* data matrices. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are data matrices (possibly with a different number of rows) whose p corresponding columns have equal variances, then the global similarity index between $\mathbf{X}_1\mathbf{D}_1$ and $\mathbf{X}_1\mathbf{D}_2$ equals that between $\mathbf{X}_2\mathbf{D}_1$ and $\mathbf{X}_2\mathbf{D}_2$.

What practical conclusions can be drawn from equation (IV.4.4)? Say $\mathbf{X}$ is an $n \times p$ data matrix whose columns have standard deviations given by vector s. Equation (IV.4.4) tells us that $s_g(\mathbf{X}, \mathbf{XD})$ is, for any diagonal matrix $\mathbf{D}$, given by the cosine of the angle between **s** and its image under **D**. Thus, diagonal matrices **D** which leave the direction of **s** approximately unchanged will produce 'globally similar' PCAs. This is in accordance with the conventional wisdom from practical applications, which tells us that comparatively minor changes in the relative sizes of the variables' standard deviations will not produce major differences in the PCAs, whereas sizable changes may radically alter the results. A further important application of (IV.4.4) will be considered in the next Subsection.

The similarity indices in either $\mathbb{R}^n$ or $\mathbb{R}^p$ are not, unfortunately, as prone as the global similarity index to a general discussion of this kind.

IV.4.3 A global comparison of Covariance and Correlation Matrix PCA

A Correlation Matrix PCA corresponds to taking the diagonal transformation matrix $\mathbf{D}_S^{-\frac{1}{2}}$ whose diagonal elements are the reciprocal standard deviations of the p variables. The discussion in the previous Subsection thus paves the way for a global comparison of the Covariance and Correlation Matrix PCAs of any data matrix.

Equation (IV.4.4) implies that *the global similarity index between a Covariance Matrix PCA and the corresponding Correlation Matrix PCA* can be written as:

$$s_g(\mathbf{X}, \mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}) = \cos(\mathbf{1}_p, \mathbf{s}) = \frac{||\mathbf{s}||_1}{\sqrt{p} \cdot ||\mathbf{s}||_2} \qquad \text{(IV.4.5)}$$

where $\mathbf{1}_p$ is the p-dimensional vector of ones and $||\cdot||_k$ $(k = 1, 2)$ are the ℓ_k-norms of the vector $\mathbf{s}$ of the standard deviations of the p variables. Thus, the global similarity index is given by the cosine of the angle between the vector $\mathbf{s}$ of standard deviations and the bisector of the positive orthant of $\mathbb{R}^p$.

This expression bears an interesting resemblance to equation (III.3.2) for the cosine of the angle in $\mathbb{M}_p$ between the covariance matrix $\mathbf{S}$ and $\mathbf{I}_p$, with the vector $\mathbf{s}$ of standard deviations now replacing the vector $\boldsymbol{\lambda}$ of the eigenvalues of $\mathbf{S}$. The discussion which followed equation (III.3.2) also applies here, since $\mathbf{s}$ can be replaced by $\alpha\mathbf{s}$ for any positive scalar α without affecting the value of (IV.4.5). Thus, if we take α to be the reciprocal of the sum of standard deviations, we can re-express $s_g(\mathbf{X}, \mathbf{X}\mathbf{D}_S^{-\frac{1}{2}})$ in terms of the variance v of the 'relative standard deviations':

$$s_g(\mathbf{X}, \mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}) = \frac{1}{\sqrt{1 + p^2 \cdot v}} \tag{IV.4.6}$$

As with equation (III.3.4), data matrices may give values of (IV.4.6) between 1 (if $v = 0$, *i.e.*, if all p variables have a common variance) and $\frac{1}{\sqrt{p}}$ (this value cannot be exactly attained for rank p data matrices, but arbitrarily close values will result from matrices for which the variance of one variable dominates those of all the remaining variables). The larger the variance of the 'relative standard deviations' $\left\{\frac{\sqrt{s_{ii}}}{\sum_{j=1}^{p}\sqrt{s_{jj}}}\right\}_{i=1}^{p}$, the larger the angle between $\mathbf{s}$ and $\mathbf{1}_p$ and, consequently, the smaller the value of the similarity index. This agrees with the conventional wisdom on the issue.

The analogy between the expressions for $\cos(\mathbf{1}_p, \mathbf{s})$ in $\mathbb{R}^p$ and $\cos(\mathbf{I}_p, \mathbf{S})$ in the

PSD cone is not a coincidence. Once the off-diagonal elements of $\mathbf{S}$ are excluded from the picture in equation (IV.4.4), the mappings between $\mathbf{XD}_k$ and $\mathbf{D}_k\mathbf{s}$ described in Section IV.2 can be replaced with mappings between $\mathbf{XD}_k$ and diagonal matrices with the elements of vector $\mathbf{D}_k\mathbf{s}$ as their diagonal elements. Correlation matrices will then be mapped onto $\mathbf{I}_p$ and the Covariance Matrix $\mathbf{S}$ onto a diagonal **p.s.d.** matrix $\mathbf{\Delta}_S$ whose diagonal elements are the coefficients of vector $\mathbf{s}$. The eigenvalues of the diagonal matrices being the diagonal elements themselves, the expressions for $\cos(\mathbf{I}_p, \mathbf{\Delta}_S)$ and $\cos(\mathbf{1}_p, \mathbf{s})$ will coincide.

The above discussion can also be, to a large extent, adapted to discuss the global similarity of a Correlation Matrix PCA with the PCAs involving alternative rescalings of the original variables which produce dimensionless data. Using Gower's (1966) [19] suggestion that the (we assume positive) *variable means* may be used to divide the data in place of standard deviations, we would be mapping the vector $\mathbf{s} \in \mathbb{R}^p$ of standard deviations onto the vector whose i-th element is $\frac{\sqrt{s_{ii}}}{\bar{x}_i}$. This is the vector of **coefficients of variation** of the p variables, a measure of relative dispersion widely used in Sampling Theory (Deming (1950) [10, (pg.75)]). The cosine of the angle formed by this vector with both $\mathbf{s}$ and $\mathbf{1}_p$ will indicate the global similarity of this alternative with both the Covariance and the Correlation Matrix PCAs of the data. This particular alternative, though, has the disadvantage of *not* being insensitive to additive changes of scale, since we are naturally talking about the mean of the uncentered variables. An alternative suggestion by Gower is to use the cube root of each variable's third (we assume central) moment. The elements of the vector **Ds** in this case are the reciprocals of the sixth root of each variable's **skewness coefficients** β_1 – or the cube root of the skewness coefficients γ_1 (Kendall *et al.* (1987) [36, (pg.106)]). In general, the n-th root of the n-th central moment can be used instead. For $n = 4$, the vector **Ds** will give the reciprocals of the fourth root of the **kurtosis coefficients** β_2 (Kendall *et al.* (1987) [36, (pg.106)]).

Examples of comparisons between Covariance and Correlation Matrix PCAs will be given at the end of Section IV.5.

A discussion along these lines involving the similarity index in $\mathbb{R}^n$ will be given in Subsection VI.2.4.

IV.4.4 Other comparisons

We now take a brief look at other comparisons which might be of interest.

Comparing in the original coordinate system

So far we have assumed that we are interested in comparing the PCAs of two different transformations of a data matrix $\mathbf{X}$, without reverting the Principal Axes in $\mathbb{R}^p$ back to the original system of coordinates.

However, we might prefer a comparison (either global or in $\mathbb{R}^p$ alone) where the effects of the transformation in $\mathbb{R}^p$ are reversed prior to the comparison. This is possible, but unfortunately in a way which is neither elegant nor practical, in that it will require prior computation of the PCA.

In Section IV.2 it was seen that the effect of applying the inverse transformation $(\mathbf{U}')^{-1}$ to the (conventional) Principal Axes in $\mathbb{R}^p$ of matrix $\mathbf{XU}'$ was to obtain the $\mathbf{M}$-orthogonal PAs in $\mathbb{R}^p$ of $\mathbf{X}$ (with $\mathbf{M} = \mathbf{U}'\mathbf{U}$), that is, the right singular vectors of an $(\frac{1}{n}\mathbf{I}_n, \mathbf{M})$-orthogonal SVD of $\mathbf{X}$. Thus, our starting point for the comparison we are now envisaging can be the $(\frac{1}{n}\mathbf{I}_n, \mathbf{M}_1)$- and $(\frac{1}{n}\mathbf{I}_n, \mathbf{M}_2)$-orthogonal SVDs of $\mathbf{X}$, as given by equation (IV.1.2):

$$\mathbf{X} = \boldsymbol{\beta}_1 \boldsymbol{\tau}_1 \mathbf{B}_1' = \sum_{i=1}^{r} \tau_i^{[1]} \beta_i^{[1]} (\mathbf{b}_i{}^{[1]})' \qquad \text{(IV.4.7)}$$

$$\mathbf{X} = \boldsymbol{\beta}_2 \boldsymbol{\tau}_2 \mathbf{B}_2' = \sum_{j=1}^{r} \tau_j^{[2]} \beta_j^{[2]} (\mathbf{b}_j{}^{[2]})' \qquad \text{(IV.4.8)}$$

Although the inner product of $\mathbf{X}$ with itself, using each of these two alternative decompositions, gives an expression which is superficially similar to (III.1.3), there is an important underlying difference. The inner products $(\mathbf{b}_i{}^{[1]})'(\mathbf{b}_j{}^{[2]})$ are no longer cosines of the angles between these vectors of $\mathbb{R}^p$ since they are not of unit $\mathbf{I}_p$-norm. Nor is there an obvious alternative inner product with which to work, since the two sets of vectors in $\mathbb{R}^p$ are of unit norm for *different* norms, resulting from the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$.

A possible alternative involves re-scaling the vectors $\{\mathbf{b}_i{}^{[1]}\}_{i=1}^{r}$ and $\{\mathbf{b}_j{}^{[2]}\}_{j=1}^{r}$ in (IV.4.7) and (IV.4.8) to have (conventional) unit norm. This is obtained by replacing matrix $\mathbf{X}$ with the matrices $\mathbf{XC}_k = \mathbf{X}(\mathbf{B}_k')^{-1}(\mathbf{B}_k'\mathbf{B}_k \circ \mathbf{I}_p)^{-1/2}\mathbf{B}_k'$, $(k = 1, 2)$, where 'o' denotes the Hadamard (entry-wise) matrix product. Skipping the details, we obtain a similarity index akin to s_g but involving the original units of measurement, by considering:

$$\cos(\mathbf{XC}_1, \mathbf{XC}_2) = \sum_{i=1}^{r}\sum_{j=1}^{r} \sqrt{\pi_i^{[1]}\pi_j^{[2]}} \cos(\boldsymbol{\beta}_i^{[1]}\boldsymbol{\beta}_j^{[2]}) \cos(\mathbf{b}_i^{[1]}\mathbf{b}_j^{[2]}) \qquad \text{(IV.4.9)}$$

where $\pi_i^{[k]} = \left(\frac{\tau_i^{[k]}}{||\boldsymbol{\tau}^{[k]}||}\right)^2$ is the i-th relative eigenvalue of $\mathbf{M}_k\mathbf{SM}_k$, $(k = 1, 2)$.

A similar analogue of the similarity index in $\mathbb{R}^p$ will (again, omitting a detailed derivation) be:

$$\frac{< \mathbf{S}_1, \mathbf{S}_2 >_{(\mathbf{I}_p, \mathbf{I}_p)}}{||\mathbf{S}||_{(M_1,M_1)} \cdot ||\mathbf{S}||_{(M_2,M_2)}} \qquad \text{(IV.4.10)}$$

where $\mathbf{S}$ denotes the original covariance matrix of $\mathbf{X}$, $< \cdot, \cdot >_{(\mathbf{I}_p,\mathbf{I}_p)}$ denotes the standard inner product in $\mathbb{M}_p$, $|| \cdot ||_{(M_k,M_k)}$, $(k = 1, 2)$, denotes the norm defined by the $(\mathbf{M}_k, \mathbf{M}_k)$-inner product (IV.1.3) and $\mathbf{S}_1$,$\mathbf{S}_2$ are the covariance matrices of $\mathbf{XC}_1$ and $\mathbf{XC}_2$, respectively.

Unfortunately, these similarity indices (IV.4.9) and (IV.4.10) are much less interesting than their counterparts (IV.4.1) and (IV.4.3), because they require prior knowledge of the $(\frac{1}{n}\mathbf{I}_n, \mathbf{M}_k)$-orthogonal SVDs of $\mathbf{X}$. The PCAs of $\mathbf{X}$ (in both metrics) will actually have to be carried out if these similarity indices are to be computed.

Low-rank comparisons

The comparisons considered so far have involved the full decompositions of the matrices. However, we are frequently interested in considering only partial results of a PCA, usually involving the first few PCs or Principal Axes in $\mathbb{R}^p$.

The obvious strategy to follow, in such cases, is the replacement of the data matrices $\mathbf{X}$ or $\mathbf{XU}'$ by the sum of those terms of their SVDs on which we wish to focus. In the case of wishing to compare the first k PCs of $\mathbf{X}$ and $\mathbf{XU}'$, for example, we replace $\mathbf{X}$ and $\mathbf{XU}'$ with their least squares rank k approximation – the first k terms in their SVD (Greenacre (1984) [21, (pg.343)]) – and compute the similarity index in $\mathbb{R}^n$ for those approximations.

The above approach will also apply to comparisons involving the original coordinates in $\mathbb{R}^p$, that is, the $(\frac{1}{n}\mathbf{I}_n, \mathbf{M})$-orthogonal SVD of $\mathbf{X}$. If the latter is given by $\mathbf{X} = \sum_{j=1}^{r} \tau_j \boldsymbol{\beta}_j \mathbf{b}_j{}'$, the first k terms of this sum are the best $\mathbf{M}$-inner product least squares approximation to $\mathbf{X}$, of rank k (Greenacre (1984) [21, (pg.345)]). This approximation is, in general, not the rank k approximation to $\mathbf{X}$ using the standard inner product in $\mathbb{R}^p$.

In both cases, the use of the previously defined similarity indices on the rank k approximations to the data matrices provides indicators for the degree of similarity of the *first* k PCs and/or PAs in $\mathbb{R}^p$.

An alternative form of partial comparison involves the use of Yanai's G.C.D. (III.2.5). With the G.C.D., attention is focused on the subspaces of $\mathbb{R}^n$ or $\mathbb{R}^p$ spanned by any k PCs or PAs in $\mathbb{R}^p$ from each transformed data matrix. No explicit use is made of the relative importance of each axis, although such considerations can be incorporated into the choice of which k vectors are to be compared.

Say $\mathbf{XU}_1{}'$ and $\mathbf{XU}_2{}'$ are two transformations of the data matrix $\mathbf{X}$, whose first k PCs are the columns of matrices $\boldsymbol{\alpha}_k^{[1]}$ and $\boldsymbol{\alpha}_k^{[2]}$. If we wish to compare the subspaces of $\mathbb{R}^n$ spanned by those two subsets of k PCs, we can take:

$$\mathrm{GCD}(\boldsymbol{\alpha}_k^{[1]}, \boldsymbol{\alpha}_k^{[2]}) = \frac{\mathrm{tr}(\mathbf{P}_{\boldsymbol{\alpha}_k^{[1]}}\mathbf{P}_{\boldsymbol{\alpha}_k^{[2]}})}{k}$$

where $\mathbf{P}_{\boldsymbol{\alpha}_k^{[j]}} = \boldsymbol{\alpha}_k^{[j]}\left[(\boldsymbol{\alpha}_k^{[j]})'(\boldsymbol{\alpha}_k^{[j]})\right]^{-1}(\boldsymbol{\alpha}_k^{[j]})' \quad (j = 1, 2)$.

Values close to 1 will indicate that the two 'Principal Subspaces' spanned by the columns of $\boldsymbol{\alpha}_k^{[1]}$ and $\boldsymbol{\alpha}_k^{[2]}$ nearly coincide.

As with the indicators from the previous Subsection, calculating these indicators requires prior knowledge of the SVDs of the matrices which are being compared.

IV.5 Multiple comparisons of PCAs

So far, all comparisons have been between *pairs* of matrices. The extension to comparisons involving more than two matrices simultaneously will now be addressed.

We might, for example, have t different transformation matrices $\{\mathbf{U}_i\}_{i=1}^t$ and wish to compare some aspect of the PCAs of the t transformed data matrices $\{\mathbf{XU'}_i\}_{i=1}^t$. It would be rewarding to find, for example, that one or several subsets of the t transformations produced results that could be classified as similar in some respect. Alternatively, we might wish to find some ordering of the transformations, with respect to a given criterion.

These questions are raising the issue of whether classical scaling and/or clustering techniques, as applied to sets of matrices, may prove useful for multiple comparisons of PCAs.

IV.5.1 The matrix of matrix correlations

In a general setting, we have t matrices $\{\mathbf{A}_i\}_{i=1}^t$ in some matrix space like $\mathbb{M}_{n\times p}$ or $\mathbb{M}_p$. Having defined an inner product — hence angles and distances — in the matrix space, it is possible to consider applying scaling or clustering methods to the t matrices.

A general discussion of such methods can be found in many Multivariate Statistics textbooks, such as Mardia, Kent and Bibby (1979) [42, (Chapters 13 and 14)], Kendall (1980) [35, (Chapter 3)] or Chatfield and Collins (1980) [8, (Chapters 10 and 11)].

The starting point for such methods is usually either a matrix of distances or a matrix of similarities between all pairs of the t entities which are being considered.

From our previous discussion, it is clear that for the purposes of comparing PCAs, it is more directly meaningful to start with the matrix of similarities defined by taking the correlations (cosines of the angles in the matrix space) between each pair of the

t matrices. We thus consider the **matrix of matrix correlations:**

$$\mathbf{\Phi} = \begin{pmatrix} 1 & \cos(\mathbf{A}_1, \mathbf{A}_2) & \cos(\mathbf{A}_1, \mathbf{A}_3) & \cdots & \cos(\mathbf{A}_1, \mathbf{A}_t) \\ \cos(\mathbf{A}_2, \mathbf{A}_1) & 1 & \cos(\mathbf{A}_2, \mathbf{A}_3) & \cdots & \cos(\mathbf{A}_2, \mathbf{A}_t) \\ \cos(\mathbf{A}_3, \mathbf{A}_1) & \cos(\mathbf{A}_3, \mathbf{A}_2) & 1 & \cdots & \cos(\mathbf{A}_3, \mathbf{A}_t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cos(\mathbf{A}_t, \mathbf{A}_1) & \cos(\mathbf{A}_t, \mathbf{A}_2) & \cos(\mathbf{A}_t, \mathbf{A}_3) & \cdots & 1 \end{pmatrix} \qquad \text{(IV.5.1)}$$

If the t matrices $\{\mathbf{A}_i\}_{i=1}^t$ are column-centered $n{\times}p$ data matrices, the elements of $\mathbf{\Phi}$ are the global similarity indices for each pair of matrices. If we are working with $p{\times}p$ covariance matrices, the elements of $\mathbf{\Phi}$ are similarity indices in $\mathbb{R}^p$. The elements of $\mathbf{\Phi}$ will be similarity indices in $\mathbb{R}^n$ if we take $\mathbf{A}_i = \mathbf{X}_i\mathbf{X}_i'$, for some set of t column-centered data matrices $\{\mathbf{X}_i\}_{i=1}^t$. And they will be values of Yanai's G.C.D. if the matrices $\{\mathbf{A}_i\}_{i=1}^t$ are matrices of orthogonal projections, as described in Subsection IV.4.4.

The matrix $\mathbf{\Phi}$ will be at the heart of the comparisons in this Section. It should be stressed that $\mathbf{\Phi}$ is a matrix of inner products (in the appropriate matrix space) between the standardized matrices $\frac{\mathbf{A}_i}{||\mathbf{A}_i||}$, $(i = 1, ..., t)$. All matrices of this sort, for any t elements of any inner product linear space, are known as **Gram matrices** and are necessarily positive semi-definite, since the quadratic form $\mathbf{x}'\mathbf{\Phi}\mathbf{x}$, $(\mathbf{x} \in \mathbb{R}^t)$, is the norm squared of the matrix $\sum_{i=1}^t x_i \frac{\mathbf{A}_i}{||\mathbf{A}_i||}$ (with x_i a scalar and the i-th element of $\mathbf{x}$), hence necessarily non-negative (see also Horn and Johnson (1985) [25, (pg.407)]). Furthermore, $\mathbf{\Phi}$ *is positive definite if and only if the* t *matrices* $\{\mathbf{A}_i\}_{i=1}^t$ *are linearly* *independent*, by a similar reasoning to the above.

It is sometimes required that measures of similarity between different entities take values between zero and one (see, for example, Seber (1984) [60, (pg.357)]). As we saw in Section III.2, or directly from equation (III.3.1), Yanai's G.C.D. and the similarity indices in $\mathbb{R}^p$ and $\mathbb{R}^n$ always take values between 0 and 1. Not so the global similarity indices, which can also take negative values as can be seen by considering for example $s_g(\mathbf{A}, -\mathbf{A})$ for any matrix $\mathbf{A}$. We shall not be unduly concerned with this problem, essentially because the specific methods which will be described in this Section do not require that all elements of the matrix $\mathbf{\Phi}$ be non-negative. It can also be argued that for most plausible comparisons between data matrices, we would expect the global similarity index to be positive (as is the case with the examples in this Section).

IV.5.2 A PCO of PCAs

A first method of multiple comparisons of PCAs consists of replacing each of the t matrices $\frac{\mathbf{A}_i}{||\mathbf{A}_i||}$ with a point in $\mathbb{R}^t$, in such a way that the distances between such points and the angles between the vectors which they define with the origin, correspond to those between the matrices $\left\{\frac{\mathbf{A}_i}{||\mathbf{A}_i||}\right\}_{i=1}^{t}$, in the matrix space to which they belong. If it is possible to give a good low-dimensional approximation to the resulting configuration of t points in $\mathbb{R}^t$, then a *graphical visualization* of the relations between the t matrices is possible.

The approach which has just been described corresponds to a **Principal Coordinate Analysis (PCO)** of the similarity matrix $\mathbf{\Phi}$ (see Gower (1966) [19]). If the spectral decomposition of the $t \times t$ matrix $\mathbf{\Phi}$ is given by:

$$\mathbf{\Phi} = \mathbf{B \Lambda B'}$$

then the t points that will represent the matrices $\left\{\frac{\mathbf{A}_i}{||\mathbf{A}_i||}\right\}_{i=1}^{t}$ are the t rows of matrix:

$$\mathbf{C} = \mathbf{B \Lambda}^{\frac{1}{2}} \tag{IV.5.2}$$

It will be noted that the t columns of $\mathbf{C}$ are the eigenvectors of $\mathbf{\Phi}$, 'scaled' so that their length squared equals their corresponding eigenvalue. It will also be noted that equation (IV.5.2) is an SVD of matrix $\mathbf{C}$, with the right singular vectors given by the columns of the identity matrix $\mathbf{I}_t$.

That the rows of matrix $\mathbf{C}$ have the desired properties is readily checked. The inner product of the i-th and j-th rows of $\mathbf{C}$ is the (i, j)-th element of matrix $\mathbf{CC'} = \mathbf{B \Lambda B'} = \mathbf{\Phi}$ and therefore equals the inner product of $\frac{\mathbf{A}_i}{||\mathbf{A}_i||}$ and $\frac{\mathbf{A}_j}{||\mathbf{A}_j||}$. In particular *each row of* $\mathbf{C}$ *is a unit-norm vector in* $\mathbb{R}^t$ *and the cosines of the angles between such vectors are the cosines of the angles in the matrix space between matrices* $\mathbf{A}_i$ *and* $\mathbf{A}_j$. The smaller the angle in $\mathbb{R}^t$ between any two such vectors, the greater the similarity (global, in $\mathbb{R}^n$ or in $\mathbb{R}^p$, in terms of the subspaces spanned by each kind of Principal Axes, according to the nature of the matrices $\{\mathbf{A}_i\}_{i=1}^{t}$) between the PCAs of the matrices which they correspond to.

The next step consists in seeking a q-dimensional approximation ($q < t$) to the configuration of t points in $\mathbb{R}^t$ given by the rows of matrix $\mathbf{C}$. This problem is a specific application of the general problem of finding a 'best' rank q approximation to $\mathbf{\Phi}$. If we replace equation (IV.5.2) with:

$$\mathbf{C}_q = \mathbf{B}_q \mathbf{\Lambda}_q^{\frac{1}{2}} \tag{IV.5.3}$$

where $\mathbf{C}_q, \mathbf{B}_q$ are txq matrices consisting of the first q columns of $\mathbf{C}$ and $\mathbf{B}$, and where $\mathbf{\Lambda}_q$ is the $q \times q$ submatrix of $\mathbf{\Lambda}$'s first q rows and columns, then $\mathbf{\Phi}_q = \mathbf{C}_q \mathbf{C}_q'$ is the best rank q approximation to $\mathbf{\Phi}$, in the sense that it minimizes:

$$\|\mathbf{\Phi} - \mathbf{A}\|^2 = \sum_{i=1}^{t} \sum_{j=1}^{t} (\phi_{ij} - a_{ij})^2$$

for any choice of rank q matrix $\mathbf{A} \in \mathbb{M}_t$ (see Subsection I.3.2). The t rows of matrix $\mathbf{C}_q$ represent points in $\mathbb{R}^q$. What has just been said is that they are a scatter of t points in $\mathbb{R}^q$ whose inner products are, globally speaking, the 'best' possible approximation to the inner products between the rows of $\mathbf{C}$. The scatter in $\mathbb{R}^q$ is therefore a 'best' q-dimensional configuration to represent the relations between the matrices $\{\mathbf{A}_i\}_{i=1}^{t}$.

It will be noted that in this presentation of PCO, the double-centering of matrix $\mathbf{\Phi}$ which usually precedes its spectral decomposition has been omitted. Double-centering was originally introduced by Gower (1966) [19, (pgs.328-329)] on the following grounds: we are interested in distances squared, as calculated from the similarity matrix $\mathbf{\Phi}$ by $d_{ij}^2 = \phi_{ii} + \phi_{jj} - 2\phi_{ij}$; these distances remain unchanged if a common constant is added to all elements of $\mathbf{\Phi}$; although the absolute magnitude of $\mathbf{\Phi}$'s elements is irrelevant, in the sense just described, they strongly influence the nature of $\mathbf{\Phi}$'s first eigenpairs and therefore any subsequent low-dimensional graphical representation; we might as well then remove such arbitrary effects by double-centering. But such considerations are not appropriate in our context. We are concerned with the association matrix itself, not with distances. Changes to the elements of $\mathbf{\Phi}$, as described above, would destroy their interpretation as the cosines of the angles between the matrices $\{\mathbf{A}_i\}_{i=1}^{t}$.

Geometrically, the case against doubly-centering $\mathbf{\Phi}$ is easily visualized. As Gower points out, the dimension-reduction stage of PCO corresponds to a PCA of the matrix $\mathbf{C}$, viewed as an $n \times p$ data matrix with $n = p = t$. The t columns of $\mathbf{C}$ are thus viewed as the 'variables'. Column-centering the 'data matrix' $\mathbf{C}$ corresponds to taking $\mathbf{P}_t\mathbf{C}$ with $\mathbf{P}_t = \mathbf{I}_t - \frac{1}{t}\mathbf{1}_t\mathbf{1}_t{}'$, the equivalent of the orthogonal projection matrix used in (I.1.1). The doubly-centered matrix $\mathbf{P}_t\mathbf{\Phi}\mathbf{P}_t$ will thus correspond to the matrix $\mathbf{X}\mathbf{X}'$ in our standard PCA notation. The matrix $\mathbf{\Phi}$ is the equivalent matrix for 'non-centered data'. Hence, our option is between a standard (column-centered) PCA of $\mathbf{C}$ or a *non-centered PCA of* $\mathbf{C}$. It is known (see for example Jolliffe (1986) [32, (pg.227)]) that a non-centered PCA, in the row space, finds the successively best-fitting q-dimensional flats *through the origin* rather than, as with standard column-centered PCA, through the center of gravity of the n points in the row space. In other words, column-centering displaces the origin of the row space to the centroid of the n-point scatter. We do *not* want this to happen in our case since the 'natural' origin for the t rows of $\mathbf{C}$ plays an important role. The angles between vectors in $\mathbb{R}^t$, whose cosines correspond to the correlations between matrices $\{\mathbf{A}_i\}_{i=1}^t$ are defined using the origin as the meeting point. Displacing the origin destroys these angles and therefore the information we are interested in.

It should be added that the t rows of matrix $\mathbf{C}_q$ (IV.5.3) correspond to the coefficients of an orthogonal projection of the t rows ('individuals') of matrix $\mathbf{C}$ onto their q first row space PAs ('PAs in $\mathbb{R}^p$' would be rather unfortunate terminology in this case, since we are now working in $\mathbb{R}^t$). The latter are the first q vectors in the canonical basis of $\mathbb{R}^t$, whereas the 'natural length' PCs of $\mathbf{C}$ are the columns of $\mathbf{C}$ itself — as can be seen from the SVD of $\mathbf{C}$ given by equation (IV.5.2).

If a two-dimensional representation of the t-point configuration is judged to be adequate, a graphical depiction of the scatter can be envisaged. If necessary, several such graphical representations may be used to accommodate additional dimensions. In Subsections IV.5.4 and IV.5.5, examples of graphical representations are given.

Volle (1981) [67, (pgs.116-118)] discusses 2-D graphical representations of the p-point scatters in $\mathbb{R}^n$ for Correlation Matrix PCA. His observations are also useful in our context. Since the t rows of $\mathbf{C}$ are unit-norm vectors, they are all on the unit hypersphere of $\mathbb{R}^t$. *Any of these points which lies exactly on the 2-dimensional* *subspace chosen for a graphical representation must be on the unit circumference of this subspace.* Points with components in the dimensions that are being ignored will, when projected onto the 2-D representation, appear in the interior of the unit circle. *The closer a point is to the origin in the two-dimensional graph, the less faithful its representation.* The cosine of the angle in the 2-D graph between two points close to the unit circumference will provide an approximation for the correlation between the matrices which gave rise to those points. But as Volle points out [67, (pg.117)], even if only one of two points is close to the unit circumference, the correlation between the matrices which they represent can still be approximated from the two-dimensional graph. In fact, the cosine of the angle between two points $\mathbf{c}_i$,$\mathbf{c}_j$ in the 'full' scatter in $\mathbb{R}^t$ can be obtained by dropping the perpendicular from one point, say $\mathbf{c}_i$, onto the vector defined by the other (say $\mathbf{c}_j$) and then measuring the line segment from the origin to the foot of the perpendicular. But this projection of $\mathbf{c}_i$ can be obtained from the 2-D graph if $\mathbf{c}_j$ is faithfully represented, by again orthogonally projecting the 2-D image of $\mathbf{c}_i$ onto the vector defined by the 2-D image of $\mathbf{c}_j$. In fact, if $\mathbf{P}_{\mathbf{c}_j}$ is the orthogonal projector onto vector $\mathbf{c}_j$ and $\mathbf{P}_2$ is the orthogonal projector onto the subspace spanned by the first two vectors in the canonical basis of $\mathbb{R}^t$, then assuming $\mathbf{c}_j$ lies exactly on $\Re(\mathbf{P}_2)$ (which is the Principal Coordinate plane), we have $\mathbf{P}_2\mathbf{c}_j = \mathbf{c}_j$ and therefore:

$$\mathbf{P}_{\mathbf{c}_j}(\mathbf{P}_2\mathbf{c}_i) = \mathbf{c}_j(\mathbf{c}_j{}'\mathbf{c}_j)^{-1}\mathbf{c}_j{}'\mathbf{P}_2\mathbf{c}_i = \mathbf{c}_j(\mathbf{c}_j{}'\mathbf{c}_j)^{-1}(\mathbf{P}_2\mathbf{c}_j)'\mathbf{c}_i = \mathbf{P}_{\mathbf{c}_j}\mathbf{c}_i$$

It should be stressed that the above application of PCO is valid when comparing any set of (similarly sized) matrices $\{\mathbf{A}_i\}_{i=1}^{t}$. As mentioned in Section III.2, it can also be used when non-centered or doubly-centered PCAs are involved.

IV.5.3 Clustering

The procedure described in the previous Subsection is helpful in identifying subgroups of matrices with similar PCAs, as measured by the similarity indices. But it may prove unsatisfactory whenever a low-dimensional representation of the t-point scatter is inadequate. In such cases, the techniques which are traditionally grouped under the name of **Cluster Analysis** might turn out to be useful.

The purpose of Cluster Analysis is, in the words of Chatfield and Collins (1980) [8, (pg.212)] "to find the 'natural groupings', if any, of a set of individuals (or objects, or points, or units, or whatever)". For our purposes, the set of t matrices $\mathbf{A}_i$ is to be inspected and any 'natural groupings' should correspond to subsets of matrices whose PCAs are similar.

There are a host of clustering methods, which are discussed in some detail in, for example, Everitt (1980) [14]. The very existence of so many methods reflects the fact that no single approach has been found to perform convincingly in all circumstances. Often, different methods provide qualitatively different information as to the existence and nature of 'natural groupings'. In order to avoid a lengthy discussion of Cluster

Analysis — which is outside the scope of this Thesis — a single clustering technique will be used in the examples which are to be considered: the **nearest neighbour** or **single-link clustering**. This decision is to some extent influenced by the discussion in Section 11.6 of Chatfield and Collins (1980) [8], where both theoretical and practical advantages of the nearest-neighbour method are pointed out. It is also influenced by the simple nature of the method and the fact that it does not require that all values in the similarity matrix be non-negative.

Like other clustering techniques, the nearest-neighbour method is based on a matrix of similarities (or dissimilarities) between each pair of elements in the set which is being studied. Once again, the matrix $\mathbf{\Phi}$ of matrix correlations (IV.5.1) will provide the starting point, in our context. The t matrices are initially viewed as t different groups. The 2 most similar groups — as measured by the highest off-diagonal elements of $\mathbf{\Phi}$ — are then merged. A new $(t\text{-}1)$x$(t\text{-}1)$ matrix of similarities is then computed, with the similarities involving multi-element groups being defined as the highest similarity for any pair of elements, one from each group (hence the name 'nearest neighbour'). Groups are successively linked up in this way until a final single group of all t elements is obtained. Once a group has been defined, the elements in that group will never be separated. New groups may only be defined by merging previously existing groups. This property, which makes the method a **hierarchical** method in the terminology of Cluster Analysis, enables the results to be graphically represented by a **dendrogram** or **branching tree**.

Examples of the application of this well-known technique are given in the next two Subsections.

IV.5.4 An example of multiple comparisons

In order to illustrate the above approaches to multiple comparisons of PCAs, we will once again turn to Kendall's soil data set, given in Table II.1.

Ten transformations of this data set will be considered, alongside the original data. They correspond to ten different ways of making the original data dimensionless, in the spirit of Correlation Matrix PCA (which is one of the transformations considered).

The eleven data matrices whose PCAs are to be compared can all be written as $\mathbf{XU}_i$, $(i = 1, ..., 11)$, where $\mathbf{X}$ is the original (column-centered) 20x4 data matrix and $\{\mathbf{U}_i\}_{i=1}^{11}$ are 4x4 diagonal matrices. Representing the i-th column of $\mathbf{X}$ by $\mathbf{x}_i$ and the ℓ_k-norms in $\mathbb{R}^n$ by $\|\cdot\|_k$, the diagonal elements of the 11 transformation matrices are given by:

$\mathbf{U}_1$	--	1	(Covariance Matrix PCA; $\mathbf{U}_1 = \mathbf{I}_4$)
$\mathbf{U}_2$	--	$1/\|\mathbf{x}_i\|_1$	
$\mathbf{U}_3$	--	$1/\|\mathbf{x}_i\|_2$	(Correlation Matrix PCA, a scaling factor aside)
$\mathbf{U}_4$	--	$1/\|\mathbf{x}_i\|_3$	
$\mathbf{U}_5$	--	$1/\|\mathbf{x}_i\|_4$	
$\mathbf{U}_6$	--	$1/\|\mathbf{x}_i\|_5$	
$\mathbf{U}_7$	--	$1/\|\mathbf{x}_i\|_{10}$	
$\mathbf{U}_8$	--	$1/\|\mathbf{x}_i\|_{20}$	

$\mathbf{U}_9$ -- $1/\|\mathbf{x}_i\|_\infty$
$\mathbf{U}_{10}$ -- $1/R(\mathbf{x}_i)$ (where $\mathrm{R}(\mathbf{x}_i)$ is the range of the i-th column of $\mathbf{X}$)
$\mathbf{U}_{11}$ -- $1/M(\mathbf{x}_i)$ (where $\mathrm{M}(\mathbf{x}_i)$ is the median of the absolute values of the elements in column i of $\mathbf{X}$)

Matrices $\mathbf{U}_2$ through $\mathbf{U}_9$ provide a sequence of transformations which are directly related since the re-scaling of each variable is always obtained by dividing it by one of its ℓ_k-norms. Matrix $\mathbf{U}_{11}$ is related to matrix $\mathbf{U}_2$, but in a different way. The denominators of the latter's diagonal elements are – a scaling factor aside – the *mean* absolute deviation from the mean, whereas those of $\mathbf{U}_{11}$ are the *median* absolute deviation from the mean. Matrix $\mathbf{U}_{10}$ corresponds to a suggestion by Gower (1966) [19, (pg.327)].

It should be noted that the ℓ_k-norm of a variable centered about its mean, for k even, is its k-th central moment (apart from a common scaling factor $\frac{1}{\sqrt[k]{n}}$). But this is not the case when k is odd, since the norms always involve the *absolute* deviations from the mean. The possibility of using the odd-order moments, rather than the norms, for k odd, was considered. But such transformations were found to be highly prone to 'unduly' highlighting one or a few of the variables with more symmetric frequency distributions. Their near-zero values would also be very sensitive to minor changes in the data. For these reasons, transformations involving odd-order moments have not been used.

The PCs of the 11 matrices $\{\mathbf{XU}_i\}_{i=1}^{11}$ will be compared using the 11x11 matrix of similarity indices in $\mathbb{R}^n$, that is, the matrix of cosines of the angles between each pair of matrices $\{\mathbf{XU}_i'\mathbf{U}_i\mathbf{X}'\}_{i=1}^{11}$. The comparisons involving only $\mathbb{R}^p$ or both spaces simultaneously will not be presented here, since they provide results that are not substantially different.

The matrix $\mathbf{\Phi}$ of similarity indices in $\mathbb{R}^n$ is given in Table IV.1. A cursory glance reveals that matrices $\mathbf{XU}_2$ through $\mathbf{XU}_{10}$ will have, overall, fairly similar PCs. On the other hand, $\mathbf{XU}_1$ and $\mathbf{XU}_{11}$ stand apart, both from the main group and from each other. The Covariance and Correlation Matrix PCs have an unimpressive similarity index (0.72).

$\mathbf{U}_j$	1	2	3	4	5	6	7	8	9	10	11
1	1.00										
2	0.69	1.00									
3	0.72	0.99	1.00								
4	0.75	0.98	1.00	1.00							
5	0.76	0.97	0.99	1.00	1.00						
6	0.77	0.96	0.98	1.00	1.00	1.00					
7	0.79	0.93	0.96	0.98	0.99	1.00	1.00				
8	0.79	0.92	0.95	0.98	0.99	0.99	1.00	1.00			
9	0.78	0.92	0.95	0.97	0.98	0.99	1.00	1.00	1.00		
10	0.75	0.95	0.98	0.99	1.00	1.00	1.00	0.99	0.99	1.00	
11	0.42	0.92	0.87	0.83	0.79	0.77	0.73	0.72	0.71	0.78	1.00

Table IV.1: Matrix of similarity indices in $\mathbb{R}^n$ for 11 transformations of Kendall's soil data

It is not surprising that the matrices $\mathbf{U}_2$ through $\mathbf{U}_9$ should perform in a fairly similar manner. We would expect variation in the values of each variable's ℓ_k-norms, for different values of k , to be fairly similar in nature. The values for the similarity indices involving $\mathbf{U}_{10}$ and $\mathbf{U}_{11}$ are less predictable and probably more data-dependent.

The first four eigenvalues, relative eigenvalues and cumulative relative eigenvalues of the matrix $\mathbf{\Phi}$ are given in Table IV.2. It is immediately apparent that a 2-D graphical representation of the 11-point scatter, as described in Subsection IV.5.2, will be appropriate. This 2-D plot is given in Figure IV.1.

Eigenvalue	Relative Eigenvalue	Cumulative relative eigenvalue
10.0783	0.9162	0.9162
0.6614	0.0601	0.9763
0.2330	0.0212	0.9975
0.0273	0.0025	1.0000

Table IV.2: First four (cumulative (relative (eigenvalues))) for the matrix of similarity indices in $\mathbb{R}^n$

All 11 matrices $\{\mathbf{XU}_i\}_{i=1}^{11}$ are reasonably well represented in Figure IV.1, since they are all on, or close to, the unit circumference. The angles between the vectors from the origin which they define will therefore give us an indication of the relative similarity of their PCs.

The overall picture which emerges from Figure IV.1 agrees with the preliminary inspection of matrix $\mathbf{\Phi}$: the matrices $\mathbf{XU}_2$ to $\mathbf{XU}_{10}$ form a group of matrices with relatively similar PCs. Both $\mathbf{XU}_1$ and $\mathbf{XU}_{11}$ stand on their own, being more similar to the large group of transformations than to each other. But a closer look at the large cluster of points also reveals a curious pattern: the matrices $\mathbf{XU}_2$ to $\mathbf{XU}_9$, which correspond to the use of the ℓ_k-norms for increasing values of k , are positioned

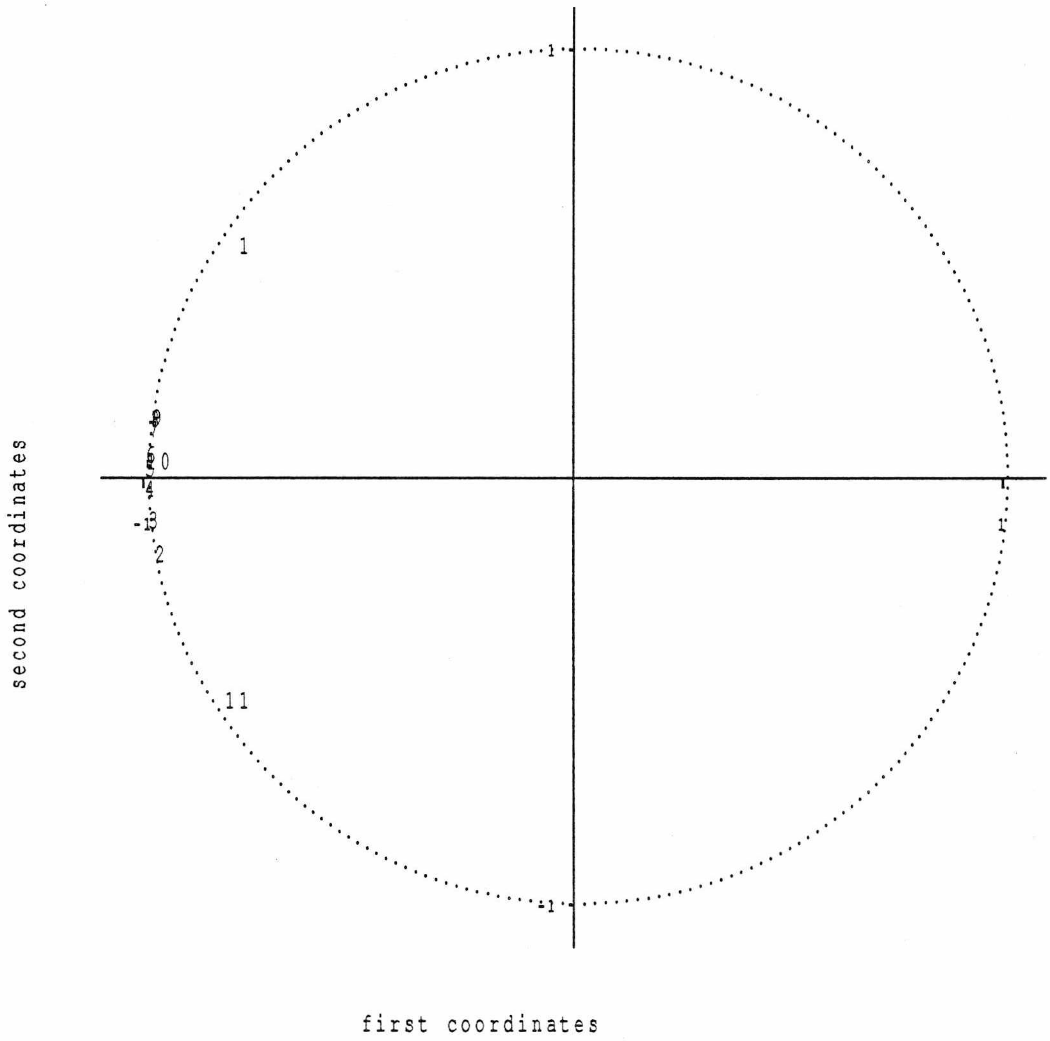

Figure IV.1: Principal Coordinates plot for the similarity indices in $\mathbb{R}^n$ of Kendall's soil data

in a relatively orderly fashion. If we follow the unit circumference in a clockwise direction, we begin with the point corresponding to the use of the ℓ_1-norm and then encounter those corresponding to the other ℓ_k-norms, for growing values of k . That the point associated with $\mathbf{XU}_{10}$ should almost coincide with the points for $\mathbf{XU}_5$ and $\mathbf{XU}_6$ appears to be more of a data-specific coincidence than anything else.

The single-link dendrogram which results from matrix $\mathbf{\Phi}$ of the similarity indices in $\mathbb{R}^n$ is given in Figure IV.2.

```
**** Nearest neighbour cluster analysis ****

**** Dendrogram ****
** Levels   100.0  90.0  80.0  70.0

            1  ..............
            8  ..            )
            9  ..)           )
            7  ..)           )
           10  ..)           )
            6  ..)           )
            5  ..)           )
            4  ..)           )
            3  ..)           )
            2  ..)..         )
           11  .....)........)..............
```

Figure IV.2: Single-link dendrogram for the matrix of similarities in $\mathbb{R}^n$ of Kendall's soil data and its transformations

The information conveyed by the dendrogram in Figure IV.2 is very much in line with the information from Figure IV.1, although the distinction between transformation 11 and the large cluster is not as clear-cut as we might have expected. This results from the fact that one element in the cluster (the point for transformation

2) is fairly close to the point for transformation 11. Their similarity index is 0.92 and the nearest-neighbour clustering method picks up this high value. Incidentally, this 'spurious' merging of two 'distinct' clusters is known as *chaining* in the terminology of Cluster Analysis. It is discussed by Chatfield and Collins (1980) [8, (pg.228)] who point it out as the main reason why alternatives to the single-link hierarchical clustering method were devised.

But the 2-D graph in Figure IV.1 is clearly superior to the dendrogram, both in highlighting that the transformations $\mathbf{U}_2$ to $\mathbf{U}_{10}$ *do* form a cluster (despite the fact that $\mathbf{U}_2$ is as similar to $\mathbf{U}_{11}$ in its effects as it is to $\mathbf{U}_8$ and to $\mathbf{U}_9$) and in picking up the orderly positioning of the ℓ_k-norm transformations within this cluster. This agrees with the preference expressed by Chatfield and Collins [8, (pg.229)] for the use of visual clustering methods.

All things considered, most of the ten alternative re-scalings which produce dimensionless data will give us reasonably similar PCs for this data set, with globally similar relative importances attached to each PC. The odd-man out in the group of transformations considered is $\mathbf{U}_{11}$, which uses the median absolute deviation from the mean as the standardizing quantity for each variable. It is the only one of the transformations considered which does not make use of the extremal values of each variable. This factor appears to be setting it apart from the others.

IV.5.5 A second example of multiple comparisons

A second example of multiple comparisons will focus on the data set given in Table III.2, where four meteorological variables are measured in the Lisbon Meteorological Station, for each of the 31 days of January 1967. It will be recalled that the variables in the data set were: maximum temperature (in °C); minimum temperature (in °C); amount of sunshine (as a proportion of total daylight hours); and rainfall (in mm.).

This time, fifteen transformations will be considered. The first 11 are the same as the ones used in the example of the previous Subsection. The last four are transformations which are specific to the nature of this data set. They are:

$\mathbf{U}_{12}$ – converts variable 3 to a percentage, rather than a proportion; in other words, it is a diagonal matrix of ones except for the third diagonal element which is 100. We expect this to significantly increase the role of this variable in the PCA, since the length of the corresponding vector in $\mathbb{R}^n$ will also increase 100-fold.

$\mathbf{U}_{13}$ – converts the original variables from metric units to SI units. Specifically, the temperatures in variables 1 and 2 are multiplied by 9/5 (the additive constant is unnecessary due to column-centering) and rainfall is divided by 25.4 to give values in inches. The proportion of daylight hours with sunshine is left unchanged. Variable 4 should see its role in the PCA greatly reduced.

$\mathbf{U}_{14}$ – combines the effects of the two previous transformation matrices.

$\mathbf{U}_{15}$ – replaces the minimum temperature with a new variable: the temperature range during the day. This is the only non-diagonal transformation matrix used in these two examples. Its transpose is given by:

$$\mathbf{U}'_{15} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

As with the previous example, only the comparison involving one kind of similarity index will be presented. The index chosen for comparison is the *global* similarity index.

The matrix $\mathbf{\Phi}$ of correlations between the 15 transformed data matrices $\{\mathbf{XU'}_i\}_{i=1}^{15}$ is given in Table IV.3. As in the previous example (and despite the different similarity index which is being considered) the transformations $\mathbf{U}_3$ to $\mathbf{U}_{10}$ seem to form a relatively homogeneous group, with $\mathbf{U}_2$ still reasonably similar. The four new transformations, however, produce some very low values of the global similarity index, as low as 0.08 (between $\mathbf{XU}_{14}$ and $\mathbf{XU}'_{15}$).

The global similarity index for the Covariance and Correlation Matrix PCA is somewhat modest (0.79), especially considering that for the four-variable data set given by $\mathbf{X}$ there is a lower bound of $\frac{1}{\sqrt{p}} = 0.50$ for this particular comparison (see Subsection IV.4.3). Even more troubling is the decidedly poor value (0.51) of the global similarity index for the (Covariance Matrix) PCAs of the data in metric units

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.00														
2	0.91	1.00													
3	0.79	0.97	1.00												
4	0.70	0.93	0.99	1.00											
5	0.65	0.90	0.98	1.00	1.00										
6	0.62	0.89	0.97	0.99	1.00	1.00									
7	0.58	0.86	0.95	0.99	1.00	1.00	1.00								
8	0.57	0.85	0.95	0.99	0.99	1.00	1.00	1.00							
9	0.56	0.85	0.95	0.98	0.99	1.00	1.00	1.00	1.00						
10	0.66	0.91	0.98	0.99	0.99	0.99	0.99	0.99	0.99	1.00					
11	0.94	0.99	0.94	0.87	0.84	0.82	0.79	0.78	0.78	0.86	1.00				
12	0.29	0.60	0.68	0.71	0.73	0.74	0.76	0.77	0.77	0.78	0.53	1.00			
13	0.51	0.66	0.75	0.78	0.78	0.77	0.76	0.76	0.76	0.71	0.62	0.17	1.00		
14	0.15	0.52	0.64	0.71	0.74	0.76	0.79	0.79	0.79	0.78	0.43	0.97	0.26	1.00	
15	0.87	0.74	0.58	0.46	0.41	0.38	0.34	0.32	0.31	0.45	0.81	0.25	0.18	0.08	1.00

Table IV.3: Matrix of global similarity indices for 15 transformations of the January 1967 Lisbon Meteorological data

($\mathbf{XU}_1$) and in SI units ($\mathbf{XU}_{13}$). This is a powerful example of the scale-dependence problem of PCA discussed in Chapter I. The significance of the example is further highlighted by the very high (0.97) value for the index involving the pair $\mathbf{XU}_{12}$,$\mathbf{XU}_{14}$. This is a comparison similar to the above in that variables 1,2 and 4 are in metric units in $\mathbf{XU}_{12}$ and in SI units in $\mathbf{XU}_{14}$. The difference lies in the fact that now variable 3 is expressed as a percentage rather than a proportion. The fact that variable 3 is now 100 times larger — and therefore playing a much larger role in the PCAs of $\mathbf{XU}_{12}$ and $\mathbf{XU}_{14}$ — has made the previously significant effects of converting from one measurement system to another almost irrelevant. The dominant role of the now much larger variable 3, in both cases, is confirmed by the very low values of s_g when comparing data sets that are identical, but for this particular change. In fact, $s_g(\mathbf{XU}_1, \mathbf{XU}_{12}) = 0.29$ and $s_g(\mathbf{XU}_{13}, \mathbf{XU}_{14}) = 0.26$.

The first five eigenvalues, relative eigenvalues and cumulative relative eigenvalues for the above matrix are given in Table IV.4.

Eigenvalues	Relative eigenvalues	Cumulative relative eigenvalues
11.8867	0.7924	0.7924
1.9840	0.1323	0.9247
1.0488	0.0699	0.9946
0.0736	0.0049	0.9995
0.0070	0.0005	1.0000

Table IV.4: First five (cumulative (relative (eigenvalues))) of the matrix of global similarity indices for the 1967 meteorological data

It will be noted that the first relative eigenvalue is substantially lower than its counterpart in the first example, which is not surprising considering the more diverse nature of the transformation matrices used in this example. But by the time the third dimension is taken into account, this gap has been almost completely bridged. For most purposes, a 2-D graphical representation should be adequate, although any conclusions to be drawn from it should be cross-checked.

Figure IV.3 gives the 'best' 2-D representation of the 15 data matrices. Most matrices appear to be well represented in this graph. The clear exception is matrix $\mathbf{XU}_{13}$, which represented the conversion of the data from the metric to SI units. The transformed data matrices $\mathbf{XU}_{12}$, $\mathbf{XU}_{14}$ and $\mathbf{XU}'_{15}$ are also less well represented than most, but not to the extent of $\mathbf{XU}_{13}$.

The overall pattern which emerges is — insofar as the comparison is possible — similar to that of Figure IV.1. The ℓ_k-norm transformations appear in a similarly orderly fashion, 'moving' around the unit circumference in a counter-clockwise direction

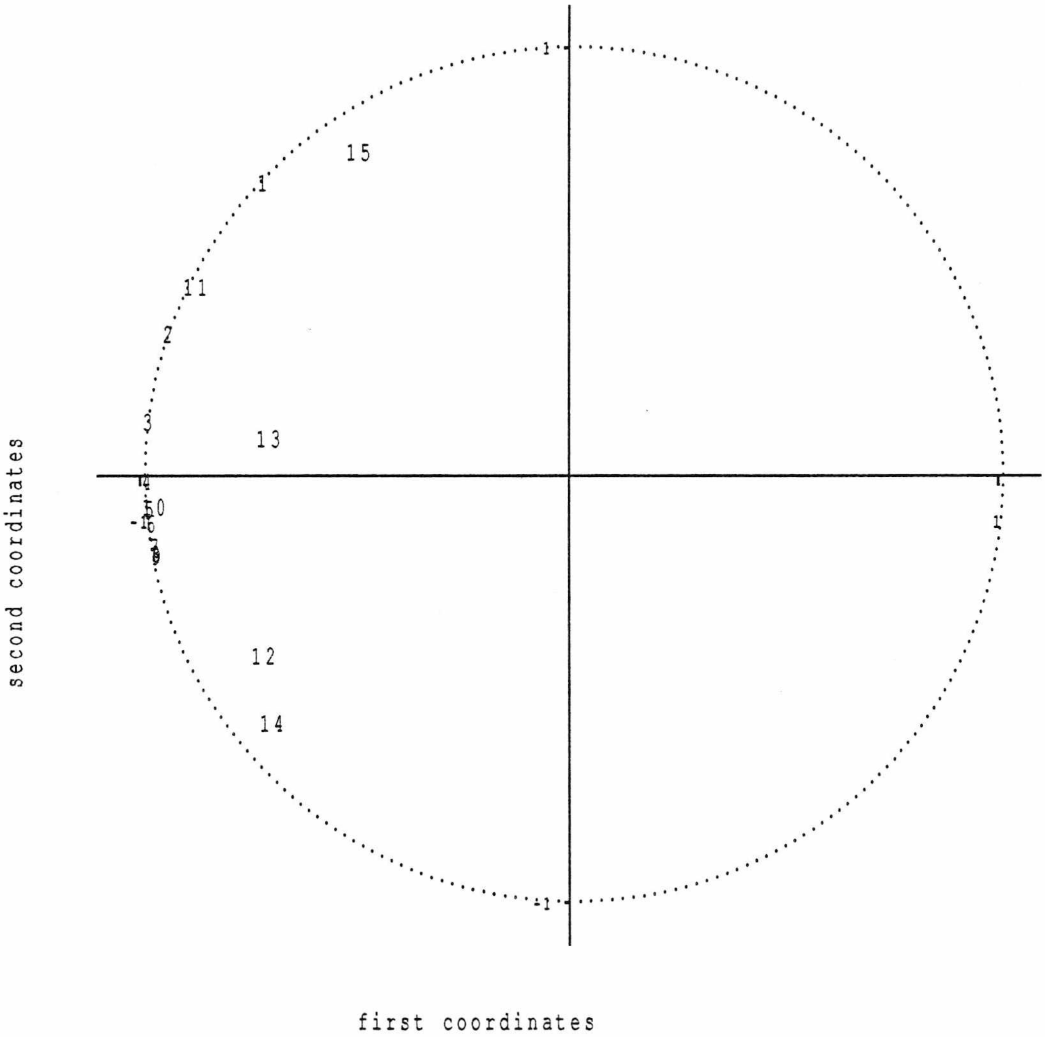

Figure IV.3: First (horizontal) and second (vertical) Principal Coordinates of the Meteorological data set

(which is purely arbitrary due to the sign-switching ambiguity of eigenvectors) as the value of k increases. However, this motion is now away from the point representing the original data matrix, rather than towards it, as in Figure IV.1. In other words, the PCA in this group most similar to the original Covariance Matrix PCA of $\mathbf{X}$ is the one involving the division by the ℓ_1-norm (and not by the ℓ_∞-norm as in the previous example).

A look at the 2-D graph which results from considering the first and *third* principal coordinates of each point is also instructive. This graph is given in Figure IV.4.

The main advantage of this graph is the faithful representation of $\mathbf{XU}_{13}$. For example, the very low values of the similarity indices between $\mathbf{XU}_{13}$, on the one hand, and $\mathbf{XU}_{12}, \mathbf{XU}_{14}$ and $\mathbf{XU}'_{15}$, on the other, are now apparent in a way which they were not in the previous graph, given the less faithful representation of these points. In particular, the fact that $\mathbf{XU}_{13}$ is on the opposite side of the origin, along the third axis, from $\mathbf{XU}_{12}, \mathbf{XU}_{14}$ and $\mathbf{XU}'_{15}$, made their relative positions in Figure IV.3 somewhat misleading.

All 15 points can be seen to lie approximately on the surface of the 3-D unit hypersphere in the subspace where their first three principal coordinates are given. This is a reflection of the high value (0.99) of the cumulative relative eigenvalue for these three dimensions. A mental merging of Figures IV.3 and IV.4 will therefore provide an almost exact representation of the 'full' 15-point scatter in $\mathbb{R}^{15}$.

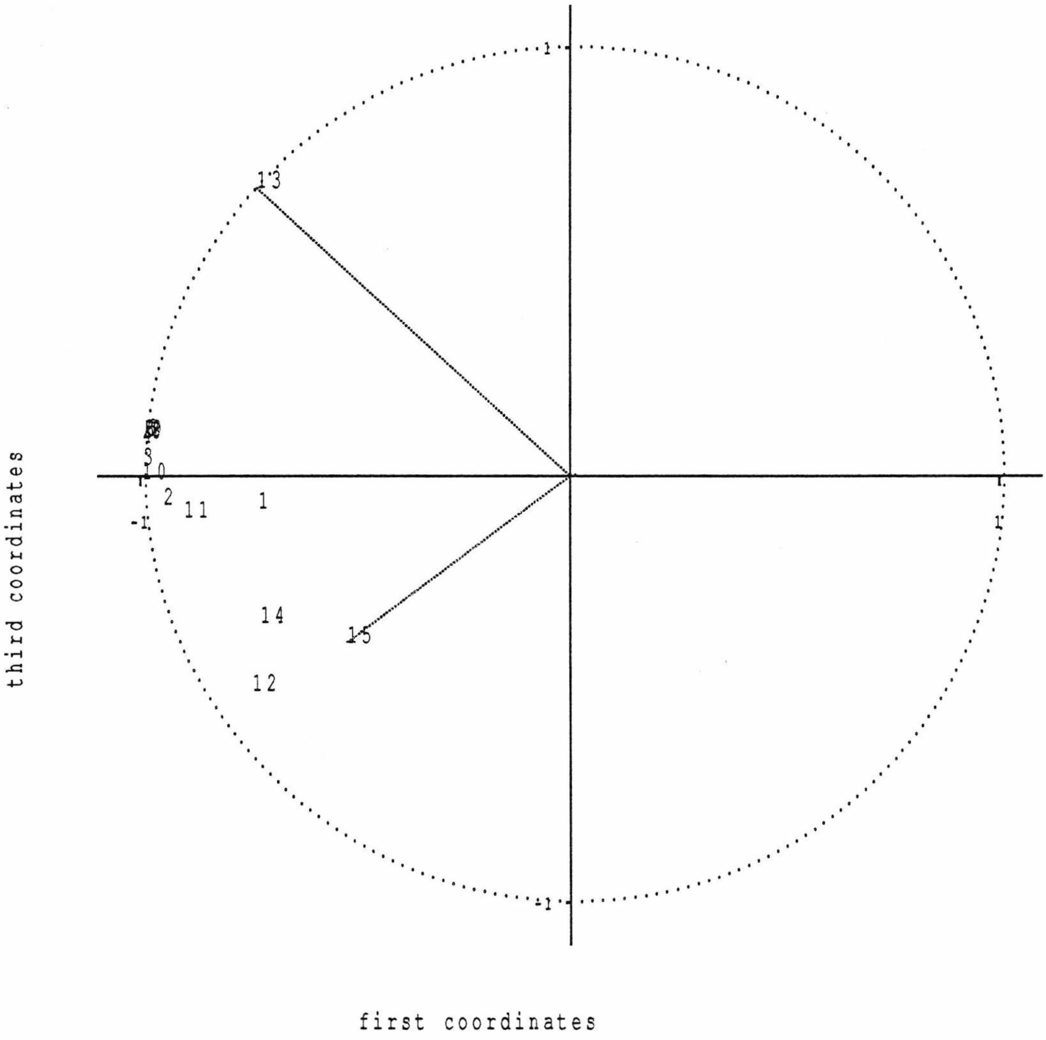

Figure IV.4: First (horizontal) and third (vertical) Principal Coordinates for the global similarities of the Meteorological data set

The dendrogram for the nearest-neighbour clustering of the 15 matrices which is produced by their matrix $\mathbf{\Phi}$ of global similarity indices is given in Figure IV.5.

```
**** Nearest neighbour cluster analysis ****

**** Dendrogram ****
** Levels   100.0  90.0  80.0  70.0

            1  .....
           11  ..   )
            2  ..)  )
            3  ..)  )
            4  ..)  )
            5  ..)  )
            6  ..)  )
            7  ..)  )
            8  ..)  )
            9  ..)  )
           10  ..)..)..
           15  ........).....
           14  ..           )
           12  ..)...........)
           13  ..............)..............
```

Figure IV.5: Single-link dendrogram for the global similarity indices of the Meteorological data

The information provided by the dendrogram is, once again, similar to, though poorer than, that provided by the Principal Coordinate Analysis. The initial clusterings ($\mathbf{XU}_2$ to $\mathbf{XU}_{11}$ and $\mathbf{XU}_{12}, \mathbf{XU}_{14}$) as well as the way these clusters incorporate the remaining points and merge are as expected. However, subtler traits such as the ordering of the ℓ_k-norm transformations are not picked up in the dendrogram.

All things considered, the multiple comparisons have thrown new light on the relative behaviour of the PCAs of each transformed data matrix.

Chapter V

Projections

The linear transformations of the data considered in the previous Chapter were assumed to be invertible. But in many applications, non-invertible linear transformations are called for. Prominent among these is the class of projections onto some subspace of either $\mathbb{R}^p$ and/or $\mathbb{R}^n$. In this Chapter we shall focus on the use of projection matrices to transform the data prior to a PCA.

V.1 Projection matrices and PCA

Appendix A introduces the concepts of projection (orthogonal or non-orthogonal, in the standard inner product) and projection matrix (idempotent matrix). Proposition A.19 lists several results concerning projection matrices and their decompositions. In Section IV.1, orthogonal projections with alternative inner products were discussed and it was shown that they could always be viewed as non-orthogonal projections in the standard inner product.

We now turn our attention to the implications of projecting the rows and/or columns of a data matrix, for Principal Component Analysis.

V.1.1 Takane & Shibayama's framework

Takane and Shibayama (1991) [64] suggest a general framework in which to tackle this issue.

They first consider two orthogonal projectors $\mathbf{P}_n$ and $\mathbf{P}_p$ onto some subspaces $\mathcal{A}_n$, $\mathcal{A}_p$ of, respectively, $\mathbb{R}^n$ and $\mathbb{R}^p$. Proposition A.19 then guarantees that $\mathbf{I}_n - \mathbf{P}_n$ and $\mathbf{I}_p - \mathbf{P}_p$ are the orthogonal projectors onto the orthogonal complements of these subspaces, $\mathcal{A}_n^\perp$ and $\mathcal{A}_p^\perp$. Any $n \times p$ data matrix $\mathbf{Y}$ can then be decomposed as:

$$\mathbf{Y} = \mathbf{P}_n\mathbf{Y}\mathbf{P}_p + (\mathbf{I}_n - \mathbf{P}_n)\mathbf{Y}\mathbf{P}_p + \mathbf{P}_n\mathbf{Y}(\mathbf{I}_p - \mathbf{P}_p) + (\mathbf{I}_n - \mathbf{P}_n)\mathbf{Y}(\mathbf{I}_p - \mathbf{P}_p) \qquad (\text{V.1.1})$$

Equation (V.1.1) is what Takane and Shibayama call the **External Analysis** of data matrix $\mathbf{Y}$. It amounts to a decomposition of $\mathbf{Y}$ into four terms: the first with columns in $\mathcal{A}_n$ and rows in $\mathcal{A}_p$; the second with columns in $\mathcal{A}_n^\perp$ and rows in $\mathcal{A}_p$; the third with columns in $\mathcal{A}_n$ and rows in $\mathcal{A}_p^\perp$; the last with columns in $\mathcal{A}_n^\perp$ and rows in $\mathcal{A}_p^\perp$. The underlying assumption is that the choice of projectors $\mathbf{P}_n$ and $\mathbf{P}_p$ is based on problem-specific considerations which then justify PCAs of some (or all) terms in (V.1.1) *individually* or of sums of some subsets of those terms. This second stage of analysis of $\mathbf{Y}$ is termed the **Internal Analysis** by Takane and Shibayama.

The above decomposition is justified by the authors on the basis of incorporating *linear constraints*, or as the authors say, *external information* which provides some grounds for assuming that $\mathbf{Y}$ has additive components in the subspaces of $\mathbb{R}^n$ and $\mathbb{R}^p$ onto which $\mathbf{P}_n$ and $\mathbf{P}_p$ project. Their discussion is framed in an inferential context, with the terms in equation (V.1.1) providing estimates for a model for the data matrix $\mathbf{Y}$. Several examples of applications of the above decomposition which can be found in the literature, are also given. In addition, there is a reference to the obvious generalization of (V.1.1) which results from further decompositions of $\mathbf{P}_n$ and $\mathbf{P}_p$. In fact, for any *partitioning of the identities* $\mathbf{I}_m = \sum_{i=1}^{k} \mathbf{P}_m^{[i]}$, $(m = n, p)$, where $\left\{P_m^{[i]}\right\}_{i=1}^{k}$ are orthogonal projection matrices in $\mathbb{R}^m$, we can write:

$$\mathbf{Y} = \sum_{i=1}^{k} \sum_{j=1}^{l} \mathbf{P}_n{}^{[i]} \mathbf{Y} \mathbf{P}_p{}^{[j]} \tag{V.1.2}$$

Takane and Shibayama do not appear to be overly concerned in their paper with the column-centering requirement of PCA. They essentially equate the PCA of a matrix to its SVD ([64, (pg.101)]). Actually, column-centering can be incorporated into the External Analysis equation itself, as can the other variants of PCA discussed in Subsection I.4.2. If we take $\mathbf{P}_n = \mathbf{P}_{\mathbf{1}_n} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n{}'$ and $\mathbf{P}_p = \mathbf{P}_{\mathbf{1}_p} = \mathbf{I}_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p{}'$ in (V.1.1), with $\mathbf{Y} = \frac{1}{\sqrt{n}}\mathbf{Z}$ for a generic $n \times p$ data matrix $\mathbf{Z}$, we have:

- **a (column-centered) PCA of Z** is an SVD of the sum of the first and third terms of the External Analysis equation, *i.e.*, an SVD of $\mathbf{P}_n\mathbf{Y}$.

- **a row-centered** (but not column-centered) **PCA of Z** is an SVD of the sum of the first and second terms in (V.1.1), *i.e.*, an SVD of $\mathbf{YP}_p$.
- **a doubly-centered PCA of Z** is an SVD of the first term of (V.1.1).
- **a non-centered PCA of Z** is an SVD of the sum of all four terms in (V.1.1).

Alternatively, we may just assume that any term or sum of terms in equation (V.1.1) which is to be subjected to PCA will be column-centered during the Internal Analysis stage. This will be equivalent to assuming that the data matrix $\mathbf{Y}$ is column-centered prior to the External Analysis decomposition whenever $\mathbf{P}_{\mathbf{1}_n}$ commutes with any $n \times n$ projection matrices used in the terms which are to be analyzed. In such cases, the following theorem tells us that these two equivalent approaches will also correspond to using a single $n \times n$ projection matrix. The Theorem is freely reproduced from Pease (1965) [49, (pg.264)]. The symbol $\mathcal{N}$ stands for the nullspace of a linear transformation (see Appendix A).

Theorem V.1 (Pease) *The product of two projection matrices* $\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2$ *is a projection matrix if* $\mathbf{Q}_1\mathbf{Q}_2 = \mathbf{Q}_2\mathbf{Q}_1$, *in which case* $\mathbf{Q}$ *projects vectors onto* $\Re(\mathbf{Q}) = \Re(\mathbf{Q}_1) \cap \Re(\mathbf{Q}_2)$ *along the subspace spanned by the vectors in* $\mathcal{N}(\mathbf{Q}_1) \cup \mathcal{N}(\mathbf{Q}_2)$. *In particular, if* $\mathbf{Q}_1$,$\mathbf{Q}_2$ *are matrices of* orthogonal *projections, then so is* $\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2$.

The last conclusion in the Theorem results from the easily checked fact that if $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are symmetric and commute, then $\mathbf{Q}$ is also symmetric.

It should be pointed out that the decomposition (V.1.1) corresponds to a repeated application of the Orthogonal Projection Theorem (Proposition A.16) in the matrix space $\mathbb{M}_{n\times p}$. This result is formally presented below, the last point of which is also given by Takane and Shibayama.

Theorem V.2 *Let* $\mathbf{Y} \in \mathbb{M}_{\mathrm{nxp}}$ *be a data matrix. Let* $\mathcal{A}_n(\mathcal{A}_p)$ *be a subspace of* $\mathbb{R}^n(\mathbb{R}^p)$. *Let* $\mathbf{P}_n(\mathbf{P}_p)$ *be the orthogonal projector onto* $\mathcal{A}_n(\mathcal{A}_p)$. *Let* $\mathbb{N}_{\mathrm{nxp}}(\mathbb{P}_{\mathrm{nxp}})$ *be the set of* $\mathrm{n{\times}p}$ *matrices whose* p *columns (*n *rows) are elements of* $\mathcal{A}_n(\mathcal{A}_p)$. *Let* $\mathbb{N}_{\mathrm{nxp}}{}^{\perp}(\mathbb{P}_{\mathrm{nxp}}{}^{\perp})$ *be the set of* $\mathrm{n{\times}p}$ *matrices whose* p *columns (*n *rows) are elements of* $\mathcal{A}_n{}^{\perp}(\mathcal{A}_p{}^{\perp})$, *the orthogonal complement of* $\mathcal{A}_n(\mathcal{A}_p)$. *Then:*

i). $\mathbb{N}_{\mathrm{nxp}}$ *and* $\mathbb{P}_{\mathrm{nxp}}$ *are subspaces of* $\mathbb{M}_{\mathrm{nxp}}$. $\mathbb{N}_{\mathrm{nxp}}{}^{\perp}$ *and* $\mathbb{P}_{\mathrm{nxp}}{}^{\perp}$ *are their orthogonal complements in* $\mathbb{M}_{\mathrm{nxp}}$.

ii). $\mathbb{M}_{\mathrm{nxp}} = (\mathbb{N}_{\mathrm{nxp}} \cap \mathbb{P}_{\mathrm{nxp}}) \oplus (\mathbb{N}_{\mathrm{nxp}}{}^{\perp} \cap \mathbb{P}_{\mathrm{nxp}}) \oplus (\mathbb{N}_{\mathrm{nxp}} \cap \mathbb{P}_{\mathrm{nxp}}{}^{\perp}) \oplus (\mathbb{N}_{\mathrm{nxp}}{}^{\perp} \cap \mathbb{P}_{\mathrm{nxp}}{}^{\perp})$. *The unique decomposition of* $\mathbf{Y} \in \mathbb{M}_{\mathrm{nxp}}$ *determined by this direct sum is given by equation (V.1.1).*

iii). Every pair of direct summands in ii) are mutually orthogonal spaces.

iv). $||\mathbf{Y}||^2 = ||\mathbf{P}_n\mathbf{Y}\mathbf{P}_p||^2 + ||(\mathbf{I}_n - \mathbf{P}_n)\mathbf{Y}\mathbf{P}_p||^2 + ||\mathbf{P}_n\mathbf{Y}(\mathbf{I}_p - \mathbf{P}_p)||^2 + ||(\mathbf{I}_n - \mathbf{P}_n)\mathbf{Y}(\mathbf{I}_p - \mathbf{P}_p)||^2$

Proof.

i). The sets $\mathbb{N}_{n\times p}$ and $\mathbb{P}_{n\times p}$ are subspaces of $\mathbb{M}_{n\times p}$ if they include all linear combinations of their elements. But that results directly from the definition of those sets, the definitions of matrix and scalar multiplication and the fact that $\mathcal{A}_n$ and $\mathcal{A}_p$ are linear spaces. The fact that any matrix in $\mathbb{N}_{n\times p}^{\perp}$ is orthogonal to any matrix in $\mathbb{N}_{n\times p}$ is also easily verified, since if $\mathbf{A} \in \mathbb{N}_{n\times p}$ and $\mathbf{B} \in \mathbb{N}_{n\times p}^{\perp}$, we have: $<\mathbf{A},\mathbf{B}>= \mathrm{tr}(\mathbf{A}'\mathbf{B}) = \sum_{i=1}^{p} \mathbf{a}_i'\mathbf{b}_i$ where $\mathbf{a}_i$,$\mathbf{b}_i$ are the i-th columns of $\mathbf{A}$ and $\mathbf{B}$, respectively. But $\mathbf{a}_i \in \mathcal{A}_n$ and $\mathbf{b}_i \in \mathcal{A}_n^{\perp}$, $\forall i = 1, ..., p$. Hence, $<\mathbf{A},\mathbf{B}>= 0$. Now suppose $\mathbf{B} \in \mathbb{M}_{n\times p}$ is some matrix such that $<\mathbf{A},\mathbf{B}>= 0$, $\forall \mathbf{A} \in \mathbb{N}_{n\times p}$. We wish to show that $\mathbf{B} \in \mathbb{N}_{n\times p}^{\perp}$, *i.e.*, that the columns of $\mathbf{B}$ are all in $\mathcal{A}_n^{\perp}$. Suppose this were not the case, that is, suppose that at least one column of $\mathbf{B}$ had a non-zero component in $\mathcal{A}_n$. We could then define the matrix $\mathbf{C} \in \mathbb{N}_{n\times p}$ to be the matrix whose j-th column is given by $\mathbf{P}_n\mathbf{b}_j$. This matrix would not be orthogonal to $\mathbf{B}$ since we would have:

$$<\mathbf{C},\mathbf{B}>= \sum_{j=1}^{p}(\mathbf{P}_n\mathbf{b}_j)'\left[\mathbf{P}_n\mathbf{b}_j + (\mathbf{I}_n - \mathbf{P}_n)\mathbf{b}_j\right] = \sum_{j=1}^{p}||\mathbf{P}_n\mathbf{b}_j||^2 > 0$$

because $\mathbf{P}_n\mathbf{b}_j \neq 0$ for at least one $\mathbf{b}_j$. Repeating the argument for matrices in $\mathbb{P}_{n\times p}$ produces the desired result: $\mathbb{N}_{n\times p}^{\perp}$ and $\mathbb{P}_{n\times p}^{\perp}$ are the orthogonal complements of $\mathbb{N}_{n\times p}$ and $\mathbb{P}_{n\times p}$, respectively. Both orthogonal complements are naturally also subspaces of $\mathbb{M}_{n\times p}$(Proposition A.8).

ii). Let $\mathbf{Y}$ be any matrix in $\mathbb{M}_{n\times p}$. We know (Proposition A.14) that we can uniquely decompose $\mathbf{Y}$ into the sum of a term in $\mathbb{N}_{n\times p}$ and a term in its orthogonal complement $\mathbb{N}_{n\times p}^{\perp}$. But any matrix in either of these spaces can also be uniquely written as the sum of a matrix in $\mathbb{P}_{n\times p}$ and a matrix in $\mathbb{P}_{n\times p}^{\perp}$.

Hence, $\mathbf{Y}$ can be uniquely decomposed into a sum of four terms in the spaces $\mathbb{N}_{n \times p} \cap \mathbb{P}_{n \times p}$, $\mathbb{N}_{n \times p} \cap \mathbb{P}_{n \times p}^{\perp}$, $\mathbb{N}_{n \times p}^{\perp} \cap \mathbb{P}_{n \times p}$ and $\mathbb{N}_{n \times p}^{\perp} \cap \mathbb{P}_{n \times p}^{\perp}$. The nature of the spaces $\mathbb{N}_{n \times p}$ (matrices of $\mathbb{M}_{n \times p}$ with columns in $\mathcal{A}_n$) implies that the unique decomposition of $\mathbf{Y}$ into a term in $\mathbb{N}_{n \times p}$ and a term in $\mathbb{N}_{n \times p}^{\perp}$ must be given by $\mathbf{Y} = \mathbf{P}_n\mathbf{Y} + (\mathbf{I}_n - \mathbf{P}_n)\mathbf{Y}$. Likewise, the unique decomposition of any matrix $\mathbf{Z} \in \mathbb{M}_{n \times p}$ into a term in $\mathbb{P}_{n \times p}$ (rows in $\mathcal{A}_p$) and a term in $\mathbb{P}_{n \times p}^{\perp}$ (rows in $\mathcal{A}_p^{\perp}$) must be given by $\mathbf{Z} = \mathbf{Z}\mathbf{P}_p + \mathbf{Z}(\mathbf{I}_p - \mathbf{P}_p)$. Hence, we have (V.1.1).

iii). A direct result of the fact that any pair of the four subspaces in the direct sum must have either a subspace of $\mathbb{N}_{n \times p}$ and a subspace of $\mathbb{N}_{n \times p}^{\perp}$ *or* a subspace of $\mathbb{P}_{n \times p}$ and a subspace of $\mathbb{P}_{n \times p}^{\perp}$.

iv). Directly from the properties of any norm in $\mathbb{M}_{n \times p}$, taking into account the pairwise orthogonality of any pair of terms (from the point above).

□

The key point in the above Theorem is made, for a particular application of decomposition (V.1.1), by Antoniadis *et al.*, (1986) [2]. They consider a two-way ANOVA and decompose a data matrix $\mathbf{Y}$ with $\mathcal{A}_n$ being the subspace spanned by the n-dimensional vector of ones, $\mathbf{1}_n$, and $\mathcal{A}_p$ the subspace spanned by the p-dimensional vectors of ones, $\mathbf{1}_p$. The four terms in (V.1.1) are, respectively, an $n \times p$ matrix of all equal elements (the overall mean of $\mathbf{Y}$); an $n \times p$ matrix of equal columns (each column giving the row means of the column-centered version of $\mathbf{Y}$);

an $n \times p$ matrix of equal rows (each row giving the column means of the row-centered version of $\mathbf{Y}$); an $n \times p$ column-centered and row-centered matrix.

Takane and Shibayama also consider ([64, (Appendix B)]) the case where the inner products defined in $\mathbb{R}^n$ and $\mathbb{R}^p$ are not the standard inner products. In that case, the projection matrices used in decomposition (V.1.1) must be replaced by $\mathbf{N}$-orthogonal and $\mathbf{M}$-orthogonal projectors, whose general form was given in Theorem IV.1. The results of Theorem V.2 remain valid, but with changes brought about by the new inner products in $\mathbb{R}^n$ and $\mathbb{R}^p$, including the use of the $(\mathbf{N},\mathbf{M})$-inner product in $\mathbb{M}_{n \times p}$ (equation (IV.1.3)). As noted in Subsection IV.1.1, orthogonal projectors with generalized inner products correspond to non-orthogonal projectors with the standard inner products.

Sabatier *et al.* (1989) [58] discuss similar ideas and note how several multivariate statistics techniques can be viewed as particular cases of PCAs for certain choices of projection transformations, possibly with alternative inner products.

V.1.2 Effects on PCA

The question of how the PCAs of $\mathbf{Y}$ and the terms resulting from an External Analysis of $\mathbf{Y}$ compare is obviously of interest.

Takane and Shibayama provide an interesting result concerning the singular values of $\mathbf{Y}$ and of $\mathbf{P}_n\mathbf{Y}\mathbf{P}_p$. If we assume that $\mathbf{Y}$ has been column-centered and that $\mathbf{P}_n$ commutes with the column-centering matrix $\mathbf{P}_{\mathbf{1}_p}$, this result will enable us to say something about the variances accounted for by the PCs of $\mathbf{Y}$. Takane and Shibayama's theorem ([64, (Appendix A)]) is given below.

Theorem V.3 (Takane & Shibayama) *Let* $\mathbf{Y} \in \mathbb{M}_{\mathrm{nxp}}$. *Let* $\mathbf{P}_n \in \mathbb{M}_{\mathrm{n}}$ *be a rank* r *orthogonal projection matrix and* $\mathbf{P}_p \in \mathbb{M}_{\mathrm{p}}$ *a rank* q *orthogonal projection matrix. If* $\sigma_j(\cdot)$ *denotes the j-th largest singular value of a matrix, we have:*

$$\sigma_{j+t}(\mathbf{Y}) \leq \sigma_j(\mathbf{P}_n\mathbf{Y}\mathbf{P}_p) \leq \sigma_j(\mathbf{Y}) \quad j = 1, ..., \mathrm{rank}(\mathbf{P}_n\mathbf{Y}\mathbf{P}_p)$$

where $t = n + p - (r + q)$.

It should be stressed that this result implies that the variance accounted for by the j-th PC of $\mathbf{P}_n\mathbf{Y}\mathbf{P}_p$ can never be larger than that accounted for by its equal-ranking counterpart of $\mathbf{Y}$. The result also applies to the case when $\mathbf{P}_n$ and/or $\mathbf{P}_p$ are replaced by $(\mathbf{I}_n - \mathbf{P}_n)$ and/or $(\mathbf{I}_p - \mathbf{P}_p)$, with only the value of t changing. The fact that the r.h.s of the above inequalities is a common upper bound for the j-th singular value of all four terms in (V.1.1) suggests that it will not, in general, be a very tight bound.

A similar result, involving the $\mathbf{N}$- and $\mathbf{M}$-orthogonal projections of the rows and columns of $\mathbf{Y}$ is also given by Takane and Shibayama ([64, (Appendix B)]).

A few words can also be said relating the PAs in $\mathbb{R}^p$ of $\mathbf{P}_n\mathbf{Y}$ to those of $\mathbf{P}_n\mathbf{Y}\mathbf{P}_p$ (possibly with $\mathbf{P}_n = \mathbf{I}_n$) and the PAs in $\mathbb{R}^n$ of $\mathbf{Y}\mathbf{P}_p$ to those of $\mathbf{P}_n\mathbf{Y}\mathbf{P}_p$ (possibly with $\mathbf{P}_p = \mathbf{I}_p$). These will , in fact, be specific applications of general results relating the eigenvectors of a generic $m \times m$ positive definite matrix $\mathbf{V}$ to those of $\mathbf{PVP}$, for some matrix of orthogonal projections $\mathbf{P} \in \mathbb{M}_m$. These results are given in the next Theorem, together with some generalizations to the case when $\mathbf{Q}$ is a matrix of non-orthogonal projections and the matrix $\mathbf{QVQ}'$ is under scrutiny.

Theorem V.4 *Let* $\mathbf{V}$ *be a positive definite matrix in* $\mathbb{S}_{\mathrm{m}}$ *and* $\mathbf{Q} \in \mathbb{M}_{\mathrm{m}}$ *be a rank* r *idempotent matrix. Then:*

i). $\mathbf{QVQ}' \in \mathbb{S}_{\mathrm{m}}$ *and therefore has a complete set of* m *orthonormal eigenvectors.*

ii). Every non-zero vector in $\mathcal{N}(\mathbf{Q}')$ *is an eigenvector of* $\mathbf{QVQ}'$ *with associated eigenvalue zero. Any set of* m *orthonormal eigenvectors of* $\mathbf{QVQ}'$ *must contain* $m - r$ *such vectors in* $\mathcal{N}(\mathbf{Q}')$.

iii). There are exactly r *eigenvectors with non-zero eigenvalues in an orthonormal set of* m *eigenvectors of* $\mathbf{QVQ}'$. *These* r *eigenvectors must all lie in* $\Re(\mathbf{Q})$.

iv). Any eigenvector of $\mathbf{V}$ *which belongs to* $\Re(\mathbf{Q}) \cap \Re(\mathbf{Q}')$ *is also an eigenvector of* $\mathbf{QVQ}'$, *with the same associated eigenvalue.*

Furthermore, if $\mathbf{P}$ *is a rank* r *matrix of* orthogonal *projections, we have:*

v). The eigenvectors of $\mathbf{PVP}$ *associated with non-zero eigenvalues are the vectors which successively maximize the Rayleigh-Ritz ratio* of the original matrix $\mathbf{V}$, *subject to the usual orthogonality constraints* and *to the added constraint of* *belonging to* $\Re(\mathbf{P})$. *The associated eigenvalues are the corresponding values of the Rayleigh-Ritz ratio.*

vi). Say $\mathbf{a}$ *is an eigenvector of* $\mathbf{PVP}$ *associated with a non-zero eigenvalue. The component of* $\mathbf{Va}$ *in* $\Re(\mathbf{P})$ *is a scalar multiple of* $\mathbf{a}$.

vii). If $\{\mathrm{x}_i\}_{i=1}^{m-r}$ *is an orthonormal basis for* $\Re(\mathbf{P})^{\perp}$, *then*

$$\mathrm{tr}(\mathbf{PVP}) = \mathrm{tr}(\mathbf{V}) - \sum_{i=1}^{m-r} \mathbf{x}_i{}'\mathbf{V}\mathbf{x}_i$$

viii). The cosine in $\mathbb{S}_{\mathrm{m}}$ *between* $\mathbf{V}$ *and* $\mathbf{PVP}$ *is given by*

$$\cos(\mathbf{PVP}, \mathbf{V}) = \frac{||\mathbf{PVP}||}{||\mathbf{V}||}$$

Proof.

i). That $\mathbf{QVQ}'$ is symmetric is trivial and the rest follows from a well-known result for symmetric matrices mentioned in Subsection I.3.1.

ii). Trivial, since $\mathbf{x} \in \mathcal{N}(\mathbf{Q}')$ implies that $\mathbf{Q}'\mathbf{x} = 0$, hence $\mathbf{QVQ}'\mathbf{x} = 0$. Since the rank of $\mathbf{Q}'$ is the rank of $\mathbf{Q}$, Proposition A.18 tells us that $\dim(\mathcal{N}(\mathbf{Q}')) = m - r$. The rest follows trivially.

iii). From ii) and the orthogonality requirements on the eigenvectors of $\mathbf{QVQ}'$, it follows that a complete set of eigenvectors of $\mathbf{QVQ}'$ must have r eigenvectors in $\mathcal{N}(\mathbf{Q}')^{\perp} = \Re(\mathbf{Q})$ (see Proposition A.18). That these eigenvectors must all have non-zero eigenvalues follows from the fact that any zero eigenvalues of $\mathbf{QVQ}'$ must imply that the corresponding eigenvector $\mathbf{a}$ annihilates the Rayleigh-Ritz ratio of $\mathbf{QVQ}'$:

$$\frac{\mathbf{a}'\mathbf{QVQ}'\mathbf{a}}{\mathbf{a}'\mathbf{a}} = 0$$

But the assumption that $\mathbf{V}$ is positive definite implies that this can only happen if $\mathbf{Q}'\mathbf{a} = 0$, *i.e.*, if $\mathbf{a} \in \mathcal{N}(\mathbf{Q}') = \Re(\mathbf{Q})^{\perp}$. Hence, the first r eigenvectors of $\mathbf{QVQ}'$ – which are in $\Re(\mathbf{Q})$ – must have non-zero eigenvalues.

iv). If $\mathbf{x} \in \Re(\mathbf{Q}) \cap \Re(\mathbf{Q}')$ is an eigenvector of $\mathbf{V}$ with eigenvalue λ, we have:

$$\mathbf{QVQ}'\mathbf{x} = \mathbf{QVx} = \lambda\mathbf{Qx} = \lambda\mathbf{x}$$

v). As with any symmetric matrix, the eigenvectors of $\mathbf{PVP}$ are successive maximizers of $\mathbf{PVP}$'s Rayleigh-Ritz ratio, subject to the usual orthogonality constraints. But the first r eigenvectors $\{\mathbf{a}_i\}_{i=1}^{r}$ of $\mathbf{PVP}$ are in $\Re(\mathbf{P})$ (as we saw in point iii)). Hence, we have:

$$\frac{\mathbf{a}_i{}'\mathbf{PVP}\mathbf{a}_i}{\mathbf{a}_i{}'\mathbf{a}_i} = \frac{\mathbf{a}_i{}'\mathbf{V}\mathbf{a}_i}{\mathbf{a}_i{}'\mathbf{a}_i} \qquad \text{with } \mathbf{a}_i \in \Re(\mathbf{P})$$

The Rayleigh-Ritz variational characterization of $\mathbf{PVP}$'s first r eigenvectors can then be re-phrased in terms of $\mathbf{V}$'s Rayleigh-Ritz ratios, with the added constraint that the vectors belong to $\Re(\mathbf{P})$.

vi). Say $\mathbf{a}_i \in \Re(\mathbf{P})$ is one of $\mathbf{PVP}$'s first r eigenvectors. We have $\mathbf{PVPa}_i = \lambda\mathbf{a}_i$ for some constant λ. But we also have $\lambda\mathbf{a}_i = \mathbf{PVPa}_i = \mathbf{PVa}_i$. Like any other vector in $\mathbb{R}^m$, the vector $\mathbf{Va}_i$ has a unique decomposition as a sum of terms in $\Re(\mathbf{P})$ and $\Re(\mathbf{P})^\perp$. Say we have $\mathbf{Va}_i = \mathrm{x}_i + \mathrm{y}_i$, with $\mathrm{x}_i \in \Re(\mathbf{P})$, $\mathrm{y}_i \in \Re(\mathbf{P})^\perp$. Then, $\mathbf{PVa}_i = \mathrm{x}_i$. Hence, $\mathrm{x}_i = \lambda\mathbf{a}_i$, that is, the component in $\Re(\mathbf{P})$ of $\mathbf{a}_i$'s image by $\mathbf{V}$ is merely a re-scaling of $\mathbf{a}_i$ itself.

vii). Proposition A.19 tells us that $\mathbf{I}-\mathbf{P}$ is an orthogonal projector on $\Re(\mathbf{P})^\perp$. Hence, if $\{\mathrm{x}_i\}_{i=1}^{m-r}$ is any set of m-r linearly independent vectors in $\Re(\mathbf{P})^\perp$ — that is, a basis for $\Re(\mathbf{P})^\perp$ — we can write $\mathbf{I}-\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, where $\mathbf{X}$ is the $m \times (m-r)$ matrix whose i-th column is x_i. In other words, $\mathbf{P} = \mathbf{I} - \mathbf{P}_X$, where $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Thus,

$$\mathrm{tr}(\mathbf{PVP}) = \mathrm{tr}[\mathbf{V} - \mathbf{VP}_X - \mathbf{P}_X\mathbf{V} + \mathbf{P}_X\mathbf{VP}_X] = \mathrm{tr}(\mathbf{V}) - \mathrm{tr}(\mathbf{VP}_X)$$

Now, if the columns of $\mathbf{X}$ are taken to be orthonormal, we can write $\mathbf{P}_X = \mathbf{XX}'$ and $\mathrm{tr}(\mathbf{VP}_X) = \mathrm{tr}(\mathbf{X}'\mathbf{VX}) = \sum_{\mathrm{i=1}}^{\mathrm{m-r}} \mathrm{x}_i{}'\mathbf{V}\mathrm{x}_i$.

viii). By definition, we have:

$$\cos(\mathbf{PVP},\mathbf{V}) = \frac{<\mathbf{PVP},\mathbf{V}>}{\|\mathbf{PVP}\|\cdot\|\mathbf{V}\|} = \frac{\mathrm{tr}(\mathbf{PVPV})}{\sqrt{\mathrm{tr}(\mathbf{PVPV})\cdot\mathrm{tr}(\mathbf{V}^2)}} = \frac{\|\mathbf{PVP}\|}{\|\mathbf{V}\|}$$

by the properties of traces and symmetric idempotent matrices.

□

Some of the results in Theorem V.4 deserve further comment. A first comment of a general nature is that the key relation being considered is that between $\mathbf{V}$ and $\mathbf{QVQ}'$ (or $\mathbf{PVP}$). The mapping

$$\mathbf{V} \rightarrow \mathbf{QVQ}'$$

is analogous to the orthogonal equivalence mappings $\mathcal{M}_A$ (equation (III.3.8)) and the congruence mappings $\mathcal{N}_U$ (equation (IV.3.3)) with the sole distinction lying in the nature of the matrices $\mathbf{A}$,$\mathbf{U}$ and $\mathbf{Q}$.

If $\mathbf{X}$ is an $n{\times}p$ data matrix and $\mathbf{Q}_p$ is a $p{\times}p$ idempotent matrix, then $\mathbf{XQ}_p{}'$ is projecting the rows of $\mathbf{X}$ onto $\Re(\mathbf{Q}_p)$. It is little surprise that the PAs in $\mathrm{I\!R}^p$ of the new configuration of n points in $\Re(\mathbf{Q}_p)$ — the eigenvectors of $\mathbf{Q}_p\mathbf{SQ}_p{}'$ — are vectors in $\Re(\mathbf{Q}_p)$ (as is guaranteed by point iii) for non-zero eigenvalues) or else vectors which account for none of the variability in the projected scatter (*i.e.*, with associated eigenvalue zero). What was not so predictable — but is pleasing — is that if any PA in $\mathrm{I\!R}^p$ of $\mathbf{X}$ happens to lie in $\Re(\mathbf{Q}_p) \cap \Re(\mathbf{Q}_p{}')$, it remains a PA in $\mathrm{I\!R}^p$ of $\mathbf{XQ}_p{}'$. If $\mathbf{Q}_p$ happens to be an *orthogonal* projector — that is, if it is also a symmetric matrix — then the requirement obviously simplifies to being in the rangespace of the projector. With this additional requirement of symmetry of the projection matrix $\mathbf{P}$, point v) gives the even more pleasing result that the PAs in $\mathrm{I\!R}^p$ of $\mathbf{XP}$ are the *vectors in* $\Re(\mathbf{P})$ which are not just 'best-fitting' for the scatter in $\Re(\mathbf{P})$ defined by $\mathbf{XP}$, but also for the scatter in $\mathrm{I\!R}^p$ defined by $\mathbf{X}$. It is therefore appropriate to speak of a 'PCA with constraints'. For matrices of non-orthogonal projections, a similar characterization in terms of $\mathbf{V}$ only is not as straightforward. The Rayleigh-Ritz ratio of $\mathbf{QVQ}'$ is given by $\frac{\mathbf{x}'\mathbf{QVQ}'\mathbf{x}}{\mathbf{x}'\mathbf{x}}$. We know that the first r successive maxima of this ratio are attained for vectors $\mathbf{x} \in \Re(\mathbf{Q})$ (point iii)), not $\Re(\mathbf{Q}')$. Hence, it need not follow that $\mathbf{Q}'\mathbf{x} = \mathbf{x}$, unless $\mathbf{x} \in \Re(\mathbf{Q}) \cap \Re(\mathbf{Q}')$.

Point vi) in the Theorem tells us, loosely speaking, that eigenvectors of $\mathbf{PVP}$ behave like 'eigenvectors' of $\mathbf{V}$ if everything but $\Re(\mathbf{P})$ is ignored.

The expression for the trace of $\mathbf{PVP}$ given by point vii) is, of course, an alternative to the traditional expressions (sum of diagonal elements or of eigenvalues of $\mathbf{PVP}$) which makes good use of the relation between $\mathbf{V}$ and its 'projected image' $\mathbf{PVP}$. In the specific case when $\Re(\mathbf{P})^{\perp}$ is of dimension 1 — say the subspace spanned by vector $\mathbf{x}$ — the trace of $\mathbf{PVP}$ will be the trace of $\mathbf{V}$ minus the square of the $\mathbf{V}$-norm of $\frac{\mathbf{x}}{||\mathbf{x}||}$. If, for example, $\mathbf{V}$ is the covariance matrix of an $n{\times}p$ data matrix $\mathbf{X}$ and $\mathbf{P}$ is the row-centering projection matrix $\mathbf{P}_{\mathbf{1}_p} = \mathbf{I}_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p{}'$, then the trace of the covariance matrix of $\mathbf{XP}_{\mathbf{1}_p}$ — that is, the total variance of the scatter of n points in the $(p-1)$-dimensional space $\Re(\mathbf{P})^{\perp}$ — will be given by

$$\mathrm{tr}(\mathbf{V}) - \frac{\mathbf{1}_p{}'\mathbf{V}\mathbf{1}_p}{||\mathbf{1}_p||^2} = \mathrm{tr}(\mathbf{V}) - \tfrac{1}{p}\mathrm{sum}(\mathbf{V})$$

where $\mathrm{sum}(\mathbf{V})$ is the sum of $\mathbf{V}$'s elements.

The above comments on the relations between the PAs in $\mathbb{R}^p$ of $\mathbf{X}$ and those of $\mathbf{XQ}_p{}'$ will also apply to the relations between the PAs in $\mathbb{R}^n$ (unit-norm PCs) of $\mathbf{X}$ and those of $\mathbf{Q}_n\mathbf{X}$ for any $n \times n$ idempotent matrix $\mathbf{Q}_n$ which commutes with the column-centering projector $\mathbf{P}_{\mathbf{1}_n}$. In that case, we take $\mathbf{V} = \frac{1}{n}\mathbf{X}\mathbf{X}'$ and the PAs in $\mathbb{R}^n$ of $\mathbf{Q}_n\mathbf{X}$ are the eigenvectors of $\mathbf{Q}_n\mathbf{V}\mathbf{Q}_n{}'$.

However, nothing has as of yet been said about the relation between the PAs in $\mathbb{R}^p$ of $\mathbf{X}$ and those of $\mathbf{Q}_n\mathbf{X}$, nor about the relation between the PCs of $\mathbf{X}$ and those of $\mathbf{X}\mathbf{Q}_p{}'$. In such cases, the results of Theorem V.4 do not apply directly, since we would be talking about the eigenvectors of $\mathbf{X}'\mathbf{Q}_n{}'\mathbf{Q}_n\mathbf{X}$ or of $\mathbf{X}\mathbf{Q}_p{}'\mathbf{Q}_p\mathbf{X}'$. But it is possible to make use of the relations between both sets of PAs of a given matrix to say more about this issue.

Theorem V.5 *Let* $\mathbf{Z}$ *be an* $\mathrm{m} \times \mathrm{q}$ *matrix of rank* r . *Let* $\{(\lambda_i, \mathbf{a}_i)\}_{i=1}^{r}$ *be the eigenpairs with non-zero eigenvalues of the matrix* $\mathbf{V} = \mathbf{Z}'\mathbf{Z}$. *Let* $\mathbf{Q}$ *be a* $\mathrm{q} \times \mathrm{q}$ *idempotent matrix of rank* s *and* $\{(\chi_j, \mathbf{q}_j)\}_{j=1}^{s}$ *the eigenpairs with non-zero eigenvalues of* $\mathbf{QVQ}'$. *Then, for all* $i = 1, ..., r$; $j = 1, ..., s$, *we have:*

i). $< \mathbf{Z}\mathbf{a}_i, \mathbf{Z}\mathbf{Q}'\mathbf{q}_j >_{\mathbb{R}^m} = \lambda_i < \mathbf{a}_i, \mathbf{Q}'\mathbf{q}_j >_{\mathbb{R}^q} = \lambda_i < \mathbf{Q}\mathbf{a}_i, \mathbf{q}_j >_{\mathbb{R}^q}$

ii). $\cos_{\mathbb{R}^m}(\mathbf{Z}\mathbf{a}_i, \mathbf{Z}\mathbf{Q}'\mathbf{q}_j) = \sqrt{\frac{\lambda_i}{\chi_j}} < \mathbf{a}_i, \mathbf{Q}'\mathbf{q}_j >_{\mathbb{R}^q} = \sqrt{\frac{\lambda_i}{\chi_j}} < \mathbf{Q}\mathbf{a}_i, \mathbf{q}_j >_{\mathbb{R}^q}$

If, furthermore, the projection matrix $\mathbf{Q}$ *is also symmetric, we have:*

iii). $< \mathbf{Z}\mathbf{a}_i, \mathbf{Z}\mathbf{Q}\mathbf{q}_j >_{\mathbb{R}^m} = \lambda_i \cos_{\mathbb{R}^q}(\mathbf{a}_i, \mathbf{q}_j)$

iv). $\cos_{\mathbb{R}^m}(\mathbf{Z}\mathbf{a}_i, \mathbf{Z}\mathbf{Q}\mathbf{q}_j) = \sqrt{\frac{\lambda_i}{\chi_j}} \cos_{\mathbb{R}^q}(\mathbf{a}_i, \mathbf{q}_j)$

v). $\cos^2_{\mathbb{R}^m}(\mathbf{Za}_i, \mathbf{ZQq}_j) \leq \frac{\lambda_i}{\chi_j}$

vi). $\cos^2_{\mathbb{R}^m}(\mathbf{a}_i, \mathbf{q}_j) \leq \frac{\chi_j}{\lambda_i}$

vii). $\chi_j = \sum_{i=1}^r \lambda_i \cos^2_{\mathbb{R}^q}(\mathbf{a}_i, \mathbf{q}_j)$

Proof.

i). We have:

$$\begin{aligned} < \mathbf{Za}_i, \mathbf{ZQ'q}_j >_{\mathbb{R}^m} &= \mathbf{a}_i'(\mathbf{Z'Z})\mathbf{Q'q}_j = \lambda_i \mathbf{a}_i'\mathbf{Q'q}_j \\ &= \lambda_i < \mathbf{a}_i, \mathbf{Q'q}_j >_{\mathbb{R}^q} = \lambda_i < \mathbf{Qa}_i, \mathbf{q}_j >_{\mathbb{R}^q} \end{aligned}$$

ii). Direct from point i) and the fact that $||\mathbf{Za}_i|| = \sqrt{\lambda_i}$ and $||\mathbf{ZQ'q}_j|| = \sqrt{\chi_j}$

The added assumption that $\mathbf{Q}$ is a matrix of *orthogonal* projections implies (Theorem V.4, point iii)) that $\mathbf{Q} = \mathbf{Q'}$ with $\mathbf{q}_j \in \Re(\mathbf{Q})$. So $\mathbf{Q'q}_j = \mathbf{Qq}_j = \mathbf{q}_j$. Points iii) and iv) flow directly from that. Points v) and vi) are trivial maximizations resulting from point iv). Finally, the Rayleigh-Ritz characterization of the eigenpairs of $\mathbf{QVQ}$ gives $\chi_j = \mathbf{q}_j'\mathbf{QVQq}_j = \mathbf{q}_j'\mathbf{Vq}_j$ the second equality again resulting from $\mathbf{q}_j \in \Re(\mathbf{Q}) = \Re(\mathbf{Q'})$. But the spectral decomposition of $\mathbf{V}$ is $\mathbf{V} = \sum_{i=1}^r \lambda_i \mathbf{a}_i \mathbf{a}_i'$. Hence, $\chi_j = \sum_{i=1}^r \lambda_i (\mathbf{a}_i'\mathbf{q}_j)^2$. □

As was seen in Chapter I, if $\mathbf{X}$ is an $n \times p$ column-centered data matrix and we take $\mathbf{Z} = \frac{1}{\sqrt{n}}\mathbf{X}$, then the vectors $\{\mathbf{Za}_i\}_{i=1}^r$ are the PCs of $\mathbf{X}$ and the vectors $\{\mathbf{ZQ'q}_j\}_{j=1}^s$ are the PCs of $\mathbf{XQ'}$. The results in points i) and ii) of the previous theorem are then the analogues of those in Theorem IV.9, restricted by the non-invertibility of $\mathbf{Q}$. The inner products in $\mathbb{R}^n$ are the covariances of the PCs. The cosines of the angles in $\mathbb{R}^n$ are the correlations between PCs. On the other hand, if we took $\mathbf{Z} = \frac{1}{\sqrt{n}}\mathbf{X}'$, then the vectors $\{\mathbf{a}_i\}_{i=1}^r$ are the unit-norm PCs of $\mathbf{X}$ and the vectors $\{\mathbf{Za}_i\}_{i=1}^r$ are the natural length PAs in $\mathbb{R}^p$ of $\mathbf{X}$ (defined in Subsection I.4.2). Likewise, the vectors $\{\mathbf{ZQ}'\mathbf{q}_j\}_{j=1}^s$ are the natural length PAs in $\mathbb{R}^p$ of $\mathbf{QX}$.

The results of the Theorem are particularly pleasing in the case of an orthogonal projection matrix $\mathbf{P}$. *The cosine of the angle between any PC of* $\mathbf{X}$ *and any (non-zero length) PC of* $\mathbf{XP}$ *is the cosine of the angle in* $\mathbb{R}^p$ *between their vectors of loadings, times a scaling factor given by the ratio of the norm of the PC of* $\mathbf{X}$ *over the norm of the PC of* $\mathbf{XP}$ (point iv)). One implication of this result is that if $\lambda_i \ll \chi_j$, then the i-th PC of $\mathbf{X}$ must be nearly orthogonal (uncorrelated) to the j-th PC of $\mathbf{XP}$. Also, if $\chi_j \ll \lambda_i$, the i-th PA in $\mathbb{R}^p$ of $\mathbf{X}$ and the j-th PA in $\mathbb{R}^p$ of $\mathbf{XP}$ must be nearly orthogonal. In addition, for $\lambda_i \simeq \chi_j$, the angles in $\mathbb{R}^p$ and the corresponding angles in $\mathbb{R}^n$ between corresponding PAs will be approximately equal.

Point vii) gives an *explicit formula for the non-zero eigenvalues of* $\mathbf{PVP}$, *as a weighted average of the eigenvalues of* $\mathbf{V}$. In fact, the weights $\cos^2(\mathbf{a}_i, \mathbf{q}_j)$ must add up (over i) to 1, since the vector $\mathbf{q}_j$ lies in the space for which the $\{\mathbf{a}_i\}_{i=1}^r$ are an orthonormal basis. The proof of this result in the Theorem could also have been obtained directly from point iv) by Pythagoras' Theorem since the vectors $\{\mathbf{Za}_i\}_{i=1}^r$ are an orthogonal basis for the subspace of $\mathbb{R}^m$ containing $\mathbf{ZPq}_j$. It is not possible to reverse the role of the eigenpairs of $\mathbf{V}$ and $\mathbf{PVP}$ in point vii), since the subspace spanned in $\mathbb{R}^m$ by the columns of $\mathbf{ZP}$ need not contain $\mathbf{Za}_i$.

A consequence of point vii) is that if $\cos^2(\mathbf{a}_i, \mathbf{q}_j) \simeq 1$, for some i and some j, then $\cos^2(\mathbf{a}_k, \mathbf{q}_j) \simeq 0$, for $k \neq i$ (since $\sum_{i=1}^{p} \cos^2(\mathbf{a}_i, \mathbf{q}_j) = 1$) and therefore $\chi_j \simeq \lambda_i$. This implies that *a necessary condition for an eigenvector of* $\mathbf{V}$ *to form a small angle with a (non-zero eigenvalue) eigenvector of* $\mathbf{PVP}$ *is that their corresponding eigenvalues be approximately equal.* Thus, the eigenvalues of $\mathbf{V}$ and $\mathbf{PVP}$ tell us whether it is possible for some eigenvectors of both matrices to be approximately equal.

But Theorem V.3 enables us to go one step further with a general conclusion that does not require knowledge of the eigenvalues in a specific case. The bounds on the singular values of $\mathbf{ZP}$ — hence on the eigenvalues of $\mathbf{PVP}$ — help us locate which eigenvectors of $\mathbf{V}$ can be close to a given eigenvector of $\mathbf{PVP}$. For example, if $\mathbf{P}$ is a projector of rank $p-1$ (as is the row-centering projector $\mathbf{P}_{\mathbf{1}_p}$) then the first eigenvector of $\mathbf{PVP}$ can only be similar to the first or second eigenvectors of $\mathbf{V}$ (with the possible exception of cases where the second eigenvalue of $\mathbf{V}$ is very close to the third, or even subsequent, eigenvalues). In addition, if k $(k > 1)$ eigenvectors of $\mathbf{V}$ have very similar counterparts among the eigenvectors of $\mathbf{PVP}$, the relative ranking of the two sets of k eigenvectors must be the same (again, with the possible exception of cases where the eigenvalues are poorly separated). These ideas will be illustrated in Subsections V.2.3 and V.3.3.

V.1.3 A special class of projectors in $\mathbb{R}^n$

In many statistical techniques we encounter projections onto subspaces of $\mathbb{R}^n$ which depend on the data values. In this Subsection we shall see how a class of such projections is always equivalent to projecting in the space $\mathbb{R}^p$ and that, conversely, all projections in $\mathbb{R}^p$ can be written as projections onto subspaces of $\mathbb{R}^n$. For simplicity, it is assumed that the inner products in both spaces are the standard inner products.

Theorem V.6 *Let* $\mathbf{Y}$ *be an* $\mathrm{n}\times\mathrm{p}$ *matrix of rank* p . *Let* $\mathbf{A},\mathbf{B} \in \mathbb{M}_{\mathrm{p}\times\mathrm{q}}$ *be rank* q *matrices. Let* $\mathbf{Q}_p \in \mathbb{M}_{\mathrm{p}}$ *be the matrix projecting onto* $\Re(\mathbf{A})$, *along* $\Re(\mathbf{B})^{\perp}$. *Let* $\mathbf{Q}_n \in \mathbb{M}_{\mathrm{n}}$ *be the matrix projecting onto* $\Re(\mathbf{YB})$ *along* $\Re(\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{A})^{\perp}$. *Then,*

$$\mathbf{Q}_n\mathbf{Y} = \mathbf{Y}\mathbf{Q}_p{}'$$

Proof. From Proposition A.19, we know that $\mathbf{Q}_p$ and $\mathbf{Q}_n$ can be written as $\mathbf{Q}_p = \mathbf{A}(\mathbf{B}'\mathbf{A})^{-1}\mathbf{B}'$ and

$$\mathbf{Q}_n = (\mathbf{YB})([\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{A}]'\mathbf{YB})^{-1}[\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{A}]' = \mathbf{YB}(\mathbf{A}'\mathbf{B})^{-1}\mathbf{A}'(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'$$

But then, $\mathbf{Q}_n\mathbf{Y} = \mathbf{YB}(\mathbf{A}'\mathbf{B})^{-1}\mathbf{A}' = \mathbf{Y}\mathbf{Q}_p$. □

What the Theorem is telling us is that *any projection of the rows of* $\mathbf{Y}$ *onto some subspace of* $\mathbb{R}^p$ *corresponds to a projection of the columns of* $\mathbf{Y}$ *onto some subspace of* $\Re(\mathbf{Y})$, that is, some subspace of the space in $\mathbb{R}^n$ spanned by $\mathbf{Y}$'s columns.

Orthogonal projections in $\mathbb{R}^p$ correspond to taking $\mathbf{B} = \mathbf{A}$. Orthogonal projections in $\mathbb{R}^n$ correspond to taking $\mathbf{A} = (\mathbf{Y}'\mathbf{Y})\mathbf{B}$. *Whenever $\Re(\mathbf{B})$ is the subspace spanned by some set of eigenvectors of $\mathbf{Y}'\mathbf{Y}$, we have orthogonal projections simultaneously in $\mathbb{R}^n$ and in $\mathbb{R}^p$.* This, of course, is at the heart of the very method of Principal Component Analysis, as discussed in Chapter I.

If $\mathbf{Y}$ is a column-centered data matrix, we can also re-phrase a special case of the above Theorem, using the **Mahalanobis inner product**, defined by the inverse of the covariance matrix (Mardia, Kent and Bibby (1979) [42, (pg.17)]).

Corollary *Let $\mathbf{X} \in \mathbb{C}_{\mathrm{n \times p}}$ be a column-centered $\mathrm{n \times p}$ data matrix of rank p and $\mathbf{S} \in \mathbb{S}_{\mathrm{p}}$ the covariance matrix it defines. Let $\mathbf{B} \in \mathbb{M}_{\mathrm{p \times q}}$ be a rank q matrix. Orthogonally projecting the columns of $\mathbf{X}$ onto $\Re(\mathbf{XB})$ is equivalent to $\mathbf{S}^{-1}$-orthogonally projecting the rows of $\mathbf{X}$ onto $\Re(\mathbf{SB})$.*

Proof. We know from the Theorem that an orthogonal projection in $\mathbb{R}^n$ onto $\Re(\mathbf{XB})$ corresponds to a projection in $\mathbb{R}^p$ onto $\Re(\mathbf{SB})$ along $\Re(\mathbf{B})^{\perp}$. The projection matrix in $\mathbb{R}^p$ can therefore be written as

$$\begin{aligned} \mathbf{Q}_p &= (\mathbf{SB})(\mathbf{B}'\mathbf{SB})^{-1}\mathbf{B}' \\ &= (\mathbf{SB})[(\mathbf{SB})'\mathbf{S}^{-1}(\mathbf{SB})]^{-1}(\mathbf{SB})'\mathbf{S}^{-1} \end{aligned}$$

Hence, by (IV.1.2), $\mathbf{Q}_p$ is an $\mathbf{S}^{-1}$-orthogonal projector onto $\Re(\mathbf{SB})$. □

V.1.4 Comparing PCAs of projected data matrices

In the last two Sections of Chapter IV, means of comparing the PCAs of two or more transformed data matrices were discussed. With the exception of comparisons involving the inversion of the transformations, those methods will also apply if some or all of the transformation matrices are projections.

In particular, the similarity indices — global, in $\mathbb{R}^n$ or in $\mathbb{R}^p$ — are valid regardless of whether or not the transformation matrix is idempotent (a projection) and of whether the full rank or a rank k approximation of the matrices is used. Likewise, the multiple comparison based on matrices of matrix correlations are valid when some or all the matrices which are being correlated arise from projections.

In later Sections we shall give examples of such comparisons. For now, we can illustrate their possible use with a global comparison of the (column-centered) PCA and the non-centered PCA of any data matrix.

A global comparison of PCA and non-centered PCA

As was seen in Subsection I.4.2, the implication of skipping the column-centering stage of PCA is that all references to variances of the p variables must be replaced by references to the non-central second order moments of those variables.

The global similarity index s_g for the comparison of a PCA and a non-centered

PCA of a given $n \times p$ data matrix $\mathbf{Y}$ can be written as:

$$s_g(\mathbf{Y}, \mathbf{P}_{\mathbf{1}_n}\mathbf{Y}) = \cos(\mathbf{Y}, \mathbf{P}_{\mathbf{1}_n}\mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{Y}'\mathbf{P}_{\mathbf{1}_n}\mathbf{Y})}{\sqrt{\mathrm{tr}(\mathbf{Y}'\mathbf{P}_{\mathbf{1}_n}\mathbf{Y}) \cdot \mathrm{tr}(\mathbf{Y}'\mathbf{Y})}} = \frac{||\mathbf{P}_{\mathbf{1}_n}\mathbf{Y}||}{||\mathbf{Y}||}$$

since $\mathbf{P}_{\mathbf{1}_n}$ is a symmetric, idempotent matrix. From the definition of the Frobenius norm, we then have:

$$s_g(\mathbf{Y}, \mathbf{P}_{\mathbf{1}_n}\mathbf{Y}) = \sqrt{\frac{\sum_{i=1}^{p} s_{ii}}{\sum_{i=1}^{p} m_{ii}}}$$

where s_{ii} is the variance of the i-th column of $\mathbf{Y}$ and m_{ii} is the non-central second order moment of the i-th column of $\mathbf{Y}$. Appropriately, the more similar are the central and non-central moments of $\mathbf{Y}$'s columns, the closer the global similarity index of the PCA with its non-central counterpart will be to its maximum value of one.

Comparable expressions can also be obtained for the similarity indices in $\mathbb{R}^n$ (using point viii) of Theorem V.4) and $\mathbb{R}^p$. Similar results can also be obtained for comparisons involving doubly-centered and row-centered PCA.

V.2 Removing isometric size from morphometric data

V.2.1 The problem and an overview of past work

Principal Component Analysis has been used in connection with the study of *size* and *shape* of biological organisms. Huxley (1972) [28] suggested that the relationship between the size y of a given organ or part of an organism and the size x of the whole organism could often be described by the **allometric growth equation**,

$$y = bx^a \tag{V.2.1}$$

The constant a represents the relative rate of growth of the organ or part. For $a = 1$ this growth is, relatively speaking, equal to that of the organism as a whole. In that case we speak of *isometric* growth. When $a > 1$, the organ's size increases faster than that of the organism and we speak of *positive allometry.* A slower relative growth of the organ, *i.e.*, $a < 1$ is termed *negative allometry.*

Since the publication in 1932 of the first edition of Huxley's book, many applications of this equation have been found (see the Introduction to the 1972 edition [28]). In 1960, Teissier [66, (pg.547)] suggested a generalization of the allometry equation to the case of m morphometric variables. Jolicoeur (1963) [31] proposed a modification to Teissier's generalization which will now be considered.

The line defined by the first eigenvector of a covariance matrix of p variables $\{\mathbf{y}_i\}_{i=1}^{p}$ can be described by the equations (see Kendall (1960) [37, (pg.7)]):

$$\frac{y_1 - \overline{y}_1}{l_1} = \frac{y_2 - \overline{y}_2}{l_2} = \cdots = \frac{y_p - \overline{y}_p}{l_p}$$

where $\mathbf{y} = (y_1, ..., y_p)$ is any point on the line, $\overline{\mathbf{y}} = (\overline{y}_1, ..., \overline{y}_p)$ is the center of gravity of the p variables in $\mathbb{R}^p$ and $\{l_i\}_{i=1}^{p}$ are the direction cosines of the line with the m axes. Jolicoeur notes that if we considered the covariance matrix of the *log-transformed* *data*, that is, if $\mathbf{y}_i = \log(\mathbf{z}_i)$ where $\{\mathbf{z}_i\}_{i=1}^{p}$ were the original variables, then the above equations become:

$$\left(\frac{z_1}{g_1}\right)^{\frac{1}{l_1}} = \left(\frac{z_2}{g_2}\right)^{\frac{1}{l_2}} = \cdots = \left(\frac{z_p}{g_p}\right)^{\frac{1}{l_p}} \tag{V.2.2}$$

where g_i is the geometric mean of the i-th log-transformed variable $\mathbf{z}_i$. It should be stressed that it is standard practice in allometric studies to work with log-transformations of the data (see, for example, Teissier (1960) [66, (pg.544)]).

Now, any single equation in (V.2.2) describes an allometric relation of the type (V.2.1) with $a = \frac{l_j}{l_k}$ the ratio of the direction cosines.

Jolicoeur thus suggests ([31, (pg.497)]) "that the most straightforward manner to generalize the allometry equation is to use the first principal component of the covariance matrix of logarithms", where 'principal component' means PA in $\mathbb{R}^p$, in our terminology. In particular, the isometric case — now corresponding to a proportionally equal growth of all p variables — is associated with the covariance matrix of log-transformed data having $\frac{1}{\sqrt{p}}\mathbf{1}_p$ as its first eigenvector. This vector we shall call the **isometric size vector in $\mathbb{R}^p$**. Any other first eigenvector corresponds to non-isometric growth for at least one pair of variables.

Jolicoeur's proposal differs from that of Teissier in that the latter suggested using the first eigenvector of the correlation, rather than the covariance, matrix. But as Jolicoeur points out ([31, (pg.497)]):

> The major reason for using the correlation matrix, however, is that it

> tends to make principal components independent from the order of magnitude and the scale of measurement of the variables. But the logarithmic transformation generally carried out on relative growth data tends in itself to make the covariance matrix independent from magnitude and scaling. Consequently, the use of the correlation matrix instead of the covariance matrix would have little additional advantage, if any, and it would make final interpretations more complicated.

In what follows we assume that covariance matrices are being used, although the use of correlation matrices would not pose major technical difficulties.

The above discussion focused on concepts in $\mathbb{R}^p$. But the PC (in $\mathbb{R}^n$) defined by an isometric size eigenvector may also prove useful in its own right (see also Teissier (1960) [66, (pgs.547-549)]). Rao (1964) [54], also citing work by other authors, calls a linear combination of p variables

$$\mathbf{x} = \sum_{i=1}^{p} c_i \mathbf{x}_i$$

a **size factor** if all the coefficients $\{c_i\}_{i=1}^{p}$ are positive and a **shape factor** if there are coefficients of different signs.

Since the nature of morphometric data entails positive covariances and correlations between the variables, any such factors defined using the first PC of the data matrix whose i-th column is $\mathbf{x}_i$ will be size factors. In fact, the first eigenvector of a **non-negative matrix** (*i.e.*, a matrix with all non-negative elements) can be taken to have only non-negative coefficients (see Minc (1988) [45, (pg.14)]). Orthogonality requirements will tend to force the remaining PCs to be shape factors — necessarily so if the first eigenvector has no zero coefficients. The size factor resulting from Jolicoeur's isometry hypothesis is $\frac{1}{\sqrt{p}}\mathbf{X}\mathbf{1}_p$. We shall henceforth call this vector of $\mathbb{R}^n$ the **isometric size factor** or **isometric size vector in $\mathbb{R}^n$**.

Somers (1989) [62] states that "*because allometric size represents bivariate shape that is correlated with isometric size, PC1 frequently contains information in both size and shape*". He therefore tackles the problem of removing isometric size relations.

Somers had previously suggested what, in his words [62, (pg.169)], was a "modification to PCA that produced a first axis summarizing variation in isometric size alone". This modification, which he called **size-constrained PCA**, essentially amounts to subtracting an eigendecomposition-like term from the correlation matrix of log-transformed data (where the vector playing the role of the 'eigenvector' is the isometric size vector $\frac{1}{\sqrt{p}}\mathbf{1}_p$) and then spectrally decomposing the residual matrix. One problem with this idea is that the residual matrix is no longer, in general, positive semi-definite and it is not clear what meaning should be attached to negative eigenvalues. Other drawbacks were pointed out by several authors and are reviewed by Somers ([62]). Prominent among these problems is the fact that the isometric vector $\frac{1}{\sqrt{p}}\mathbf{1}_p$ is not orthogonal to the new 'size-constrained PAs in $\mathbb{R}^p$' and the isometric size factor $\frac{1}{\sqrt{p}}\mathbf{X}\mathbf{1}_p$ is not uncorrelated with (orthogonal in $\mathbb{R}^n$ to) the new 'size-constrained PCs'.

The first of these problems is dealt with by Somers in [62, (pg.171)]. Based on personal suggestions by W. Krzanowski and J.C. Gower, Somers considers doubly-centering either the correlation or the covariance matrix of the log-transformed data. As we saw in Section V.1, this corresponds to an eigendecomposition of $\mathbf{P}_{\mathbf{1}_p}\mathbf{R}\mathbf{P}_{\mathbf{1}_p}$ or $\mathbf{P}_{\mathbf{1}_p}\mathbf{S}\mathbf{P}_{\mathbf{1}_p}$, where $\mathbf{R}$,$\mathbf{S}$ are the correlation and covariance matrices of the log-transformed data. Since $\mathbf{P}_{\mathbf{1}_p}\mathbf{R}\mathbf{P}_{\mathbf{1}_p}$ is no longer a correlation matrix, Jolicoeur's preference for working with covariance matrices is all the more appropriate.

Since $\mathcal{N}(\mathbf{P}_{\mathbf{1}_p})$ is the one-dimensional subspace spanned by $\mathbf{1}_p$, we can see from Theorem V.4 that $\mathbf{P}_{\mathbf{1}_p}\mathbf{S}\mathbf{P}_{\mathbf{1}_p}$ will have a single zero eigenvalue (if $\mathbf{S}$ is of full rank) whose corresponding eigenvector is $\frac{1}{\sqrt{p}}\mathbf{1}_p$. The orthogonality constraints imply that the remaining $p-1$ PAs in $\mathbb{R}^p$ of $\mathbf{P}_{\mathbf{1}_p}\mathbf{S}\mathbf{P}_{\mathbf{1}_p}$ are indeed orthogonal to $\mathbf{1}_p$, as required. (It should be said that any attempt to replace $\mathbf{P}_{\mathbf{1}_p}\mathbf{S}\mathbf{P}_{\mathbf{1}_p}$– or $\mathbf{P}_{\mathbf{1}_p}\mathbf{R}\mathbf{P}_{\mathbf{1}_p}$ – with their corresponding correlation matrices would destroy these characteristics).

But as Ranatunga (1989) [53] points out, the PCs defined by these first $p-1$ PAs in $\mathbb{R}^p$ of $\mathbf{P}_{\mathbf{1}_p}\mathbf{S}\mathbf{P}_{\mathbf{1}_p}$ are *not*, in general, orthogonal to (uncorrelated with) the isometric size factor $\frac{1}{\sqrt{p}}\mathbf{X}\mathbf{1}_p$. Theorem V.6 explains why this is so. Doubly-centering corresponds to a projection of the columns of $\mathbf{X}$ onto $\Re(\mathbf{X}\mathbf{S}^{-1}\mathbf{1}_p)^{\perp}$ along $\Re(\mathbf{X}\mathbf{1}_p)$. Hence, the PCs of $\mathbf{X}\mathbf{P}_{\mathbf{1}_p}$ are orthogonal to $\mathbf{X}\mathbf{S}^{-1}\mathbf{1}_p$, not to $\mathbf{X}\mathbf{1}_p$ (unless $\mathbf{1}_p$ is already an eigenvector of $\mathbf{S}$, hence of $\mathbf{S}^{-1}$).

Let us take a closer look at this point. Doubly-centering $\mathbf{S}$ implies that the data are, in $\mathbb{R}^p$, orthogonally projected onto $\Re(\mathbf{1}_p)^\perp$. The eigenvectors of $\mathbf{P}_{\mathbf{1}_p}\mathbf{S}\mathbf{P}_{\mathbf{1}_p}$ are loadings for PCs of $\mathbf{X}\mathbf{P}_{\mathbf{1}_p}$. If $\mathbf{a} \in \Re(\mathbf{1}_p)^\perp$ is one such eigenvector, then the PC is $\mathbf{X}\mathbf{P}_{\mathbf{1}_p}\mathbf{a} = \mathbf{X}\mathbf{a}$. But for the eigenvector $\frac{1}{\sqrt{p}}\mathbf{1}_p$, the associated PC is $\mathbf{X}\mathbf{P}_{\mathbf{1}_p}\mathbf{1}_p = \mathbf{0} \in \mathbb{R}^n$, not $\mathbf{X}\mathbf{1}_p$. Thus, whilst the eigenvector $\mathbf{a}$ must be orthogonal to the eigenvector $\mathbf{1}_p$, the PC (of $\mathbf{X}\mathbf{P}_{\mathbf{1}_p}$) $\mathbf{X}\mathbf{a}$ is not, in general, orthogonal to the vector $\mathbf{X}\mathbf{1}_p$ (which is not a PC of either $\mathbf{X}$ or $\mathbf{X}\mathbf{P}_{\mathbf{1}_p}$).

The lack of orthogonality of the PCs of $\mathbf{X}\mathbf{P}_{\mathbf{1}_p}$ with the isometric size factor $\mathbf{X}\mathbf{1}_p$ will tend to worsen as the angle between $\mathbf{X}\mathbf{1}_p$ and $\mathbf{X}\mathbf{S}^{-1}\mathbf{1}_p$ (to which the PCs are orthogonal) grows. The cosine of this angle can be written as:

$$\cos(\mathbf{X}\mathbf{1}_p, \mathbf{X}\mathbf{S}^{-1}\mathbf{1}_p) = \frac{p}{\sqrt{\mathrm{sum}(\mathbf{S}) \cdot \mathrm{sum}(\mathbf{S}^{-1})}}$$

where sum$(\cdot)$ denotes the sum of all matrix elements. Thus, the larger the product of the sums of elements in $\mathbf{S}$ and $\mathbf{S}^{-1}$, the worse we shall tend to be in terms of the uncorrelatedness of the PCs of $\mathbf{X}\mathbf{P}_{\mathbf{1}_p}$ with the isometric size factor.

Noting that the 'shape PCs' of $\mathbf{X}\mathbf{P}_{\mathbf{1}_p}$ were not uncorrelated with the isometric size factor $\mathbf{X}\mathbf{1}_p$, Ranatunga suggested ([53, (Chapter 3)]) an alternative method of tackling the problem. In our terminology, her suggestion is that the rows of the (log-transformed) data matrix $\mathbf{X}$ be orthogonally projected onto $\Re(\mathbf{S}\mathbf{1}_p)^\perp$, rather than onto $\Re(\mathbf{1}_p)^\perp$. In other words, that we carry out a PCA of $\mathbf{X}\mathbf{P}_{S\mathbf{1}_p}$, where $\mathbf{P}_{S\mathbf{1}_p}$ is the orthogonal projection matrix onto the orthogonal complement of the *vector of row* *sums of* **S**. This projector can be written (see Proposition A.12) as:

$$\mathbf{P}_{S\mathbf{1}_p} = \mathbf{I}_p - (\mathbf{S1}_p)(\mathbf{1}_p{}'\mathbf{S}^2\mathbf{1}_p)^{-1}(\mathbf{S1}_p)' = \mathbf{I}_p - \frac{\mathbf{S1}_p\mathbf{1}_p{}'\mathbf{S}}{\text{sum}(\mathbf{S}^2)} \tag{V.2.3}$$

where $\text{sum}(\mathbf{S}^2)$ is the sum of all elements of $\mathbf{S}^2$.

That Ranatunga's method achieves its stated goal of obtaining shape factors which are uncorrelated to the isometric size factor $\frac{1}{\sqrt{p}}\mathbf{X1}_p$ is best seen using Theorem V.6. Ranatunga's projector $\mathbf{P}_{S\mathbf{1}_p}$ is a projector in $\mathbb{R}^p$ onto $\Re(\mathbf{S1}_p)^\perp$, along $\Re(\mathbf{S1}_p)$. Hence, combining Theorem V.6 with Proposition A.19 (point iv)) we can conclude that $\mathbf{XP}_{S\mathbf{1}_p} = \mathbf{Q}_n\mathbf{X}$ where $\mathbf{Q}_n$ projects the columns of $\mathbf{X}$ onto $\Re(\mathbf{X1}_p)^\perp$ along $\Re(\mathbf{XS1}_p)$. Thus, $\mathbf{Q}_n$ can be written as:

$$\mathbf{Q}_n = \mathbf{I}_n - (\mathbf{XS1}_p)[(\mathbf{X1}_p)'(\mathbf{XS1}_p)]^{-1}(\mathbf{X1}_p)' = \mathbf{I}_n - \frac{\mathbf{XS1}_p\mathbf{1}_p{}'\mathbf{X}'}{\text{sum}(\mathbf{S}^2)} \tag{V.2.4}$$

The columns of $\mathbf{XP} = \mathbf{Q}_n\mathbf{X}$, and by Theorem V.4.iii) its PCs, must therefore lie in $\Re(\mathbf{X1}_p)^\perp$, ensuring the required orthogonality.

But if Ranatunga's method does put right the situation in $\mathbb{R}^n$, it does so at the expense of the orthogonality of the PAs in $\mathbb{R}^p$ of the projected data matrix with the isometry vector $\frac{1}{\sqrt{p}}\mathbf{1}_p$. In fact, the PAs in $\mathbb{R}^p$ of $\mathbf{XP}$ are either $\mathbf{S1}_p$ (with associated eigenvalue zero) or orthogonal to $\mathbf{S1}_p$ — with no guarantee of orthogonality with $\mathbf{1}_p$ (again, unless $\mathbf{1}_p$ is already an eigenvector of $\mathbf{S}$). Ranatunga notes this, but stresses that the lack of orthogonality in $\mathbb{R}^p$ seems to be relatively minor, for the examples considered, and can be safely ignored. Whilst this is true (essentially because for positive matrices, as are morphometric covariance matrices, $\mathbf{S1}_p$ will always be in the same orthant as $\mathbf{1}_p$, thus making $\mathbf{1}_p$ an 'approximation' to $\mathbf{S1}_p$, in particular when the row sums of $\mathbf{S}$ are approximately equal), a modification to the method will enable us to achieve orthogonality with the isometric size vectors in both spaces, at no extra cost.

V.2.2 A new solution

In this Subsection we propose a solution to the problem of removing isometric size from a data set which ensures that the PAs *in both* $\mathbb{R}^p$ *and* $\mathbb{R}^n$ will be orthogonal to the isometry vectors.

The key notion is to merge the relevant aspects of the doubly-centering and the Ranantunga approaches, retaining those aspects of each which guarantee the required orthogonality in each space.

The Krzanowski/Gower approach consists of projecting the rows of $\mathbf{X}$ *onto* $\Re(\mathbf{1}_p)^{\perp}$, along $\Re(\mathbf{1}_p)$. The target subspace in this projection ensures that the non-zero-eigenvalue eigenvectors of the projected data's covariance matrix are orthogonal to $\mathbf{1}_p$. On the other hand, Ranatunga's method consists in projecting the rows of $\mathbf{X}$ onto $\Re(\mathbf{S1}_p)^{\perp}$, *along* $\Re(\mathbf{S1}_p)$. From Theorem V.6 it emerges that the direction *along* which this projection is made is the key aspect in ensuring that the columns of the projected data matrix — hence its PCs — will be orthogonal to $\mathbf{X1}_p$. This property will be preserved even if the rows of $\mathbf{X}$ are projected (non-orthogonally) *onto* some other subspace, provided the projection is along the direction of the vector of row sums of $\mathbf{S}$.

We can then merge the two methods and *project the rows of* $\mathbf{X}$ *onto* $\Re(\mathbf{1}_p)^{\perp}$ *along* $\Re(\mathbf{S1}_p)$. This means taking $\mathbf{XQ}_p{}'$, where

$$\mathbf{Q}_p = \mathbf{I}_p - (\mathbf{S1}_p)(\mathbf{1}_p{}'\mathbf{S1}_p)^{-1}\mathbf{1}_p{}' = \mathbf{I}_p - \frac{\mathbf{S1}_p\mathbf{1}_p{}'}{\mathrm{sum}(\mathbf{S})} \tag{V.2.5}$$

Theorem V.6 tells us that this is equivalent to *orthogonally projecting the columns of* $\mathbf{X}$ *onto* $\Re(\mathbf{X1}_p)^{\perp}$. In other words, we have $\mathbf{XQ}_p{}' = \mathbf{P}_{X\mathbf{1}_p}\mathbf{X}$ where

$$\mathbf{P}_{X\mathbf{1}_p} = \mathbf{I}_n - (\mathbf{X1}_p)(\mathbf{1}_p{}'\mathbf{X}'\mathbf{X1}_p)^{-1}(\mathbf{X1}_p)' \tag{V.2.6}$$

Hence, applying Theorem V.4 to the appropriate projected matrices ($\mathbf{Q}_p\mathbf{SQ}_p{}'$ and $\mathbf{P}_{X\mathbf{1}_p}(\frac{1}{n}\mathbf{XX}')\mathbf{P}_{X\mathbf{1}_p}$) we conclude that $\mathbf{XQ}_p{}'$ has $p-1$ Principal Axes in both $\mathbb{R}^n$ and $\mathbb{R}^p$ with non-zero associated eigenvalues. The PAs in $\mathbb{R}^p$ are orthogonal to $\mathbf{1}_p$ and the PCs (PAs in $\mathbb{R}^n$) are orthogonal to $\mathbf{X1}_p$.

This solution is no more involved than either of the previous methods and produces the same number of 'shape factors' $(p-1)$ as before. Appropriately, if the isometry vector in $\mathbb{R}^p$ is already an eigenvector of $\mathbf{S}$ (in which case the isometry factor is a PC of the original data), the three methods coincide. In that case, orthogonality of the remaining PAs in both spaces with the isometry vectors is already guaranteed by conventional PCA and all three methods will merely remove the term in the SVD of the data matrix which corresponds to the isometry vectors in both spaces.

The nature of projector (V.2.5) indicates that the results of the analysis which is now being suggested are, in a sense, intermediate between the results of doubly-centered PCA and of Ranatunga's approach. The essence of all 3 projection-based methods suggested for the removal of isometric size is summarized in Table V.1.

Projection Method	Column Space ($\mathbb{R}^n$)		Row Space ($\mathbb{R}^p$)	
	Onto	Along	Onto	Along
Doubly-centered PCA	$\Re(\mathbf{XS}^{-1}\mathbf{1}_p)^\perp$	$\Re(\mathbf{X1}_p)$	$\Re(\mathbf{1}_p)^\perp$	$\Re(\mathbf{1}_p)$
New method	$\Re(\mathbf{X1}_p)^\perp$	$\Re(\mathbf{X1}_p)$	$\Re(\mathbf{1}_p)^\perp$	$\Re(\mathbf{S1}_p)$
Ranatunga	$\Re(\mathbf{X1}_p)^\perp$	$\Re(\mathbf{XS1}_p)$	$\Re(\mathbf{S1}_p)^\perp$	$\Re(\mathbf{S1}_p)$

Table V.1: Summary of three projection-based methods of removing isometric size from morphometric data

V.2.3 An example

Over the past few years, Prof. Byron Morgan has collected anatomical measurements of final year undergraduate students in Statistics at the University of Kent at Canterbury. One such data set is given in Table V.2. The covariance matrix of the log-transformed data is given in Table V.3.

The eigenvectors of the covariance matrix $\mathbf{S}$ are given in Table V.4. In Table V.5, we can find the eigenvectors of the 'projected covariance matrix' $\mathbf{Q}_p\mathbf{S}\mathbf{Q}_p{}'$, where $\mathbf{Q}_p$ is the non-orthogonal projector (V.2.5). In both Tables, the absolute and relative eigenvalues (as a percentage) are also indicated.

Table V.5 confirms that the isometry vector $\frac{1}{\sqrt{p}}\mathbf{1}_p$ is an eigenvector of $\mathbf{Q}_p\mathbf{S}\mathbf{Q}_p{}'$ with

Student	Chest	Waist	Hand	Head	Height	Forearm	Wrist
1	38	34	9.5	24	72	13	7.5
2	38	30	8	23	70	11	7
3	40	32	8	23	71	10	7
4	32	22	5.5	21	60	10.5	6
5	38	30.5	8	23.25	69	11	6
6	38	36	8	23	72	10.5	6.5
7	36	31	9	23	70	11	7
8	32	26.5	7	22	65	10	6
9	50	42	10	25	70.5	13	8
10	38	32	10	23	71	11	7
11	34	24	7.5	22.5	64	9	6
12	32	26	7	22	63	9.5	6
13	32	24	7	22	62	9	6
14	34	26.5	7	22	64	9.5	6
15	36	32.5	8.3	22.6	70	12.3	6.7
16	38	32	9	23.5	73	11	6.5
17	40	33	9.5	24	75.5	12.3	8
18	35	28	7	22.5	64	9.5	6
19	36	30	9	23.5	70	12.5	6.5
20	38	33	9	23	72.5	11.5	7
21	39	32	8.5	23	73	11	7
22	38	32	7	24	70.5	10	6.5

Table V.2: Anatomical measurements (in inches) of final year undergraduate Statistics students at UKC (1985)

eigenvalue zero. Hence, the remaining eigenvectors of $\mathbf{Q}_p\mathbf{S}\mathbf{Q}_p{}'$ are necessarily orthogonal to the isometry vector in $\mathbb{R}^p$, as was also the case with doubly-centered PCA, but not with Ranatunga's method. This implies that the sum of the coefficients of these eigenvectors is always zero, as can be checked in Table V.5 (rounding-off errors aside).

There is a striking similarity between the first six (absolute) eigenvalues of the

Variables	Chest	Waist	Hand	Head	Height	Forearm	Wrist
Chest	0.973	(0.893)	(0.716)	(0.868)	(0.735)	(0.637)	(0.807)
Waist	1.274	2.093	(0.795)	(0.862)	(0.854)	(0.707)	(0.784)
Hand	1.033	1.681	2.137	(0.787)	(0.827)	(0.719)	(0.790)
Head	0.320	0.466	0.430	0.140	(0.794)	(0.656)	(0.758)
Height	0.445	0.759	0.742	0.182	0.377	(0.681)	(0.776)
Forearm	0.690	1.122	1.153	0.269	0.459	1.203	(0.763)
Wrist	0.730	1.039	1.058	0.259	0.437	0.768	0.840

Table V.3: Covariance matrix ($\times 10^{-2}$) and correlation matrix (in brackets) of the log-transformed anatomical data

	Eigenvectors						
Variables	1	2	3	4	5	6	7
Chest	0.35	-0.47	-0.08	-0.39	-0.58	-0.32	0.26
Waist	0.54	-0.53	-0.01	0.47	0.23	0.38	-0.02
Hand	0.53	0.52	0.62	0.02	-0.22	0.09	0.05
Head	0.13	-0.08	0.03	-0.02	-0.12	-0.26	-0.95
Height	0.22	-0.01	0.10	0.12	0.54	-0.78	0.17
Forearm	0.36	0.46	-0.75	0.24	-0.17	-0.08	0.02
Wrist	0.32	0.07	-0.18	-0.74	0.48	0.26	-0.08
λ_k (%)	0.0634 (81.7%)	0.0056 (7.2%)	0.0044 (5.7%)	0.0022 (2.9%)	0.0012 (1.5%)	0.0006 (0.8%)	0.0002 (0.3%)

Table V.4: Eigenpairs of the Covariance Matrix $\mathbf{S}$ of the log-transformed anatomical data

'projected covariance matrix' $\mathbf{Q}_p\mathbf{S}\mathbf{Q}_p{}'$ and the last six eigenvalues of the original covariance matrix $\mathbf{S}$. The corresponding eigenvectors are also fairly similar. This reflects the fact that the isometry vector $\frac{1}{\sqrt{p}}\mathbf{1}_p$ which is being removed by the projection is close to the first eigenvector of $\mathbf{S}$ (the cosine of the angle between them is 0.93).

A measure of this overall similarity can be obtained by setting the largest singular value of the original (column-centered, log-transformed) data matrix $\mathbf{X}$ to zero — effectively subtracting from $\mathbf{X}$ the first term in its SVD — and then computing the

	Eigenvectors						
Variables	1	2	3	4	5	6	7
Chest	-0.46	-0.04	-0.33	-0.61	-0.20	0.35	0.38
Waist	-0.52	0.06	0.54	0.19	0.50	0.04	0.38
Hand	0.53	0.67	0.07	-0.25	0.20	0.12	0.38
Head	-0.08	0.05	-0.01	-0.12	-0.24	-0.88	0.38
Height	-0.00	0.12	0.13	0.54	-0.67	0.28	0.38
Forearm	0.46	-0.71	0.30	-0.19	0.03	0.10	0.38
Wrist	0.07	-0.15	-0.69	0.44	0.39	-0.01	0.38
λ_k	0.0056	0.0044	0.0022	0.0012	0.0006	0.0002	0.0000
(%)	(39.0%)	(30.9%)	(15.7%)	(8.3%)	(4.4%)	(1.6%)	(0.0%)

Table V.5: Eigenpairs of the covariance matrix $\mathbf{Q}_p\mathbf{S}\mathbf{Q}_p{}'$ of the projected anatomical data

similarity index in $\mathbb{R}^p$ of the resulting matrix with the projected data matrix $\mathbf{X}\mathbf{Q}_p{}'$. This index has an impressive value of 0.9941.

Despite this very high value of the similarity index in $\mathbb{R}^p$, some individual eigenvectors of $\mathbf{S}$ — in particular those with smaller associated eigenvalues — can be somewhat far from orthogonality with the isometric vector. Specifically, the cosines of the angles between $\mathbf{1}_p$ and the last six eigenvectors of $\mathbf{S}$ (which are the sum of these eigenvectors' coefficients times $\frac{1}{\sqrt{p}}$ - and can always be taken to be positive thanks to sign-switching) are 0.02, 0.10, 0.11, 0.06, 0.26 and 0.21.

As far as $\mathbb{R}^n$ is concerned, we know that the first six PCs of $\mathbf{X}\mathbf{Q}_p{}'$ are orthogonal to the isometric size factor $\mathbf{X}\mathbf{1}_p$. How did the original PCs of $\mathbf{X}$ perform in this respect? The last six PCs of $\mathbf{X}$ had the following correlations (cosines of the angles in $\mathbb{R}^n$) with $\mathbf{X}\mathbf{1}_p$: 0.01, 0.03, 0.02, 0.01, 0.03 and 0.01. The first PC had a correlation of 1 to two decimal places (0.999 to three) with $\mathbf{X}\mathbf{1}_p$. Thus, we were already close to having an isometric size factor in $\mathbb{R}^n$, with the original data.

All things considered, the method now proposed appears to have performed well. The stated goals of (exact) orthogonality with the isometric size vectors in both spaces have been achieved with what appears to be as little change to the PCA of $\mathbf{X}$ as was possible.

How does this performance compare with those of the other two methods suggested in Section V.1? The relatively high value of the cosine of the angle between $\mathbf{1}_p$ and $\mathbf{S1}_p$ (0.93) suggests that the differences between them might not be very significant. On the other hand, the fairly low cosine of the angle between vectors $\mathbf{X1}_p$ and $\mathbf{XS}^{-1}\mathbf{1}_p$ – a mere 0.24 – suggests that doubly-centered PCA, which produces PCs that are uncorrelated with $\mathbf{XS}^{-1}\mathbf{1}_p$, rather than with $\mathbf{X1}_p$, might not perform all that similarly to the other two methods. This is indeed the case.

Table V.6 gives the global similarity indices between all pairs of PCAs of $\mathbf{XP}_{\mathbf{1}_p}$, $\mathbf{XQ}_p{}'$, $\mathbf{XP}_{S\mathbf{1}_p}$and of $\mathbf{X}$ with the first term in its SVD removed. We denote the latter matrix as $\mathbf{XP}_{\mathbf{a}_1}$, since it corresponds to orthogonally projecting the rows of $\mathbf{X}$ onto the orthogonal complement of the first eigenvector of $\mathbf{S}$, $\mathbf{a}_1$.

The differences between the PCAs of $\mathbf{XP}_{\mathbf{1}_p}$ and of the other three matrices are most noticeable — for our purposes — in the correlations of $\mathbf{XP}_{\mathbf{1}_p}$'s (non-zero length) PCs with the isometric size factor $\mathbf{X1}_p$. They are: 0.92, 0.05, 0.24, 0.17, 0.05 and 0.12. The correlation of $\mathbf{XP}_{\mathbf{1}_p}$'s first PC with the isometric size factor (0.92) is

s_g	$\mathbf{XP}_{\mathbf{a}_1}$	$\mathbf{XP}_{\mathbf{1}_p}$	$\mathbf{XQ}_p{}'$	$\mathbf{XP}_{S\mathbf{1}_p}$
$\mathbf{XP}_{\mathbf{a}_1}$	1.0000			
$\mathbf{XP}_{\mathbf{1}_p}$	0.7822	1.0000		
$\mathbf{XQ}_p{}'$	0.9947	0.7931	1.0000	
$\mathbf{XP}_{S\mathbf{1}_p}$	0.9998	0.7894	0.9952	1.0000

Table V.6: Global similarity indices for the PCAs of four projections of the anatomical data

surprisingly high. This PC is a shape factor (the coefficients of the first eigenvector of $\mathbf{P}_{\mathbf{1}_p}\mathbf{S}\mathbf{P}_{\mathbf{1}_p}$ are of different signs: -0.06, 0.49, 0.60, -0.53, -0.32, -0.05, -0.13), but a shape factor which is highly correlated with the isometric size factor. And all the remaining shape factors of $\mathbf{XP}_{\mathbf{1}_p}$ are more correlated with $\mathbf{X1}_p$ than their counterparts for the original data matrix $\mathbf{X}$ (the relevant correlations were given above).

Clearly, the performance of doubly-centered PCA in terms of removing isometric size in $\mathbb{R}^n$ is quite poor. But this results from the very nature of $\mathbf{P}_{\mathbf{1}_p}$'s 'companion projector' in $\mathbb{R}^n$: it only seeks orthogonality with $\mathbf{XS}^{-1}\mathbf{1}_p$, not with $\mathbf{X1}_p$, and we have already seen how the angle between these two vectors is large, for this data set.

However, the performance of doubly-centered PCA in $\mathbb{R}^p$ also merits some comments. Table V.7 gives the similarity indices in $\mathbb{R}^p$ for all pairs of the above four analyses and is equally clear in singling out doubly-centered PCA as the odd-man out. In contrast with the other methods doubly-centering appears to be tinkering quite considerably with the PAs in $\mathbb{R}^p$ of $\mathbf{X}$ in order to ensure orthogonality with the isometric size vector $\mathbf{1}_p$.

The relatively minor differences between the Ranatunga method and the method

s_p	$\mathbf{XP}_{\mathbf{a}_1}$	$\mathbf{XP}_{\mathbf{1}_p}$	$\mathbf{XQ}_p{}'$	$\mathbf{XP}_{S\mathbf{1}_p}$
$\mathbf{XP}_{\mathbf{a}_1}$	1.0000			
$\mathbf{XP}_{\mathbf{1}_p}$	0.6944	1.0000		
$\mathbf{XQ}_p{}'$	0.9941	0.7219	1.0000	
$\mathbf{XP}_{S\mathbf{1}_p}$	1.0000	0.6957	0.9947	1.0000

Table V.7: Similarity indices in $\mathbb{R}^p$ for four projections of the anatomical data

which is being proposed agree with Ranatunga's comments on the fact that the PAs in $\mathbb{R}^p$ produced by her method tended to be 'almost' orthogonal to $\mathbf{1}_p$. But the theoretical appeal of the new method, its guarantee of success for *all* morphometric data sets and the fact that simultaneous orthogonality is obtained at no extra cost, all combine to suggest that it is a valid alternative.

V.3 Removing underlying functions in $\mathbb{R}^p$

Another area where projection matrices have been used in connection with PCA involves data sets which are time series or, more generally, observations of a given phenomenon at p different values of some independent (in the regression analysis sense of the word) variable. It might be known, or assumed, that part of the variation of the observed values corresponds to some additive component which can be described by a function. Rather than carrying out a PCA of the observed values, it might be preferred to remove this functional component from the data prior to the analysis. In the words of Winsberg (1988) [68, (pg.130)]:

> The obvious or known variation in the data may often be described (...). Additional subtle variation in the data may be suspected and it is sometimes useful to remove obvious or known variation in order to uncover the unknown or subtle variation (...). To attain the desired goal, a '**filter**' is incorporated (...) which removes the obvious or known variation.

There is an obvious analogy between this problem and the one raised — in a different context — by Somers (1989) [62] and which was addressed in Section V.2. But there are also important differences which we now consider.

V.3.1 Focusing on $\mathbb{R}^p$

The class of $n\times p$ data sets where the p 'variables' are, in effect, different observations of the same quantity at p different points in time, space or some other independent variable, play a special role in the context of PCA.

Each row of an $n\times p$ data matrix of this kind can be viewed as a function, or more precisely, as a set of p observations on a given function. In this context, the most appropriate graphical representation of the data is a two-dimensional graph with the horizontal axis indicating the p values of the independent variable and the vertical axis denoting observed values. Each of the n rows will then appear as a set of p points on the graph which can be joined up if, and in such a way as, it is considered appropriate to produce n curves. For simplicity, we shall often refer to these n rows as 'curves' in what follows, in the knowledge that they are only p observations on such functions which cannot therefore uniquely determine any given curve.

Unlike previous data sets which have been considered, there is not now much practical interest in the space $\mathbb{R}^n$. Principal Components, which are linear combinations of the p data columns, will have little significance, since our attention is now focused on what happens 'horizontally'. By contrast, PAs in $\mathbb{R}^p$ acquire a special significance. They too can be viewed as functions observed at the p values of the independent variable. By definition, the PAs in $\mathbb{R}^p$ will be 'best-fitting' such observed functions, subject to orthogonality constraints. This means that they will successively minimize the sum of squares of distances at all p points between observed curves and themselves. It also means (see the discussion in Chapter I) that they will successively account for the maximum variability of the n curves and thus provide (successively orthogonal) patterns of variability of the n curves.

Thus, reference to a 'PCA' of such data sets is, in reality, reference to an Analysis in $\mathbb{R}^p$, for only PAs in $\mathbb{R}^p$ are really meaningful in this context.

In addition, column-centering should be used with caution. It amounts to replacing each of the n observed 'curves' with its deviation from the mean 'curve', and therefore implicitly removing a 'basic model curve'. Analyzing the curves themselves, rather than the residuals from this 'mean curve' therefore implies skipping the column-centering stage of PCA.

Thus, references to a PCA of such data sets will often mean a *non-centered PCA in* $\mathbb{R}^p$, *i.e.*, a spectral decomposition of the matrix of (non-centered) second order moments for the observations on the functions at each of the p points.

V.3.2 Filters

Besse and Ramsay (1986) [5] tackled the problem of removing certain classes of functions from the observed data. Their method – which is quite intricate and laborious – is based on proving a very elegant equivalence between a PCA of a certain transformation of the data matrix and the vectors which would be obtained if the following steps were taken: (i) replacing the n observed rows of the data matrix with interpolating splines ; (ii) annihilating ('filtering out') additive components involving the functions which are to be removed from the data ; (iii) carrying out a Functional PCA (see Ramsay (1982) [51]) of the residual functions. The results of both approaches are equivalent in the sense that it is possible to define an inner-product-preserving isomorphism between the space of interpolants and the space of rows ($\mathbb{R}^p$), with special inner products in each space defined by the method itself.

Winsberg and Kruskal (1986) [69] noted the considerable technical difficulties with Besse & Ramsay's approach, along with some disadvantages (see also Winsberg (1988) [68, (pgs. 132-133)]) and looked for simpler alternative methods of removing the variability which can be accounted for by known underlying functions. One such alternative consists in *orthogonally projecting the rows of the data matrix onto the orthogonal complement of the subspace spanned in* $\mathbb{R}^p$ *by the vectors of* p *observations of the functions which are to be removed.* Specifically, say we wish to filter out s functions $\{f_k(t)\}_{k=1}^s$ from the data. Let $\{t_j\}_{j=1}^p$ denote the p points at which observations are being made. We then construct the $p \times s$ matrix $\mathbf{C}$ whose k-th column has the p observations of $f_k(t)$:

$$\mathbf{C} = \begin{pmatrix} f_1(t_1) & \cdots & f_k(t_1) & \cdots & f_s(t_1) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ f_1(t_j) & \cdots & f_k(t_j) & \cdots & f_s(t_j) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ f_1(t_p) & \cdots & f_k(t_p) & \cdots & f_s(t_p) \end{pmatrix}$$

Using matrix $\mathbf{C}$, we define the orthogonal projector onto $\Re(\mathbf{C})^\perp$, *i.e.*,

$$\mathbf{P}_C = \mathbf{I}_p - \mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}' \qquad \text{(V.3.1)}$$

It is assumed that the s columns of $\mathbf{C}$ are linearly independent, but if that turned out not to be the case, the inverse of $\mathbf{C}'\mathbf{C}$ in (V.3.1) can be replaced with its Moore-Penrose generalized inverse (see also Rao (1980) [56, (pg.12)]), without affecting the nature and role of $\mathbf{P}_C$.

Winsberg and Kruskal's method then amounts to determining the right singular vectors ('non-centered PCA PAs in $\mathbb{R}^p$') of the projected (residual) data matrix $\mathbf{YP}_C$. The crucial idea here is no different from what has been discussed so far in this

Chapter, and can be viewed as an application of the External Analysis decomposition (V.1.1). The projector $\mathbf{P}_C$ is an example of a filter in the terminology of Winsberg (1988) [68]. It will filter out any component in each row of $\mathbf{Y}$ which is in $\Re(\mathbf{C})$, *i.e.*, which is a linear combination of $\mathbf{C}$'s columns, in such a way that the remaining residual is orthogonal to $\Re(\mathbf{C})$. Naturally, the s columns of $\mathbf{C}$ can be replaced with any other s linearly independent vectors in $\Re(\mathbf{C})$.

In effect, this idea amounts to replacing each row of the data matrix $\mathbf{Y}$ with its residual after being linearly regressed on the columns of $\mathbf{C}$. Using the standard notation of linear regression, $\mathbf{YP}_C$ can be written as $\mathbf{Y}-\hat{\mathbf{Y}}\mathbf{C}'$, with $\hat{\mathbf{Y}} = \mathbf{YC}(\mathbf{C}'\mathbf{C})^{-1}$. Each row of $\hat{\mathbf{Y}}$ can therefore be written (in column vector form) as $(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\mathbf{y}_i{}^{row}$. This is known to be (see for example Basilevski (1983) [4, (pg.145)]) the vector of regression coefficients of $\mathbf{y}_i{}^{row}$ on the s columns of $\mathbf{C}$.

Winsberg (1988) [68] gives an example of the method at work. In the next Subsection we shall consider another example of the method.

V.3.3 The 100 km race data

Eighty runners completed the 1984 Lincoln 100 km race. In this Subsection we consider the 80x10 data set giving the time each runner took to cover each 10km stretch in the race. This data set is also analyzed in Jolliffe (1986) [32, (pgs.82-85;211)]. Due to its size, it is reproduced separately, in Appendix B.

A model for the times in long-distance races will now be considered. Removing the 'model race times' from the observed data — in the spirit of Winsberg and Kruskal's approach — and then carrying out a PCA of the residual times will enable us to look for the main sources of variability in the deviations from the model race times.

The Hindley model

Hindley (1984) [24] studied the times of winners in long distance races and suggested a basic model for their performance. The model assumes that the best runners *maintain a constant ratio of present speed to average speed so far in the race.* Denoting distance by x and time by t , this assumption is that

$$\frac{dx}{dt} = s\frac{x}{t} \qquad \text{(V.3.2)}$$

where s is the constant which Hindley calls the **slow-down index** for a runner.

The assumption implies that the time t taken to cover a given distance x is:

$$t = t_*\left(\frac{x}{x_*}\right)^{\frac{1}{s}} \qquad \text{(V.3.3)}$$

where (x_*, t_*) is an 'initial' condition, that is, a point on the curve (which, however, cannot be the time and distance at the beginning of the race!).

Before the 1984 race, Hindley — one of the race organizers — proposed the model to the participants, suggesting it as a strategy for good running. We can now try to analyze the deviations from the model in the actual race times and enquire as to the main sources of variability in these deviations.

In order to apply the Winsberg and Kruskal filters, that is, replace the data with residuals from regression models, we may follow two alternative strategies.

One option is to consider that all runners have a given common slow-down index s, but that the different abilities of each runner will be reflected in different 'reference times' t_* at some chosen distance x_*. This implies that we take the matrix $\mathbf{C}$ to have a single column ($\mathbf{c}$), whose j-th coefficient, corresponding to the stretch from x_{j-1} to x_j is:

$$c_j = x_j^{\frac{1}{s}} - x_{j-1}^{\frac{1}{s}} \qquad (x_0 = 0) \qquad \text{(V.3.4)}$$

(since the times in the data set are for each 10km stretch and not cumulative times). The projection (regression and removal of fit) carried out by $\mathbf{P}_C$ will then fit a value of $\frac{t_*}{x_*^{\frac{1}{s}}}$ for each runner — the value which ensures the least squares fit for the model (V.3.3) to that particular row of observations. The vector of these fitted values is $\mathbf{YC}(\mathbf{C}'\mathbf{C})^{-1} = \mathbf{Y}\frac{\mathbf{c}}{||\mathbf{c}||^2}$. A PCA of $\mathbf{YP}_C$ would, in this case, study the variability of the deviations from model (V.3.3) for a fixed value of s .

A second option can be envisaged if it is considered that the assumption of a common slow-down index is unrealistic. In that case, the problem becomes intrinsically non-linear. We can still fit different values of s for each runner, using the Winsberg and Kruskal approach, but at the price of moving to logarithmic coordinates. In fact, equation (V.3.3) is a direct analogue of the allometry equation (V.2.1). A logarithmic transformation would linearize the equation to:

$$\log t = \log \frac{t_*}{x_*^{\frac{1}{s}}} + \frac{1}{s} \log x \qquad \text{(V.3.5)}$$

hence making the scalar $\frac{1}{s}$ the regression coefficient of $\log t$ on $\log x$. By working with cumulative times after each 10km stretch — rather than the times for each stretch — and converting to log-distances and log-times, we can provide estimates for both the constant term $\log \frac{t_*}{x_*^{\frac{1}{s}}}$ and the coefficient $\frac{1}{s}$ *for each runner*. This implies taking the matrix $\mathbf{Z}$ of the logarithmic cumulative times and a projection matrix $\mathbf{P}_C$ where $\mathbf{C}$ has two columns: a column of ones $(\mathbf{1}_p)$ for the constant term and a column of logarithmic cumulative distances, whose j-th entry is $\log \mathbf{x}_j$.

The disadvantages of this second option are clear. It is unpleasant to work with log-times and log-distances, in particular when they are to be submitted to a PCA, thus making interpretation of results particularly intractable. It is true that we can always transform back by exponentiation, but this does not solve all problems. Exponentiating the PAs in $\mathbb{R}^p$ of the log-transformed data would give us non-orthogonal 'curves'. Exponentiating the data after the fit, but prior to the PCA, would give us orthogonal axes. However, the removed fits would not be least squares fits, nor would they be orthogonal projections.

All things considered, we shall proceed with the first option and make some considerations concerning the second alternative at the end.

Fitting the Hindley model

The choice of a common slow-down index s is aided by Hindley (1984) [24]. Having studied the performance of top runners, he recommends $s = 0.97$ as a good value for the index. A value of $s \simeq 1$ implies a nearly linear relation between distance and time. Smaller values of s will imply 'more than linear' growth of time *vis-a-vis* distance, that is, a slowing down of the pace of the race.

The spectral decomposition of the matrix of (non-centered) second order moments of $\mathbf{YP}_C$, where $\mathbf{Y}$ is the data matrix in Appendix B and $\mathbf{P}_C$ is defined by (V.3.1) and (V.3.4) with $s = 0.97$, reveals that three dimensions are needed to account for 90% of the variability of the residual curves, and four if we wish to reach 95%. The eigenpairs of this matrix are given in Table V.8.

Figure V.1 gives the projections of the scatter of $n = 80$ points in $\mathbb{R}^p$ onto the first two such eigenvectors, enabling us to have an idea of how the 80 runners perform in terms of the two main sources of variation. The number for each runner reflects their ranking on arrival.

It is in the nature of the method that the last (zero-eigenvalue) eigenvector is the 'unit-norm model race', that is, the vector of model times for each 10km stretch (with $s = 0.97$), scaled to have sum of squares equal to one. It can be seen to be a fairly constant set of values, with a gradual growth in successive stretches. The 'model races' which are filtered out for each runner are scalar multiples of this track record.

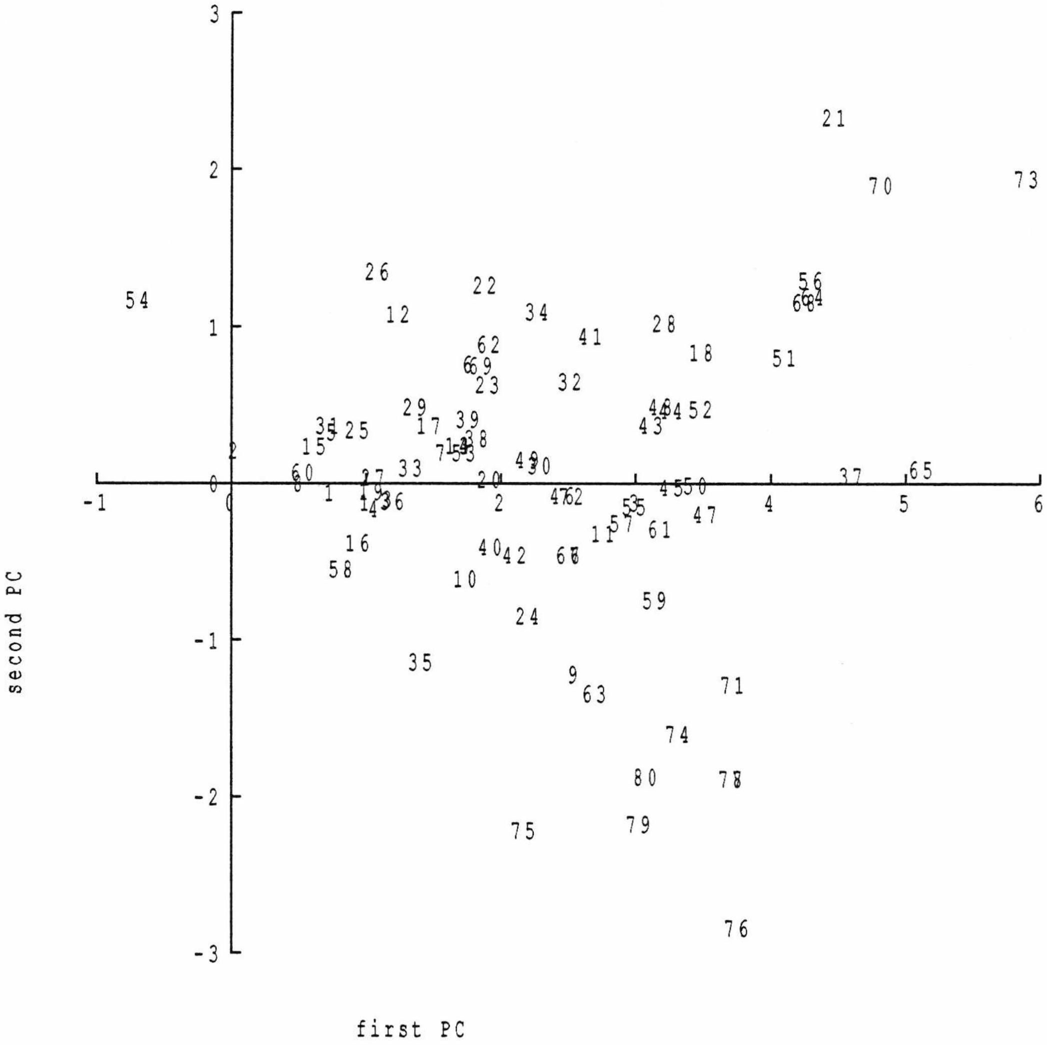

Figure V.1: Least-squares two-dimensional representation of the scatter of 80 residual time series for the 100km race data

	Eigenvectors									
Km	1	2	3	4	5	6	7	8	9	10
10	-0.34	0.22	0.12	0.20	0.16	0.41	0.29	-0.09	0.64	0.29
20	-0.32	0.21	0.12	0.24	0.07	0.34	0.06	0.19	-0.72	0.31
30	-0.39	0.08	0.06	0.08	-0.20	-0.34	-0.48	-0.59	-0.01	0.31
40	-0.26	-0.07	0.06	-0.05	-0.11	-0.39	-0.20	0.76	0.21	0.32
50	-0.18	-0.33	-0.02	-0.39	-0.12	-0.26	0.68	-0.18	-0.14	0.32
60	0.07	-0.42	-0.23	-0.42	-0.12	0.58	-0.38	0.01	0.04	0.32
70	0.15	-0.07	-0.58	0.20	0.67	-0.21	-0.03	-0.04	-0.03	0.32
80	0.43	-0.39	0.17	0.64	-0.33	-0.00	0.09	-0.03	0.03	0.32
90	0.41	0.14	0.66	-0.31	0.40	-0.07	-0.11	-0.06	-0.01	0.32
100	0.38	0.65	-0.34	-0.16	-0.42	-0.02	0.10	0.02	0.02	0.33
λ_k	540.95	72.91	38.74	33.15	19.50	8.06	5.63	3.05	1.45	0.00
π_k	0.748	0.101	0.054	0.046	0.027	0.011	0.008	0.004	0.002	0.000
$\sum \pi_i$	0.748	0.849	0.902	0.948	0.975	0.986	0.994	0.998	1.000	1.000

Table V.8: Eigenvectors and (cumulative (relative (eigenvalues))) of the matrix of second order moments for the projected 100km race data

The first eigenvector, which accounts for some 75% of the total variability of the residual races, reflects the fact that most runners got off to a faster start than the model strategy would recommend. This is reflected in the almost universally negative signs of the first few columns of matrix $\mathbf{YP}_C$, as opposed to the almost universally positive signs in the last few columns. The runners on one end of this eigenvector (the left-most points in Figure V.1) are the exception to this rule: runner 54 (who scored positive residuals in the first 40kms and negative residuals at 60, 70 and 100km); runner 2 (positive at 10, 20 and 40km, negative at 70 and 90km); and to a lesser extent, runners 8, 15, 60, 31 and 58 (all of whom scored at least one positive residual in the first 40kms). It should be noted that the removal of the model race has implied that the order of arrival of each runner (*i.e.*, total time taken to run the 100kms) is no longer a significant factor of variation, as was the case with a conventional PCA of $\mathbf{Y}$ (see Jolliffe (1986) [32, (pg.85)]).

The second eigenvector essentially contrasts the time for the last stretch with the times for the fifth, sixth and eighth readings, but with a lesser role also being played by the first two 10km stretches. The significance of this eigenvector (which accounts for about 10% of the total variability) is less clear. Extreme values on this eigenvector go to runners who started off significantly faster than their model target and then oscillated considerably around the target values in the latter half of the race. On one end of the eigenvector (the top of Figure V.1) we find those among this category of runners who ran a 'boom and bust' race, with fast times through most of the middle part of the race and a significantly slower time in the last stretch. Prominent examples of this track record are runners 21, 70 and 73. On the other end (the bottom of Figure V.1) we find the 'second wind' runners who, having slowed down earlier on in the race than the first group, managed to actually speed up in the last ten kilometers. The two most extreme cases of this behaviour (runners 75 and 76) were the runners whose residual time in the last 10km was the fastest.

The next three eigenvectors provide different contrasts between patterns of behaviour in the race, and in particular in the second half of the race. This is not surprising since most of the variation in the data is centered on the latter half of the race (see Appendix B).

It may be noted that a runner who ran a 'model race' (with all residual times zero) would be at the origin of $\mathbb{R}^p$, hence of any graph like that of Figure V.1. Computing the distance from the origin for each runner (the ℓ_2-norm of each row of $\mathbf{YP}_C$) we see that the 10 runners who best kept to the model race (with $s = 0.97$) were, in this order, 2, 8, 15, 1, 60, 5, 31, 19, 27 and 16 (this information is not always faithfully reflected in Figure V.1, since all 10 dimensions are used in computing the distances, and only two in the graph). Thus, 'model racing' and order of arrival do not appear to be very strongly related.

Before moving on, a few words about the PCA of $\mathbf{Y}$, the matrix of observed times (without any filtering). The eigenvectors and corresponding eigenvalues of the matrix of (non-centered) second order moments of $\mathbf{Y}$'s columns are given in Table V.9.

Comparing Tables V.8 and V.9, it can be seen that the last eight eigenvalues in Table V.9 are, broadly speaking, similar to the last eight non-zero eigenvalues (*i.e.*, the second to ninth eigenvalues) in Table V.8. The discussion following Theorem V.5 tells us that it is possible that the two sets of corresponding eigenvectors are also similar. A preliminary inspection of these eigenvectors reveals some broadly similar traits but also, with one or two exceptions, noticeable differences of detail.

A sufficient condition for the eigenvectors of the matrix $\mathbf{V}$ of second order moments of the original data to also be eigenvectors of matrix $\mathbf{P}_C\mathbf{V}\mathbf{P}_C$ is (Theorem V.4,

	Eigenvectors									
Km	1	2	3	4	5	6	7	8	9	10
10	0.25	0.33	0.33	0.01	-0.10	0.15	-0.43	0.24	-0.03	0.66
20	0.27	0.37	0.34	0.03	-0.11	0.03	-0.32	0.10	-0.01	-0.74
30	0.27	0.36	0.17	0.01	-0.06	-0.20	0.36	-0.52	0.56	0.12
40	0.28	0.32	0.02	-0.04	0.11	-0.18	0.45	-0.06	-0.75	0.06
50	0.29	0.17	-0.33	-0.14	0.32	-0.14	0.22	0.69	0.34	-0.03
60	0.32	-0.00	-0.47	0.04	0.43	-0.22	-0.53	-0.39	-0.07	0.02
70	0.34	-0.10	-0.11	0.62	0.11	0.65	0.19	-0.04	0.02	-0.03
80	0.37	-0.15	-0.41	0.02	-0.80	-0.16	-0.03	0.05	-0.03	0.01
90	0.37	-0.30	0.10	-0.74	0.06	0.44	0.05	-0.14	-0.00	-0.04
100	0.36	-0.61	0.48	0.22	0.13	-0.43	0.02	0.11	0.01	0.03
λ_k	35167.35	106.40	70.08	37.26	29.15	18.45	7.94	5.43	2.38	1.32
π_k	0.992	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
$\sum \pi_i$	0.992	0.995	0.997	0.998	0.999	1.000	1.000	1.000	1.000	1.000

Table V.9: Eigenvectors and (cumulative (relative (eigenvalues))) of the matrix of second order moments for the original 100km race data

point iv)) that they belong to $\Re(\mathbf{P}_C)$, *i.e.*, that they be orthogonal to the 'model race' vector $\mathbf{c}$ (see equation (V.3.4)). It is instructive to look at the cosines of the angles between the last eight eigenvectors of $\mathbf{V}$ and the vector $\mathbf{c}$, as well as at the cosines between the j-th eigenvector of $\mathbf{V}$ ($j = 3, ..., 10$) and the $(j-1)$-th eigenvector of $\mathbf{P}_C\mathbf{V}\mathbf{P}_C$, $\mathbf{q}_{j-1}$, whose eigenvalues are, as we saw above, fairly similar. These values are given in Table V.10.

	j-th eigenvector of $\mathbf{V}$							
	3	4	5	6	7	8	9	10
$\mathbf{c}$	0.029	0.012	0.029	0.016	0.005	0.006	0.011	0.007
$\mathbf{q}_{j-i}$	0.957	0.919	0.892	0.975	0.995	0.986	0.958	0.972

Table V.10: Cosines of angles between the last 8 eigenvectors of the matrix $\mathbf{V}$ of second order moments of the race data and (a) the 'model race' vector $\mathbf{c}$; (b) the eigenvectors of the 'projected covariance matrix' $\mathbf{P}_C\mathbf{V}\mathbf{P}_C$ whose eigenvalues are most similar

Although the result of Theorem V.4 (point iv)) is only valid if the eigenvectors of $\mathbf{V}$ are *exactly* orthogonal to the vector $\mathbf{c}$, we would expect some variational robustness. Table V.10 suggests that although, roughly speaking this might be true, it is best to be cautious in this respect. The fifth eigenvector of $\mathbf{V}$ is nearly on $\Re(\mathbf{P}_C)$ since it forms an angle of approximately $88°22'$ with the model vector $\mathbf{c}$. However, the eigenvector of $\mathbf{P}_C\mathbf{V}\mathbf{P}_C$ whose eigenvalue is most similar (the fourth) forms an angle of about $26°50'$ with it. Although this particular pair of vectors has among the least similar corresponding eigenvalues (29.15 and 33.15, respectively), other examples of analogous behaviour can be found. For example, the tenth eigenvector of $\mathbf{V}$, which forms an even smaller angle with $\Re(\mathbf{P}_C)$ (about $0°23'$) and the corresponding (ninth) eigenvector of $\mathbf{P}_C\mathbf{V}\mathbf{P}_C$, which forms an angle of about $13°40'$ with that subspace. Their associated eigenvalues are very similar (1.32 and 1.45). Thus, even minor deviations of an eigenvector of $\mathbf{V}$ from $\Re(\mathbf{P}_C)$ may imply noticeable changes in a 'corresponding' eigenvector of $\mathbf{PVP}$.

The difference in the traces of the matrices $\mathbf{V}$ and $\mathbf{P}_C\mathbf{V}\mathbf{P}_C$ is approximately 34 723. From Theorem V.4 (point vii)) we know that this is the value of $\mathbf{V}$'s Rayleigh-Ritz ratio determined by the 'model race' vector $\mathbf{c}$. The fact that it is also close to the largest eigenvalue of $\mathbf{V}$ suggests that $\mathbf{c}$ is probably a good approximation to $\mathbf{V}$'s largest eigenvector, $\mathbf{a}_1$. The cosine of the angle between them is 0.9936, which corresponds to an angle of approximately $6°30'$. Thus, *the first PA in* $\mathbb{R}^p$ *of*

$\mathbf{Y}$ *is a good approximation of the model vector* $\mathbf{c}$. This is, in itself, an interesting validation of the model (and the parameter $s = 0.97$) in that it states that the model is approximately the main factor of variability in the data.

What has just been said also suggests a simple way of estimating a value for the common slow-down index s from the data itself, rather than using a given value, such as Hindley's suggested value ($s = 0.97$). We can compute $\mathbf{V}$'s first eigenvector $\mathbf{a}_1$ — which is the vector of $\mathbb{R}^p$ which best fits the data in the least squares sense — and then determine the value of s which gives rise to the 'model race' vector $\mathbf{c}$ (with j-th element $c_j = k\left((10j)^{1/s} - [10(j-1)]^{1/s}\right)$, for some constant k and $j = 1, ..., 10$) which best approximates $\mathbf{a}_1$. Choosing a value of s along these lines by the least-squares criterion gives rise to a problem in non-linear regression. But by analogy with the discussion that led to equation (V.3.5), we may make this a problem in linear regression by a logarithmic transformation. Regressing the logarithms of the partial sums of $\mathbf{a}_1$'s coefficients on the logarithms of the cumulative distances at which observations are made, we have a fitted value for $1/s$. Alternatively, we can reverse the roles of log-cumulative times and log-cumulative distances to obtain a direct estimate for s . Neither of these fitted values is a least-squares fit in the original coordinate system, but they do have the virtue of using information provided by the data.

Applying these two latter strategies to our example leads to fitted values of 1.0910 for $1/s$ (hence 0.9166 for s) and 0.9154 for s . A PCA with a value of $s = 0.916$

(the mean of the above values for s) will not be very different from that described above, using $s = 0.97$. Their similarity index in $\mathbb{R}^p$ is 0.92. Nonetheless, there is a slight improvement in the approximation between the new 'model race' vector (with $s = 0.916$) and $\mathbf{V}$'s first eigenvector $\mathbf{a}_1$. The angle between them is down to about $4°25'$ (the cosine of this angle is 0.9970).

On the other hand, a non-linear regression of $\mathbf{a}_1$ on $\mathbf{c}$, as calculated by the default option on GENSTAT's 'FITNONLINEAR' command, produces a fitted value of $s = 0.8712$. This somewhat lower value gives rise to a fitted model vector $\mathbf{c}$ whose angle with $\mathbf{a}_1$ is down to approximately $3°40'$ (a cosine of 0.9980). The differences in a PCA using this value for s and those using the previous values are now becoming considerable. The similarity index in $\mathbb{R}^p$ for the PCAs using $s = 0.916$ and $s = 0.8712$ is 0.87, whilst for the PCAs using $s = 0.97$ and $s = 0.8712$ it is as low as 0.68.

Since we are essentially interested in illustrating the potential of projective transformations, rather than in the specific data set which is now being considered, a more detailed analysis of the PCAs with these estimated values of s will be omitted.

Finally, a few words about the second option for removing the Hindley model, discussed at the beginning of this Subsection.

We have already noted how the advantage of relaxing the constraint of a common slow-down index is obtained at the expense of moving to log-transformed data. Another consequence will be that a second dimension is forcefully removed from the data (matrix **C** has two columns in this case). The major advantage lies in the possibility of obtaining different slow-down indices for each runner, which in turn should provide for the possibility of removing more of the variability in the data by filtering out the model.

Rather than produce a lengthy presentation of the results with this option, we shall merely illustrate these two advantages. The trace of the matrix of second order moments of the log cumulative times is 24 321.62. The trace of its counterpart for the projected data is a mere 0.85. Almost all the variability in the log cumulative times has thus been removed by filtering out the fitted values of the linear regression on the log cumulative distances. The regression coefficients for the eighty rows are fits of $1/s$. Taking reciprocals, we have fitted values for each runner's slow-down index. These values for each s are given in Table V.11. There is a general trend for smaller indices as the order at arrival grows, which is not surprising. Also, the common value of $s = 0.916$ which resulted from trying to fit **V**'s first eigenvector to the model (with a similar linear regression of log-transformed cumulatives) seems to be an appropriate 'average value' of the indices in Table V.11, whose mean is 0.924. Runner fifty-four's value (1.019) is greater than one, implying that he actually has a 'speed-up' index. The exceptional nature of this runner's track record has already been noted. It should be added that the first half of his race was covered in more time (318 minutes) than the second half (305 minutes).

Runner	s	Runner	s	Runner	s	Runner	s
1	0.958	21	0.911	41	0.920	61	0.896
2	0.983	22	0.954	42	0.920	62	0.938
3	0.942	23	0.932	43	0.901	63	0.865
4	0.934	24	0.903	44	0.898	64	0.908
5	0.968	25	0.976	45	0.919	65	0.880
6	0.966	26	0.979	46	0.914	66	0.928
7	0.960	27	0.946	47	0.897	67	0.928
8	0.958	28	0.951	48	0.902	68	0.859
9	0.900	29	0.961	49	0.937	69	0.925
10	0.933	30	0.928	50	0.901	70	0.875
11	0.890	31	0.959	51	0.886	71	0.881
12	0.983	32	0.935	52	0.906	72	0.925
13	0.956	33	0.943	53	0.937	73	0.868
14	0.956	34	0.930	54	1.019	74	0.899
15	0.976	35	0.921	55	0.897	75	0.899
16	0.951	36	0.928	56	0.908	76	0.834
17	0.958	37	0.894	57	0.917	77	0.829
18	0.950	38	0.962	58	0.967	78	0.829
19	0.949	39	0.942	59	0.893	79	0.863
20	0.935	40	0.935	60	0.969	80	0.855

Table V.11: Fitted values of the slow-down index s for each of the 80 runners in the 100km race

There is an interesting coherence between these fitted values of s and the positions of each runner on the first principal axis which resulted from the first option (*i.e.*, from filtering out a model race with a constant s-value of 0.97 for all runners) and which are depicted in Figure V.1. In fact, the correlation between these two sets of 80 values is fairly high in magnitude (0.80 – the sign is irrelevant due to sign-switching). A graph of these 80 pairs of values is given in Figure V.2.

Thus, PCAs of projected data can contribute to an exploratory analysis and validation of a model for certain data sets.

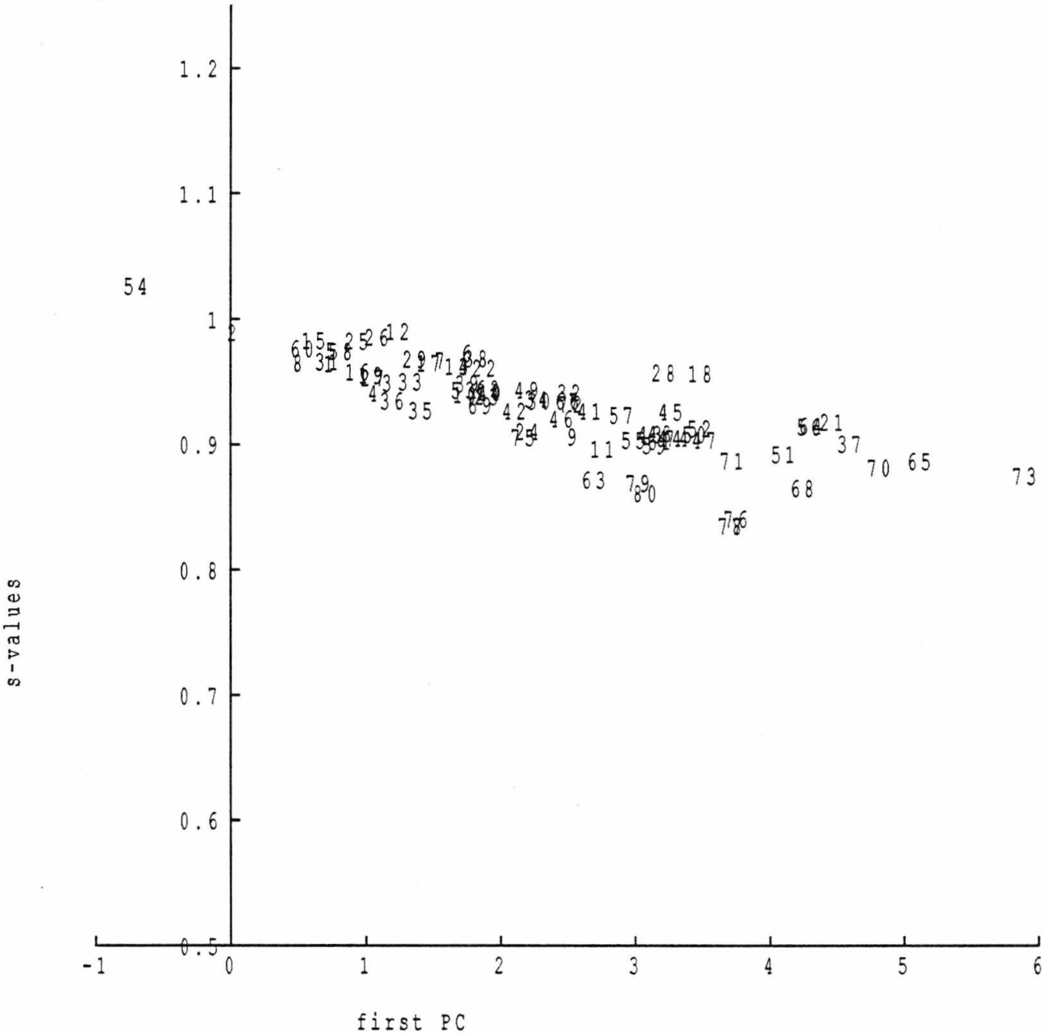

Figure V.2: Scatter of the 80 runners' coordinates on the first principal axis of the residual data after removal of a model race with slow-down index $s = 0.97$ vs. their estimated values of the slow-down index

Chapter VI

A Scale-invariant Component Analysis

The scale dependence of Principal Components was discussed in Section I.5. As was pointed out at the time, it is probably the most serious drawback with PCA, posing fundamental problems in applying the method to data sets where the variables are measured in different units or when different re-scalings of each variable are conceivable.

In this Chapter an alternative 'Component Analysis', which is insensitive to re-scalings of each variable, will be considered. It elaborates on ideas of Hotelling (1933) [27] and Meredith and Millsap (1985) [43]. The p original variables will be replaced by p new variables (*i.e.*, components) which are linear combinations of the original variables, uncorrelated among themselves and which successively optimize a given criterion that is unaffected by changes of scale.

The components thus obtained are not new variables: they turn out to be the Correlation Matrix PCs of the data set. But whereas these components are not *Principal* Components of any scaling of the variables other than standardization (I.1.3), they are the optimum components for *all* scalings under the new criterion. Hence, they represent a scale-invariant solution to a new problem.

VI.1 The method

VI.1.1 A property of Correlation Matrix PCs

The key idea behind the new method is summarized in a passing comment in Hotelling's original 1933 paper [27] which, unfortunately, does not seem to have got the attention it deserves. At the end of his introductory section to the article, Hotelling writes ([27, (pg.422)]):

> An easily verified property of the method [Correlation Matrix PCA] is that the first of our principal components has a greater mean square correlation with the tests [variables] than does any other variable; and that among all variables uncorrelated with the first $q-1$ components ($q = 2, 3, ..., p$), that having the greatest mean square correlation with the

> tests [variables] is the q-th principal component. The argument is similar to that of the next section [derivation of the Correlation Matrix PCs], and will not be given explicitly.

Let us take a closer look at what Hotelling is saying. Say $\mathbf{X}$ is an $n \times p$ column-centered data matrix. All potential 'components' (linear combinations of the columns of $\mathbf{X}$) can be written as $\mathbf{Xc}$ for some vector of coefficients $\mathbf{c} \in \mathbb{R}^p$. The correlations (cosines of the angles in $\mathbb{R}^n$) between $\mathbf{Xc}$ and the p original variables are the coefficients of the vector

$$\mathbf{r} = \left(\tfrac{1}{\sqrt{n}} \mathbf{D}_S^{-\frac{1}{2}} \mathbf{X}' \right) \left(\frac{\mathbf{Xc}}{\sqrt{\mathbf{c}'\mathbf{X}'\mathbf{Xc}}} \right) \qquad \text{(VI.1.1)}$$

where $\mathbf{D}_S^{-\frac{1}{2}}$ is the diagonal matrix of reciprocal standard deviations.

The sum of squares of such correlations is the norm squared of $\mathbf{r}$, *i.e.*, :

$$||\mathbf{r}||^2 = \frac{(\mathbf{Xc})' \left[\frac{1}{n} (\mathbf{X}\mathbf{D}_S^{-\frac{1}{2}})(\mathbf{D}_S^{-\frac{1}{2}}\mathbf{X}') \right] (\mathbf{Xc})}{(\mathbf{Xc})'(\mathbf{Xc})} \qquad \text{(VI.1.2)}$$

Our problem is finding the vectors $\mathbf{Xc} \in \Re(\mathbf{X})$ which successively maximize $||\mathbf{r}||^2$, subject to orthogonality requirements. This is a constrained maximization problem, since we require that the solutions belong to the p-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}$.

Let us first consider the unconstrained problem of successively maximizing the Rayleigh-Ritz ratio of matrix $\mathbf{BB}' = \frac{1}{n}(\mathbf{X}\mathbf{D}_S^{-\frac{1}{2}})(\mathbf{D}_S^{-\frac{1}{2}}\mathbf{X}')$, *i.e.*, of maximizing

$$\frac{\mathbf{y}'\mathbf{BB}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}$$

for $\mathbf{y} \in \mathbb{R}^n$. This problem is well-known to be solved by the eigenpairs of $\mathbf{BB}'$. From Section I.4 (equation (I.4.17)) we know that the non-zero eigenvalues of this matrix – assuming $\mathbf{X}$ is of rank p – are the eigenvalues of $\mathbf{B}'\mathbf{B} = \mathbf{R}$, the correlation matrix for the data. Furthermore, the corresponding eigenvectors of $\mathbf{BB}'$ can be written (without the unit-norm requirement and with $\mathbf{R}$'s j-th eigenvector denoted by $\mathbf{b}_j$) as:

$$\mathbf{y}_j = \mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}\mathbf{b}_j \tag{VI.1.3}$$

Thus, two important conclusions follow:

i). the non-zero-eigenvalue solutions to the unconstrained problem are already vectors in $\Re(\mathbf{X})$ and are therefore also solutions to the original constrained problem of successively maximizing (VI.1.2) (which could also be solved by applying point v) of Theorem V.4, with $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{V} = \mathbf{BB}'$);

ii). *the solutions (VI.1.3) are* none other than *the Correlation Matrix PCs of* $\mathbf{X}$. *The eigenvalues of the correlation matrix are* the values of $||\mathbf{r}||^2$, *i.e., the sums of squares of the correlations (SSCs) between each Correlation Matrix PC and the* p *original variables.*

Hotelling's conclusions are thus proved in our descriptive PCA setting. Although requiring a different proof when working with random variables, the basic result remains valid in an inferential setting (as was considered by Hotelling in [27]).

It should be stressed that an analogous reasoning to that used above will provide an *optimality property of Covariance Matrix PCs* which does not appear to be as well known as those discussed in Chapter I (Subsection I.4.2). This is proved in the following theorem.

Theorem VI.1 *Let* $\{\mathrm{x}_i\}_{i=1}^{p}$ *be* p *variables. The* p *linear combinations of these variables (with unit sum of squared coefficients) which successively maximize the sum of squared covariances with all* p *variables, subject to the constraint of orthogonality among themselves, are the (Covariance Matrix) Principal Components of the variables. The square of the j-th eigenvalue of the covariance matrix determined by* $\{\mathrm{x}_i\}_{i=1}^{p}$ *is the sum of squared covariances of the j-th PC with all* p *variables.*

Proof. Let $\mathbf{X}$ be the matrix whose columns are the variables $\{\mathrm{x}_i\}_{i=1}^{p}$. Let $\mathbf{Xc}$ be any linear combination of those variables. The covariances between $\mathbf{Xc}$ and the columns of $\mathbf{X}$ are given by the (p-dimensional) vector $\frac{1}{n}\mathbf{X}'\mathbf{Xc} = \mathbf{Sc}$, where $\mathbf{S}$ is the covariance matrix of the set $\{\mathrm{x}_i\}_{i=1}^{p}$. Maximizing the sum of squares of these covariances implies maximizing $\mathbf{c}'\mathbf{S}^2\mathbf{c}$. With the unit-norm constraint on $\mathbf{c}$, we have that the optimal $\mathbf{c}$'s are the unit-norm eigenvectors $\{\mathbf{a}_j\}_{j=1}^{p}$ of $\mathbf{S}^2$, hence of $\mathbf{S}$. Thus, the vectors $\mathbf{Xa}_j$ are the (Covariance Matrix) PCs of $\mathbf{X}$. The sum of squared covariances of the j-th PC with all p variables is $\mathbf{a}_j{}'\mathbf{S}^2\mathbf{a}_j = \mu_j^2$, where μ_j is the

j-th eigenvalue of $\mathbf{S}$. □

The consequences of Hotelling's passing comment characterizing Correlation Matrix PCs are far-reaching.

A first important consequence is that it provides a justification for the use of Correlation Matrix PCA which is not usually mentioned in the literature. As was recalled in Chapter I (Subsection I.5.3), Gower (1966) [19] raised the question of why, in our quest for dimensionless data, the particular normalization of dividing by the standard deviation should be preferred to other alternative options which also ensured that goal. This is not a trivial question, since alternative normalizations will produce different PCs. The arguments in favour of standardization (I.1.3) — hence of Correlation Matrix PCA — are essentially the familiarity of correlations and – a not necessarily desirable feature – the fact that it gave equal weight to all p variables in the analysis. But another overriding justification is now obvious. The PCs which result from the standardization (I.1.3) are components which, besides being optimal in the usual sense for Principal Components, are also optimal in the sense of maximizing the sum of squared correlations with all p variables. In particular, the first Correlation Matrix PC of a set of variables $\{\mathbf{x}_i\}_{i=1}^{p}$ is the variable which is 'globally most correlated with all p variables'.

But there is a more fundamental way of looking at the implications of Hotelling's characterization, which will be considered in the next Subsection.

VI.1.2 Changing the PCA criterion

We have seen how the Correlation Matrix PCs of p variables successively maximize the sum of squared correlations with those variables, subject to orthogonality constraints.

But the nature of correlations makes them insensitive to changes of scales in the variables. Thus, the *Correlation Matrix PCs have this property regardless of whether the* p *variables are in their original units, standardized or re-scaled in any other way.* Hence, replacing the usual PCA criteria for determining the p new uncorrelated variables (components) with the new criterion of successively maximizing the SSC of each component with all p variables will result in a *truly scale-invariant set of components.* These components will no longer be Principal Components (except when the data are standardized) in that they no longer successively maximize fitted variance (or the other PCA criteria). But they do optimize the new criterion which is 'as statistically respectable' as those of PCA.

We shall henceforth refer to this new set of components as the **Most Correlated Components (MCCs)** and the method which obtains them as **Most Correlated Component Analysis (MCCA).**

It pays to consider the geometry of this change of criterion. As has been seen before, Principal Components can be defined as the vectors of $\mathbb{R}^n$ which either successively minimize the sum of squared *distances* between the lines which they define and the vectors representing the p original variables or which (in their 'natural length' version) are the longest possible linear combinations of the original variables, with unit sum of squared loadings. Changing the scales on any number of variables is equivalent to re-scaling the vectors representing them in $\mathbb{R}^n$ (Section I.2). Unless all variables are re-scaled in the same way, such changes will result in a different set of PCs. For example, a conversion of units in one variable from meters to centimeters will imply a 100-fold increase in the length of the respective vector in $\mathbb{R}^n$, which will 'attract' the first PC towards it and thereby affect the remaining PCs as well. But there is one geometric concept which remains unchanged by a magnification/contraction of any vector: the angles which that vector forms with all other vectors in $\mathbb{R}^n$. The new criterion is focusing precisely on this fact. The cosine of the angle between two vectors in $\mathbb{R}^n$ representing variables centered about their means is the correlation between those variables (Section I.2). By focusing on the sum of squared cosines of the angles between vectors, the new criterion is insensitive to changes in the magnitude of such vectors. We can loosely speak of a criterion which 'minimizes angles' rather than 'minimizes distances' between variables and components.

The fact that the MCCs are the Correlation Matrix PCs raises the question of whether it is justifiable to speak of a new method, rather than a further characterization of Correlation Matrix PCs.

The fundamental aspect of Hotelling's characterization is that it enables us to relate Correlation Matrix PCs to the (non-standardized) original variables. But this was not done by working with the criteria and concepts of PCA. The criteria *defining* PCA were replaced when making the connection between Correlation Matrix PCs and non-standardized variables. In replacing PCA's criteria with alternative criteria we are, in effect, considering a new method. This much is stressed by Krzanowski (1988) [38, (section 2.2.7)], who discusses various other methods "within the general categorization of 'projection methods'" [38, (pg.71)]. A similar reasoning presided over the coining of the term **Projection Pursuit** (see, for example, Jones and Sibson (1987) [33]). The essence of Projection Pursuit is to generalize the criteria for choosing subspaces onto which the data are to be orthogonally projected (although the focus is on $\mathbb{R}^p$ [33, (pg.3)], rather than on $\mathbb{R}^n$, and on finding subspaces that enhance certain structural traits of the data set, such as clusters). The PCA criterion of maximizing fitted variance thus becomes one of many possible alternative criteria, called **projection indices** or **indices of interestingness** [33]. MCCA replaces PCA's index of interestingness in $\mathbb{R}^n$ with an alternative. It is fortunate that in so doing, we can avoid the need for numerical, as opposed to analytical, solutions which characterizes most projection indices [38, (pg. 78)].

A similar view is also taken by Meredith and Millsap (1985) [43] in an interesting paper. They speak of 'Component Analyses' referring to any number of alternative criteria used to determine a set of uncorrelated linear combinations of p given variables. One criterion which they consider is, in effect, an alternative to Hotelling's characterization.

Meredith and Millsap look at the regression of variables onto linear combinations of themselves, much as Hotelling did in his original derivation of PCA (see the discussion in Subsection I.4.1). In our descriptive setting, this corresponds to orthogonally projecting the columns of a data matrix $\mathbf{X}$ onto subspaces of $\Re(\mathbf{X})$ spanned by certain linear combinations $\mathbf{XW}$ of those columns, where $\mathbf{W}$ is some $p \times q$ matrix. In other words, it corresponds to orthogonally projecting the columns of $\mathbf{X}$ onto $\Re(\mathbf{XW})$. One criterion considered by Meredith and Millsap is to choose $\mathbf{W}$ in such a way that the sum of squared correlations (SSC) between the columns of $\mathbf{X}$ *and their projections on* $\Re(\mathbf{XW})$ is maximized. In other words, they seek the 'most faithful' or 'best' representation of the columns of $\mathbf{X}$, as defined by the *maximum sum of squared multiple correlations between the columns of* $\mathbf{XW}$ *and each of the columns of* $\mathbf{X}$.

This criterion will generate the same set of vectors (the Correlation Matrix PCs) as the one considered previously if we take $\mathbf{W}$ to have a single column and then increase the number of columns in steps of one, always retaining those vectors which solved the problem in the previous step. Otherwise, the columns of $\mathbf{W}$ will only be defined up to a rotation.

In fact, regardless of the size of $\mathbf{W}$ and assuming only that it is of rank q, the projections of $\mathbf{X}$'s columns onto $\Re(\mathbf{XW})$ are (see Proposition A.19) the columns of

$\hat{\mathbf{X}}$, defined (with $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$) as:

$$\begin{aligned} \hat{\mathbf{X}} = \mathbf{P}_{XW}\mathbf{X} &= (\mathbf{XW})[(\mathbf{XW})'(\mathbf{XW})]^{-1}(\mathbf{XW})'\mathbf{X} \\ \Longleftrightarrow \hat{\mathbf{X}} &= \mathbf{XW}(\mathbf{W}'\mathbf{SW})^{-1}\mathbf{W}'\mathbf{S} \qquad \text{(VI.1.4)} \end{aligned}$$

In PCA, we can speak of minimizing the sum of squared distances from the columns of $\mathbf{X}$ to the columns of $\hat{\mathbf{X}}$. This would amount to seeking a matrix $\mathbf{W}$ which minimized:

$$\begin{aligned} \tfrac{1}{n}\|\mathbf{X} - \hat{\mathbf{X}}\|^2 &= \operatorname{tr}(\mathbf{S}) - \operatorname{tr}\left(\tfrac{1}{n}\hat{\mathbf{X}}'\hat{\mathbf{X}}\right) \\ &= \operatorname{tr}(\mathbf{S}) - \operatorname{tr}(\mathbf{W}'\mathbf{S}^2\mathbf{W}[\mathbf{W}'\mathbf{SW}]^{-1}) \qquad \text{(VI.1.5)} \end{aligned}$$

since $\mathbf{X}'\hat{\mathbf{X}} = \mathbf{X}'\mathbf{P}_{XW}\mathbf{X} = \hat{\mathbf{X}}'\hat{\mathbf{X}}$. An equivalent formulation is to maximize the second term in (VI.1.5), *i.e.*, to maximize $\frac{1}{n}\|\hat{\mathbf{X}}\|^2$, the sum of variances of the fitted vectors. It is known (see, for example, [43, (pgs.496-497)]) that the maximization of the second term in (VI.1.5) lies in taking the columns of $\mathbf{W}$ to be any q vectors spanning the same subspace of $\mathbb{R}^p$ as the first q eigenvectors of $\mathbf{S}$. By following the step-by-step approach described previously we end up with precisely the first q eigenvectors of $\mathbf{S}$ as the columns of $\mathbf{W}$. (Incidentally, minimizing $\|\mathbf{X} - \hat{\mathbf{X}}\|^2$ can also be interpreted as minimizing the sum of squared distances in $\mathbb{R}^p$ between the *rows* of $\mathbf{X}$ and those of $\hat{\mathbf{X}}$; equation (VI.1.4) and the discussion which followed Theorem V.6 then tells us that the above solution also corresponds to orthogonally projecting the rows of $\mathbf{X}$ onto $\Re(\mathbf{W})$, as is known).

But the Meredith and Millsap approach focuses on the angles in $\mathbb{R}^n$ between the j-th column of $\mathbf{X}$, $\mathbf{x}_j$, and the corresponding column of $\hat{\mathbf{X}}$, $\mathbf{P}_{XW}\mathbf{x}_j$. The cosine of this angle, which is the *multiple correlation coefficient between* $\mathbf{x}_j$ *and the* q *columns of* $\mathbf{XW}$, can be seen to be:

$$\cos(\mathbf{x}_j, \mathbf{P}_{XW}\mathbf{x}_j) = \frac{<\mathbf{x}_j, \mathbf{P}_{XW}\mathbf{x}_j>}{||\mathbf{x}_j|| \cdot ||\mathbf{P}_{XW}\mathbf{x}_j||} = \frac{||\mathbf{P}_{XW}\mathbf{x}_j||}{||\mathbf{x}_j||} \tag{VI.1.6}$$

$$= \sqrt{\left\langle \frac{\mathbf{x}_j}{||\mathbf{x}_j||}, \mathbf{P}_{XW}\frac{\mathbf{x}_j}{||\mathbf{x}_j||} \right\rangle} \tag{VI.1.7}$$

The square of this cosine is the j-th diagonal element of the matrix:

$$\mathbf{D}_S^{-\frac{1}{2}}(\ \tfrac{1}{n}\mathbf{X}'\hat{\mathbf{X}})\mathbf{D}_S^{-\frac{1}{2}} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{SW}(\mathbf{W}'\mathbf{SW})^{-1}\mathbf{W}'\mathbf{SD}_S^{-\frac{1}{2}} \tag{VI.1.8}$$

(although the off-diagonal elements of this matrix are *not* the cosines squared of the angle between one variable and the projection of another, since the last equality in (VI.1.6) is not valid in that case). The sum of squared cosines is therefore the trace of the matrix (VI.1.8) and it is this trace which we wish to maximize.

With no loss of generality, we can write $\mathbf{W} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{U}$ and obtain the equivalent problem of maximizing:

$$\mathrm{tr}(\mathbf{RU}[\mathbf{U}'\mathbf{RU}]^{-1}\mathbf{U}'\mathbf{R}) = \mathrm{tr}(\mathbf{U}'\mathbf{R}^2\mathbf{U}[\mathbf{U}'\mathbf{RU}]^{-1}) \tag{VI.1.9}$$

This being an analogue of the second term in (VI.1.5), the solutions consist in taking the columns of $\mathbf{U}$ to be vectors spanning the same subspace of $\mathbb{R}^p$ as the first

q eigenvectors of $\mathbf{R}$. The step-by-step approach used before will give us precisely the first q eigenvectors of $\mathbf{R}$ as the columns of $\mathbf{U}$. This implies that $\mathbf{XW} = \mathbf{XD}_S^{-\frac{1}{2}}\mathbf{U}$ will be the first q Correlation Matrix PCs of $\mathbf{X}$. The sum of their associated eigenvalues will be the sum of squared multiple correlation coefficients between the p variables and the q MCCs. The subspace spanned by the first q MCCs thus ensures the highest possible sum of squared correlations between the p variables and their projections. We can speak of the **q-dimensional Most Correlated Subspace**.

Thus, Meredith and Millsap give a criterion which also produces the Correlation Matrix PCs without explicit reference to the standardized variables. Their results are also valid in an inferential context (the context in which they work).

Meredith and Millsap do not appear to be aware of the connection between their criterion and the property of Correlation Matrix PCs mentioned by Hotelling. But they are clearly aware of the significance of their criterion and its scale invariance. They say [43, (pg.497)]:

> We also observe that the principle leads to a scale invariant method (...). The foregoing result is compelling and astonishing in its simplicity. To the best of the writers' knowledge, it is entirely novel, which is surprising if true.

Unfortunately, the authors do not appear to have returned to the subject.

Okamoto (1969) [47], whilst investigating optimality properties of (Covariance

Matrix) PCs, also obtained an expression which is related to Meredith and Millsap's criterion. Denoting by $\mathcal{R}^2(\mathbf{x}_i, \mathbf{Y})$ the squared multiple correlation coefficient between the i-th variable $\mathbf{x}_i$ and a set of q variables represented by $\mathbf{Y}$, Okamoto notes that taking $\mathbf{Y}$ to be the first q PCs of the set $\{\mathbf{x}_i\}_{i=1}^p$ will maximize the quantity

$$\sum_{i=1}^{p} s_{ii}\mathcal{R}^2(\mathbf{x}_i, \mathbf{Y})$$

Okamoto did not elaborate on this result. However, taking $s_{ii} = 1$, $\forall i = 1, ..., p$, that is, working with standardized variables, implies in particular that Correlation Matrix PCs maximize the sum of squared multiple correlation coefficients with all variables, *i.e.*, Meredith and Millsap's result. Fixing the variables of $\mathbf{Y}$ (the Correlation Matrix PCs) and considering only the sum $\sum_{i=1}^p \mathcal{R}^2(\mathbf{x}_i, \mathbf{Y})$, the precise normalization of the variables $\{\mathbf{x}_i\}_{i=1}^p$ becomes unimportant since multiple correlations are insensitive to changes of scale in $\mathbf{x}_i$.

In the next Section we shall explore MCCA in greater detail.

VI.2 MCCA in detail

VI.2.1 Variables, MCCs and PCs

So far, we have focused on how one or several MCCs perform with regards to all p variables. However, it is also possible to quantify the performance of one or several MCCs with regards to any subset of the p variables in a relatively simple fashion.

At the same time, a fuller understanding of MCCA will also require a comparative analysis with PCA, enabling us to understand how PCs perform in terms of MCCA's 'minimizing angles' criterion and how MCCs perform in terms of, for example, PCA's 'maximizing fitted variance' criterion.

In the following five Theorems these questions are considered. It is pleasing that answers can be found by looking at the diagonal elements in the terms of the spectral decompositions of the covariance matrix $\mathbf{S}$ or the correlation matrix $\mathbf{R}$.

Theorem VI.2 *Let $\{\mathrm{x}_i\}_{i=1}^{p}$ be* p *variables and $\{\phi_j\}_{j=1}^{p}$ their MCCs. Let $\mathbf{R}$ be the correlation matrix of the variables $\{\mathrm{x}_i\}_{i=1}^{p}$; $\mathbf{R}_j$ the j-th term and $\mathbf{R}_{[q]}$ the sum of the first* q *terms in $\mathbf{R}$'s spectral decomposition. Then:*

i). the squared correlation between the variable x_i and the j-th MCC ϕ_j is given by the i-th diagonal element of $\mathbf{R}_j$.

ii). the squared multiple correlation coefficient between the i-th variable x_i and the first q *MCCs $\{\phi_j\}_{j=1}^{q}$ is given by the i-th diagonal element of $\mathbf{R}_{[q]}$.*

Proof.

i). It is a standard result (see equation (II.2.1) with $s_{ii} = 1$) that the correlation between the i-th variable and the j-th Correlation Matrix PC is given by $\sqrt{\mu_j}b_{ij}$, where $(\mu_j, \mathbf{b}_j)$ is the j-th eigenpair of $\mathbf{R}$ and b_{ij} is the i-th coefficient of $\mathbf{b}_j$. This expression is obviously valid whether or not x_i has been standardized.

Hence, the correlation between the *i*-th variable and the *j*-th MCC, which we will denote by ρ_{ij}, is given by $\rho_{ij} = \mu_j b_{ij}^2 = (\mathbf{R}_j)_{(i,i)}$. Alternatively, we could work from equations (VI.1.7) and (VI.1.8) as in the proof of point ii).

ii). We seek the squared correlation between $\mathbf{x}_i$ and its orthogonal projection onto the subspace spanned by the first q MCCs. In equations (VI.1.7) and (VI.1.8) we saw how this was given by the *i*-th diagonal element of the matrix

$$\mathbf{D}_S^{-\frac{1}{2}}\mathbf{SW}(\mathbf{W}'\mathbf{SW})^{-1}\mathbf{W}'\mathbf{SD}_S^{-\frac{1}{2}}$$

with $\mathbf{W} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{B}_q$, where $\mathbf{B}_q$ is the $p \times q$ matrix of $\mathbf{R}$'s first q eigenvectors. Hence, it is the *i*-th diagonal element of the matrix

$$\mathbf{RB}_q(\mathbf{B}_q'\mathbf{RB}_q)^{-1}\mathbf{B}_q'\mathbf{R} = \mathbf{B}_q \boldsymbol{\mu}_q \boldsymbol{\mu}_q^{-1} \boldsymbol{\mu}_q \mathbf{B}_q = \mathbf{R}_{[q]} \qquad \text{(VI.2.1)}$$

where $\boldsymbol{\mu}_q$ is the $q \times q$ diagonal matrix of $\mathbf{R}$'s first q eigenvalues.

□

It should be noted that nothing in the proof of Theorem VI.2 restricts the conclusion in point ii) only to the *first* q MCCs, although this is undoubtedly the most interesting case. Any other subset of q MCCs can be considered, in which case $\mathbf{R}_{[q]}$ must be replaced by the sum of the q corresponding terms in $\mathbf{R}$'s spectral decomposition.

Point ii) of Theorem VI.2 considers multiple correlations between $\mathbf{x}_i$ and q MCCs. From a repeated application of point i) we see that this squared multiple correlation is also the sum of squared correlations between $\mathbf{x}_i$ and each individual MCC. We can also consider the similar problem reversing the roles of variables and MCCs. The sum of squared correlations between the j-th MCC and any m variables is, from point i), the sum of the corresponding m diagonal terms of $\mathbf{R}_j$. For $m = p$ we have the 'MCCA criterion' discussed in Section VI.1. But, unfortunately, this is *not* the squared multiple correlation between the MCC and the m variables, despite the superficial similarity with the previous case. Inverting the roles of MCCs and variables in the discussion leading to (VI.2.1) produces an analogue of equation (II.4.1) for the squared multiple correlation between $\boldsymbol{\phi}_j$ and m variables in the set $\mathcal{S}$:

$$r_m = \sqrt{\mu_j} \cdot \sqrt{\mathbf{d}_j^{[\mathcal{S}]'}(\mathbf{S}_{[\mathcal{S}]})^{-1}\mathbf{d}_j^{[\mathcal{S}]}} \qquad \text{(VI.2.2)}$$

where $\mathbf{d}_j^{[\mathcal{S}]}$ is the vector of $\mathbb{R}^m$ with the loadings in $\mathbf{D}_S^{-\frac{1}{2}}\mathbf{b}_j$ which correspond to the m variables with indices in subset $\mathcal{S}$, and $\mathbf{S}_{[\mathcal{S}]}$ is the corresponding submatrix of $\mathbf{S}$.

The squared correlations in points i) and ii) must naturally take values between zero and one. However, the fact that they are (necessarily non-negative) diagonal elements of $\mathbf{R}_j$ and $\mathbf{R}_{[q]}$ also implies that they cannot exceed μ_j and $\sum_{j=1}^{q} \mu_j$, respectively, *i.e.*, the traces of $\mathbf{R}_j$ and $\mathbf{R}_{[q]}$.

It should be stressed that the off-diagonal elements of $\mathbf{R}_j$ and $\mathbf{R}_{[q]}$ will not have similarly interesting interpretations. For example, the (i, h)-th element of $\mathbf{R}_j$ is, by definition, $\mu_j b_{ij} b_{hj}$. From equation (II.2.1) this can be seen to be the product of the correlation between $\mathbf{x}_i$ and $\boldsymbol{\phi}_j$ with the correlation between $\mathbf{x}_h$ and $\boldsymbol{\phi}_j$, a quantity of no particular interest.

We now consider how subsets of PCs perform in terms of multiple correlations with individual variables. The result is a pleasing analogue of the above Theorem.

Theorem VI.3 *Let $\{\mathrm{x}_i\}_{i=1}^p$ be p variables and $\{\boldsymbol{\xi}_j\}_{j=1}^p$ their PCs (with any given normalization). Let $\mathbf{S}$ be the covariance matrix of the variables $\{\mathrm{x}_i\}_{i=1}^p$; $\mathbf{S}_j$ the j-th term and $\mathbf{S}_{[q]}$ the sum of the first q terms in the spectral decomposition of $\mathbf{S}$. Let $\mathbf{D}_S^{-\frac{1}{2}}$ be the diagonal matrix of reciprocal standard deviations of the variables. Then:*

i). the squared correlation between the i-th variable x_i and the j-th PC $\boldsymbol{\xi}_j$ is given by the i-th diagonal element of $\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_j\mathbf{D}_S^{-\frac{1}{2}}$.

ii). the squared multiple correlation between the i-th variable x_i and the first q PCs, $\{\boldsymbol{\xi}_j\}_{j=1}^q$, is given by the i-th diagonal element of $\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_{[q]}\mathbf{D}_S^{-\frac{1}{2}}$.

Proof.

i). Equation (II.2.1) tells us that the correlation between x_i and $\boldsymbol{\xi}_j$ is $a_{ij}\sqrt{\frac{\lambda_j}{s_{ii}}}$, where $(\lambda_j, \mathbf{a}_j)$ is the j-th eigenpair of $\mathbf{S}$, a_{ij} is the i-th element of $\mathbf{a}_j$ and s_{ii} is the i-th diagonal element of $\mathbf{S}$. Hence, the correlation squared is

$$\rho_{ij}^2 = \lambda_j \frac{a_{ij}^2}{s_{ii}} = (\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_j\mathbf{D}_S^{-\frac{1}{2}})_{(i,i)}$$

As in Theorem VI.2, an alternative proof is possible along the lines of the proof of point ii).

ii). The squared multiple correlation of x_i with the first q PCs is the i-th diagonal element of matrix $\mathbf{D}_S^{-\frac{1}{2}}\mathbf{SW}(\mathbf{W}'\mathbf{SW})^{-1}\mathbf{W}'\mathbf{SD}_S^{-\frac{1}{2}}$, with $\mathbf{W} = \mathbf{A}_q$, the $p \times q$ matrix whose columns are the first q eigenvectors of $\mathbf{S}$. We then have:

$$\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}\mathbf{A}_q(\mathbf{A}_q{}'\mathbf{S}\mathbf{A}_q)^{-1}\mathbf{A}_q{}'\mathbf{S}\mathbf{D}_S^{-\frac{1}{2}} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{A}_q\boldsymbol{\Lambda}_q\boldsymbol{\Lambda}_q{}^{-1}\boldsymbol{\Lambda}_q\mathbf{A}_q{}'\mathbf{D}_S^{-\frac{1}{2}} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_{[q]}\mathbf{D}_S^{-\frac{1}{2}}$$

where $\boldsymbol{\Lambda}_q$ is the $q \times q$ diagonal matrix with the first q eigenvalues of $\mathbf{S}$.

□

The comments which followed Theorem VI.2 also apply to Theorem VI.3, with the obvious adjustments. The new and interesting aspect of Theorem VI.3 is that *replacing MCCs with PCs led to replacing the terms in* $\mathbf{R}$*'s spectral decomposition with the corresponding terms in the spectral decomposition of* $\mathbf{S}$ *under a congruence defined by* $\mathbf{D}_S^{-\frac{1}{2}}$. It is as if, loosely speaking, we focus on the covariance matrix $\mathbf{S}$ and take the appropriate terms in the spectral decomposition either before or after the standardizing congruence. This congruence acts as a normalizer which ensures that the elements being considered are all smaller than or equal to 1.

The sum of squared multiple correlations of the first q PCs with all p variables (the 'MCCA criterion') is therefore given by the trace of $\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_{[q]}\mathbf{D}_S^{-\frac{1}{2}}$.

We now turn our attention to how MCCs and PCs compare in terms of the criteria which PCA optimize. Among the several such criteria, we choose to work with one that produces particularly pleasing results for our comparison. We shall consider the '*fitted variances*', that is, the variances of the projections of each variable onto the subspaces spanned by given subsets of the PCs and MCCs.

The next Theorem describes how PCs perform in terms of their own criterion.

Thus, it is a Theorem which does not involve MCCs. But the analogy with Theorems VI.2 and VI.3 is all too obvious. Besides, it paves the way for Theorem VI.5 where the performance of MCCs in terms of this criterion will be considered.

Theorem VI.4 *Let $\{\mathbf{x}_i\}_{i=1}^{p}$ be* p *variables and $\{\boldsymbol{\xi}_j\}_{j=1}^{p}$ their PCs (with any given normalization). Let $\mathbf{S}$ be the covariance matrix of the variables $\{\mathbf{x}_i\}_{i=1}^{p}$; $\mathbf{S}_j$ the j-th term and $\mathbf{S}_{[q]}$ the sum of the first* q *terms in the spectral decomposition of $\mathbf{S}$. Then:*

i). the variance of the orthogonal projection of the i-th variable $\mathbf{x}_i$ onto the subspace spanned by the j-th PC $\boldsymbol{\xi}_j$ is given by the i-th diagonal element of $\mathbf{S}_j$.

ii). the variance of the orthogonal projection of the i-th variable $\mathbf{x}_i$ onto the subspace spanned by the first q *PCs, $\{\boldsymbol{\xi}_j\}_{j=1}^{q}$, is given by the i-th diagonal element of $\mathbf{S}_{[q]}$.*

Proof.

i). Again, using the notation from Subsection IV.1.2, we can write the projection of the i-th variable $\mathbf{x}_i$ onto the subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{XW}$ as:

$$\mathbf{P}_{XW}\mathbf{x}_i = \mathbf{XW}(\mathbf{W}'\mathbf{X}'\mathbf{XW})^{-1}\mathbf{W}'\mathbf{X}'\mathbf{x}_i = \mathbf{XW}(\mathbf{W}'\mathbf{SW})^{-1}\mathbf{W}'\mathbf{Se}_i \qquad \text{(VI.2.3)}$$

where $\mathbf{e}_i$ is the i-th vector in the canonical basis of $\mathbb{R}^p$. The variance of this projected vector is given by:

$$var(\mathbf{P}_{XW}\mathbf{x}_i) = \frac{1}{n}||\mathbf{P}_{XW}\mathbf{x}_i||^2 = \mathbf{e}_i'\mathbf{SW}(\mathbf{W}'\mathbf{SW})^{-1}\mathbf{W}'\mathbf{Se}_i \qquad \text{(VI.2.4)}$$

PCs result from taking $\mathbf{W}$ to have eigenvectors of $\mathbf{S}$ in its columns. Assuming $\mathbf{W} = \mathbf{a}_j$, the j-th eigenvector of $\mathbf{S}$, we have:

$$var(\mathbf{P}_{XW}\mathbf{x}_i) = \mathbf{e}_i{}'\lambda_j\mathbf{a}_j\mathbf{a}_j{}'\mathbf{e}_i = \lambda_j a_{ij}^2 = (\mathbf{S}_j)_{(i,i)}$$

where a_{ij} is the i-th element of $\mathbf{a}_j$.

ii). This case corresponds to taking $\mathbf{W} = \mathbf{A}_q$ in (VI.2.3), where $\mathbf{A}_q$ is the $p\times q$ matrix whose columns are the first q eigenvectors of $\mathbf{S}$. We then have, from (VI.2.4):

$$var(\mathbf{P}_{XW}\mathbf{x}_i) = \mathbf{e}_i{}'\mathbf{S}_{[q]}\mathbf{e}_i = (\mathbf{S}_{[q]})_{(i,i)}$$

□

Although the quantities being considered now are variances rather than correlations, there is a nice parallel with the previous two Theorems. The terms or sum of terms in the spectral decomposition of $\mathbf{S}$ now replace those of $\mathbf{R}$ or their own congruences under $\mathbf{D}_S^{-\frac{1}{2}}$, to provide the information which is required. As with previous Theorems, point ii) remains valid if *any* q – rather than the first q – PCs are considered, with $\mathbf{S}_{[q]}$ being replaced by the sum of the q corresponding terms in the spectral decomposition of $\mathbf{S}$.

The i-th diagonal term of $\mathbf{S}_j$ lies between zero and $\min\{\lambda_j, s_{ii}\}$. The i-th diagonal term of $\mathbf{S}_{[q]}$ lies between zero and $\min\{\sum_{j=1}^{q}\lambda_j, s_{ii}\}$. Adding up any m diagonal elements of $\mathbf{S}_j$ or $\mathbf{S}_{[q]}$ will give the sum of fitted variances for the corresponding m variables. Taking $m = p$ produces the criterion which PCA optimizes.

The off-diagonal elements of $\mathbf{S}_j$ are the covariances between two fitted variables, or between the original variable and the projection of a different variable (as can be seen by adapting the proof of the last Theorem). This is a slightly more interesting situation than when considering the MCCA criterion.

Finally, we shall consider the performance of MCCs and Most Correlated Subspaces in terms of the 'PCA criterion' of fitted variances. The result is again pleasing.

Theorem VI.5 *Let $\{\mathbf{x}_i\}_{i=1}^{p}$ be p variables and $\{\boldsymbol{\phi}_j\}_{j=1}^{p}$ their MCCs. Let $\mathbf{R}$ be the correlation matrix of the variables $\{\mathbf{x}_i\}_{i=1}^{p}$, $\mathbf{R}_j$ the j-th term and $\mathbf{R}_{[q]}$ the sum of the first q terms in $\mathbf{R}$'s spectral decomposition. Let $\mathbf{D}_S^{\frac{1}{2}}$ be the diagonal matrix of standard deviations of the variables. Then:*

i). the variance of the orthogonal projection of the i-th variable $\mathbf{x}_i$ onto the subspace spanned by the j-th MCC $\boldsymbol{\phi}_j$ is given by the i-th diagonal element of $\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_j\mathbf{D}_S^{\frac{1}{2}}$.

ii). the variance of the orthogonal projection of the i-th variable $\mathbf{x}_i$ onto the subspace spanned by the first q MCCs, $\{\boldsymbol{\phi}_j\}_{j=1}^{q}$, is given by the i-th diagonal element of $\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_{[q]}\mathbf{D}_S^{\frac{1}{2}}$.

Proof.

i). The value we seek is given by equation (VI.2.4) with $\mathbf{W} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{b}_j$, where $\mathbf{b}_j$ is $\mathbf{R}$'s j-th eigenvector. Hence, we have:

$$\begin{aligned} var\left(\mathbf{P}_{\mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}\mathbf{b}_j}\mathbf{x}_i\right) &= \mathbf{e}_i'\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}\mathbf{b}_j(\mathbf{b}_j'\mathbf{R}\mathbf{b}_j)^{-1}\mathbf{b}_j'\mathbf{R}\mathbf{D}_S^{\frac{1}{2}}\mathbf{e}_i \\ &= \mathbf{e}_i'\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_j\mathbf{D}_S^{\frac{1}{2}}\mathbf{e}_i = (\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_j\mathbf{D}_S^{\frac{1}{2}})_{(i,i)} \end{aligned}$$

ii). Also from (VI.2.4), but taking $\mathbf{W} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{B}_q$, where $\mathbf{B}_q$ is the $p \times q$ matrix whose columns are the first q eigenvectors of $\mathbf{R}$, we get:

$$var\left(\mathbf{P}_{\mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}\mathbf{B}_q}\mathbf{x}_i\right) = \mathbf{e}_i{}'\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_{[q]}\mathbf{D}_S^{\frac{1}{2}}\mathbf{e}_i$$

□

Similar comments to those following Theorem VI.4 apply.

The essence of the previous four Theorems can be nicely summarized as in Table VI.1. For each of the four combinations of one method (PCA or MCCA) with one criterion (fitted variance or sum of squared correlations), the focus is on the diagonal elements of terms or sums of terms of the spectral decompositions of either $\mathbf{S}$ or $\mathbf{R}$, or congruences thereof. Denoting these terms or sums of terms by $\mathbf{R}_{(*)}$ and $\mathbf{S}_{(*)}$ enables us to grasp the symmetries which characterize the comparative study.

	PCA	MCCA
PCA criterion (fitted variance)	$\mathbf{S}_{(*)}$	$\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_{(*)}\mathbf{D}_S^{\frac{1}{2}}$
MCCA criterion (SSC)	$\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_{(*)}\mathbf{D}_S^{-\frac{1}{2}}$	$\mathbf{R}_{(*)}$

Table VI.1: Matrices whose diagonal elements are meaningful indicators for the performance of PCA and MCCA

Table VI.1 shows that the performance of PCA is associated with the decomposition of the covariance matrix or a fixed congruence ($\mathcal{N}_{\mathbf{D}_S^{-\frac{1}{2}}}$ in the notation of Chapter IV – see (IV.3.3)) of this decomposition. The performance of MCCA is associated with the decomposition of the correlation matrix or a fixed congruence ($\mathcal{N}_{\mathbf{D}_S^{\frac{1}{2}}}$) of this decomposition. If we read by rows, it is not surprising to find that the fitted variance criterion is associated with matrices 'of the order of magnitude of a spectral decomposition of the covariance matrix', whereas SSC is given by matrices 'of the order of magnitude of a spectral decomposition of a correlation matrix'. Looking at the diagonals of the Table, we see that the performance of each method under its own criterion is given by the 'pure' terms in the spectral decompositions, whereas the performance under the other method's criterion is given by the congruences of these terms.

But it is possible to say something further concerning the performance of *both methods* as far as *fitted variance* is concerned.

The criterion for PCA is to maximize global fitted variance. In Theorems VI.4 and VI.5 we saw how the diagonal elements of $\mathbf{S}_j$ and $\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_j\mathbf{D}_S^{\frac{1}{2}}$ gave us the breakdown *for each variable* of the variance fitted on the j-th PC and the j-th MCC. In judging the performance of these components as far as each variable is concerned, we might prefer to replace the (absolute) fitted variance of each variable with the *proportion* of each variable's variance that is accounted for by the fit onto a given component. The next Theorem considers this point.

Theorem VI.6 *Let $\{\mathrm{x}_i\}_{i=1}^{p}$ be* p *variables. Let* $\mathbf{S}$ *and* $\mathbf{R}$ *be their covariance and correlation matrices, respectively. Then:*

i). the proportion of the variance of the i-th variable, x_i, which is accounted for by its projection onto the j-th PC is given by the i-th diagonal element of

$\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_j\mathbf{D}_S^{-\frac{1}{2}}$, *where* $\mathbf{S}_j$ *is the j-th term in the spectral decomposition of* $\mathbf{S}$.

ii). the proportion of the variance of the i-th variable, $\mathbf{x}_i$, *which is accounted for by its projection onto the j-th MCC is given by the i-th diagonal element of* $\mathbf{R}_j$, *where* $\mathbf{R}_j$ *is the j-th term in the spectral decomposition of* $\mathbf{R}$.

iii). the proportion of the variance of the i-th variable, $\mathbf{x}_i$, *which is accounted for by its projection onto the subspace spanned by the first* q *PCs is given by the i-th diagonal element of* $\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_{[q]}\mathbf{D}_S^{-\frac{1}{2}}$, *where* $\mathbf{S}_{[q]}$ *is the sum of the first* q *terms in the spectral decomposition of* $\mathbf{S}$.

iv). the proportion of the variance of the i-th variable, $\mathbf{x}_i$, *which is accounted for by its projection onto the subspace spanned by the first* q *MCCs is given by the i-th diagonal element of* $\mathbf{R}_{[q]}$, *where* $\mathbf{R}_{[q]}$ *is the sum of the first* q *terms in* $\mathbf{R}$*'s spectral decomposition.*

Proof. We are interested in making the results of points i) and ii) in both Theorems VI.4 and VI.5 relative to the variance of the i-th variable. One way of dividing each diagonal element of the matrices $\mathbf{S}_{(*)}$ and $\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_{(*)}\mathbf{D}_S^{\frac{1}{2}}$ (where the asterisk stands for either j or $[q]$) by the variance of the corresponding variable is to pre- and post-multiply those matrices by $\mathbf{D}_S^{-\frac{1}{2}}$. The results are therefore proved. □

The most striking feature about this result is, of course, that in both the case of PCs and MCCs we are taken back to the diagonal elements of the matrices which were considered in Theorems VI.2 and VI.3. The implication is that the proportion of variance of a variable which is accounted for by its orthogonal projection onto a PC, an MCC or a subspace spanned by subsets of PCs and MCCs is equal to the correlation squared between the variable and its fit onto those components or sets of components. Actually, this relation remains true when the variable is orthogonally projected in $\mathbb{R}^n$ (linearly regressed) onto *any vector or any subspace of* $\mathbb{R}^n$, and had already been given in equation (VI.1.6) for projections onto subspaces of $\Re(\mathbf{X})$. By again resorting to the geometry in $\mathbb{R}^n$, we see that this proportion is merely the definition of the cosine between $\mathbf{x}_i$ and any orthogonal projection of $\mathbf{x}_i$. In fact,

$$\cos(\mathbf{x}_i, \mathbf{P}\mathbf{x}_i) = \frac{||\mathbf{P}\mathbf{x}_i||}{||\mathbf{x}_i||} \tag{VI.2.5}$$

and the ratio on the r.h.s. of (VI.2.5) is the square root of the proportion of fitted variance.

The fact that the relation (VI.2.5) holds regardless of the subspace onto which orthogonal projections are being targeted implies that *a third optimal property characterizing MCCs* (*i.e.*, Correlation Matrix PCs) *has been found.*

Theorem VI.7 *Let* $\{\mathbf{x}_i\}_{i=1}^{p}$ *be* p *variables. The* p *uncorrelated linear combinations of these variables which successively maximize the sum of squares of the proportion of each variable's variance which is preserved by an orthogonal projection (regression) onto each of those linear combinations, is given by the MCCs (Correlation Matrix PCs) of the variables.*

Proof. We wish to maximize $\sum_{i=1}^{p} \frac{||\mathbf{Px}_i||^2}{||\mathbf{x}_i||^2}$, where $\mathbf{Px}_i$ is the orthogonal projection of $\mathbf{x}_i$ onto any given vector $\mathbf{y} \in \mathbb{R}^n$. But (VI.2.5) implies that this quantity is

$$\sum_{i=1}^{p} \cos^2(\mathbf{x}_i, \mathbf{Px}_i) = \sum_{i=1}^{p} \cos^2(\mathbf{x}_i, \mathbf{y})$$

This is the quantity which was to be successively maximized in the original derivation of MCCA (Subsection VI.1.1). Hence, the MCCs are the vectors which solve the problem. □

Thus, a further connection between PCA and MCCA has been found. MCCs can also be interpreted in terms of maximizing fitted variance, although the precise criterion used will no longer be the sum of each fitted variable's variance, but rather the sum of squared proportions of each variable's total variance which is being fitted.

The amount of information contained in the diagonal elements of the spectral decompositions of **S** and **R**, concerning the performance of PCs and MCCs, is a pleasant surprise.

VI.2.2 MCCA and dimensionality reduction

One frequent application of PCA is the reduction of dimensionality of a data set, *i.e.*, replacing the p original variables with q $(q < p)$ uncorrelated variables which capture most of the variability in the original data.

But it is also relevant to replace the p original variables with q uncorrelated variables which are 'globally most correlated' with the original variables. Thus, *MCCA* *also has an important role to play in dimensionality reduction.* Replacing p variables with their first q MCCs, rather than PCs, amounts to taking a different view of what is the most 'faithful' or interesting q-dimensional approximation to the information conveyed by the original variables.

It is possible to quantify the 'quality' of a given reduction in dimensionality via MCCA, in much the same way as is done in PCA with the 'percentage variance accounted for' by a given subset of PCs.

We saw in Subsection VI.1.1 (and Theorem VI.2) how each eigenvalue μ_j of the correlation matrix $\mathbf{R}$ represented the sum of squared correlations (SSC) between the corresponding MCC, $\mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}\mathbf{b}_j$, and all p variables. The general bounds zero and $\mathrm{tr}(\mathbf{R}) = \mathrm{p}$ on an eigenvalue of any correlation matrix reflect the extreme scenarios in this respect: uncorrelatedness with all p variables and perfect correlation with all p variables (which would naturally imply that all the variables coincide or are merely reflections of each other about the origin). The larger an eigenvalue, the more 'globally correlated' will the associated MCC be with the p variables. Hence, *the relative eigenvalue $\pi_j = \mu_j/p$ reflects the quality of a given MCC as a substitute for the* p *variables, on a scale from zero to one.*

Furthermore, *the quality of the replacement of the* p *variables with* q *MCCs can also be summarized by adding the MCC's associated relative eigenvalues,* as is the case with PCs. *The q-th relative eigenvalue of* $\mathbf{R}$ is the sum of the mean squared correlations between the p variables and each of the q MCCs, *i.e.*, the *mean squared multiple correlation between each of the variables and the* q *MCCs*, as was seen before.

It is interesting to note that, as with PCA, these indicators of how well correlated with the variables are one or several MCCs depend only on the correlation matrix $\mathbf{R}$ and not on the precise data set which generated $\mathbf{R}$.

A general pattern which emerges from many data sets is that the first few cumulative relative eigenvalues of covariance matrices tend to be larger than their counterparts in the corresponding correlation matrices. In other words, the relative eigenvalues of correlation matrices tend to be more similar in magnitude than those of the covariance matrices which are associated with it. The implication of this trend for our purposes is that *more MCCs than PCs will tend to be needed to ensure the same 'quality of fit'*, by their respective criteria.

But the above statement must be qualified in two ways: (i) the criteria of 'quality of fit' used on both cases are different and, strictly speaking, not directly comparable; (ii) the trend does not apply to *all* covariance and correlation matrices.

A simple counterexample to the general trend is given by the covariance matrix

$$\mathbf{S} = \begin{pmatrix} a^2 & \rho a^2 & 0.00 \\ \rho a^2 & a^2 & 0.00 \\ 0.00 & 0.00 & b^2 \end{pmatrix}$$

The relative eigenvalues of this covariance matrix are $\frac{b^2}{2a^2+b^2}$ and $\frac{a^2(1\pm\rho)}{2a^2+b^2}$. For the corresponding correlation matrix (which results from taking $a = b = 1$) they are $(1+\rho)/3, 1/3, (1-\rho)/3$. The relative eigenvalue $(1+\rho)/3$ of the correlation matrix is the largest if $\rho > 0$. The eigenvalue $\frac{a^2(1+\rho)}{2a^2+c^2}$ is the largest if $a^2(1+\rho) > c^2$. With these assumptions, the first relative eigenvalue of the correlation matrix is larger than its counterpart of the covariance matrix if $\frac{1+\rho}{3} > \frac{a^2(1+\rho)}{2a^2+c^2}$, *i.e.*, if $c^2 > a^2$. Thus, whenever $a^2 < c^2 < a^2(1+\rho)$, the quality of fit with a single MCC exceeds that of a single (covariance matrix) PC.

VI.2.3 MCCA and $\mathbb{R}^p$

What has been said so far about MCCA centers almost exclusively on the space $\mathbb{R}^n$. The MCCs are vectors in $\mathbb{R}^n$ and the criterion used to determine them applies in that space.

In PCA, there is an important and gratifying interplay between the spaces $\mathbb{R}^n$ and $\mathbb{R}^p$. There is a simultaneous solution of least squares problems in both spaces. Unfortunately, MCCA does not behave quite so nicely in $\mathbb{R}^p$.

One set of vectors in $\mathbb{R}^p$ that arose from the discussion in the previous section was the set of vectors of coefficients in the linear combinations (loadings) defining each MCC. These were the vectors $\{\mathbf{D}_S^{-\frac{1}{2}}\mathbf{b}_j\}_{j=1}^{p}$ where $\mathbf{b}_j$ is the j-th eigenvector of the correlation matrix $\mathbf{R}$.

A first comment about these vectors of loadings is that, unlike the MCCs which they define, they are scale-dependent. The loadings will change with re-scalings of the variables, as a result of the presence of $\mathbf{D}_S^{-\frac{1}{2}}$ (nor could it be otherwise if the components are to be scale-invariant). In itself, this would greatly reduce the importance of these vectors in a method whose main advantage over PCA is its scale-invariance.

A second draw-back with the vectors of loadings is that, unlike their PCA counterparts, they are not orthogonal (unless the variables are all of unit variance). They are $\mathbf{D}_S$-orthonormal and $\mathbf{S}$-orthogonal (although not $\mathbf{S}$-orthonormal), yet nothing of great interest can be said for the standard inner product in $\mathbb{R}^p$.

A third and more important problem is that these vectors do not appear to be the solutions to any interesting optimization problem in $\mathbb{R}^p$. It will be recalled that, in PCA, the vectors of loadings for the PCs were also the Principal Axes in $\mathbb{R}^p$, *i.e.*, the orthonormal vectors defining the successive subspaces which best fitted the data in the sense of minimizing the sum of squared distances from the rows of $\mathbf{X}$ to the subspaces. This was a direct result of the fact that orthogonally projecting the columns of $\mathbf{X}$ onto $\Re(\mathbf{XA}_q)$ – where $\mathbf{A}_q$ has the first q eigenvectors of $\mathbf{S}$ as its columns – was equivalent to orthogonally projecting the rows of $\mathbf{X}$ onto $\Re(\mathbf{A}_q)$ (see also Theorem V.6 and the subsequent comments). Attempting an analogous approach, *i.e.*, considering the orthogonal projections of the columns of $\mathbf{X}$ onto $\Re(\mathbf{XD}_S^{-\frac{1}{2}}\mathbf{B}_q)$ – where $\mathbf{B}_q$ is $\mathbf{A}_q$'s counterpart for the correlation matrix $\mathbf{R}$ – will lead to a non-orthogonal projection of the rows of $\mathbf{X}$ in $\mathbb{R}^p$. From Theorem V.6, we have that the orthogonal projection of the columns of $\mathbf{X}$ onto the subspace spanned by the first q MCCs corresponds to projecting the rows of $\mathbf{X}$ onto $\Re(\mathbf{S}\mathbf{D}_S^{-\frac{1}{2}}\mathbf{B}_q)$ along $\Re(\mathbf{D}_S^{-\frac{1}{2}}\mathbf{B}_q)^{\perp}$. Now the matrix $\mathbf{S}\mathbf{D}_S^{-\frac{1}{2}}\mathbf{B}_q$ can be written as $\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}\mathbf{B}_q = \mathbf{D}_S^{\frac{1}{2}}\mathbf{B}_q\boldsymbol{\mu}_q$, where $\boldsymbol{\mu}_q$ is the $q\times q$ diagonal matrix of $\mathbf{R}$'s first q eigenvalues. The latter diagonal matrix merely weights the columns of $\mathbf{D}_S^{\frac{1}{2}}\mathbf{B}_q$ and is therefore irrelevant in defining the subspace which they span. Hence, we are projecting the rows of $\mathbf{X}$ onto $\Re(\mathbf{D}_S^{\frac{1}{2}}\mathbf{B}_q)$ along $\Re(\mathbf{D}_S^{-\frac{1}{2}}\mathbf{B}_q)^{\perp}$. In general, $\Re(\mathbf{D}_S^{\frac{1}{2}}\mathbf{B}_q) \neq \Re(\mathbf{D}_S^{-\frac{1}{2}}\mathbf{B}_q)$ (as can be seen by taking $q = 1$). Thus, *the vectors of loadings do not even define the successive subspaces of* $\mathbb{R}^p$ *onto which the data are projected.* A second set of vectors in $\mathbb{R}^p$, the vectors $\{\mathbf{D}_S^{\frac{1}{2}}\mathbf{b}_k\}_{k=1}^{p}$, has entered the stage.

Related to the previous point is the fact that the vectors of loadings no longer play a role in the factorization of the data matrix which is most meaningful for MCCA.

Let the usual SVD of $\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}$ be given by $\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{D}_S^{-\frac{1}{2}} = \boldsymbol{\beta}\boldsymbol{\mu}\mathbf{B}'$. For the purposes of MCCA, we want to consider the matrix $\mathbf{X}$ and not the standardized data matrix $\mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}$. Hence we can write:

$$\tfrac{1}{\sqrt{n}}\mathbf{X} = \boldsymbol{\beta}\boldsymbol{\mu}(\mathbf{D}_S^{\frac{1}{2}}\mathbf{B})' \qquad \text{(VI.2.6)}$$

The columns of $\boldsymbol{\beta}$ are the unit-norm MCCs (Correlation Matrix PCs) of $\mathbf{X}$. The diagonal elements of $\boldsymbol{\mu}$ are the square roots of the SSCs of the corresponding MCC with all p variables. But the columns of $\mathbf{D}_S^{\frac{1}{2}}\mathbf{B}$ are not the vectors of loadings of the

MCCs: they are the new set of vectors which was discussed above.

This second set of vectors in $\mathbb{R}^p$ which has appeared will be referred to as the **companion vectors** of the vectors of loadings. Like the vectors of loadings, the companion vectors are scale-dependent and non-orthogonal (in the usual inner product and unless the variables all have unit variances). A curious property of these two sets of vectors in $\mathbb{R}^p$ is that they are what we might call *cross-orthonormal, i.e.*,

$$\mathbf{b}_i'\mathbf{D}_S^{-\frac{1}{2}}\mathbf{D}_S^{\frac{1}{2}}\mathbf{b}_j = \delta_{ij} \qquad \text{(VI.2.7)}$$

since the eigenvectors $\{\mathbf{b}_j\}_{j=1}^{p}$ of $\mathbf{R}$ are an orthonormal set of vectors.

The fact that the vectors of loadings and their companion vectors do not appear to play a meaningful role in MCCA is undoubtedly a major drawback of the method, when compared with PCA. Many applications and convenient properties of PCA will not be paralleled in MCCA. Furthermore, MCCA will be meaningless with data sets for which the focus is on $\mathbb{R}^p$, such as the 'sampled functions' or time-series data sets considered in Section V.3.

At the same time, these drawbacks are further arguments in favour of viewing MCCA as a distinct method, rather than just a new way of interpreting Correlation Matrix PCs.

VI.2.4 Miscellanea on MCCA

In this Subsection we consider some loose ends of MCCA.

Units of MCCs

In Subsection I.5.1 the issue of units of measurement of PCs was addressed. We now consider a similar issue for MCCs.

Each MCC is a linear combination of the variables $\{\mathbf{x}_i\}_{i=1}^{p}$. The loading for the i-th variable on the j-th MCC is given by $\frac{b_{ij}}{\sqrt{s_{ii}}}$, where s_{ii} is the variance of $\mathbf{x}_i$ and b_{ij} is the i-th element of $\mathbf{b}_j$, the j-th eigenvector of the correlation matrix $\mathbf{R}$. The coefficients b_{ij} being adimensional, we have that

$$\phi_j = \sum_{i=1}^{p} b_{ij} \frac{\mathbf{x}_i}{\sqrt{s_{ii}}}$$

involves the sum of p dimensionless terms. MCCs are therefore dimensionless variables, regardless of the nature of the p original variables. This is a much nicer situation than the one we faced in (Covariance Matrix) PCA.

Interpretation of MCCs

The interpretation of PCs by relating them to the p original variables was the topic of Chapter II. The problems raised at the time, concerning the inadequacy of interpretations based only on the loadings for each variable in the linear combination which defined the PCs, are all the more pertinent now that we have scale-dependent loadings for scale-invariant vectors. Any attempt at such 'interpretations' would produce scale-dependent interpretations of the given set of components!

The general principle for interpreting PCs suggested at the end of Chapter II, namely regressing PCs on subsets of variables, is also valid for MCCs. Unfortunately, this principle implies that there is no easy solution to the problem of deciding which subset of the p variables is most suitable for this purpose, for any given MCC, as was seen in the context of PCA.

Similarity of MCCA and PCA

The fact that the Most Correlated Components of a set of variables are its Correlation Matrix PCs implies that *an overall indicator of the degree of similarity between a PCA and an MCCA of a given data matrix* $\mathbf{X}$ *is provided by the similarity index in* $\mathbb{R}^n$ *between* $\mathbf{X}$ *and* $\mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}$.

From equation (IV.4.2) we then have that this overall indicator can be written as:

$$s_n(\mathbf{X}, \mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}) = \frac{\mathrm{tr}(\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}^2\mathbf{D}_S^{-\frac{1}{2}})}{\sqrt{\mathrm{tr}(\mathbf{S}^2)\cdot \mathrm{tr}(\mathbf{R}^2)}} \qquad \text{(VI.2.8)}$$

A little algebraic manipulation will enable us to write $s_n(\mathbf{X}, \mathbf{X}\mathbf{D}_S^{-\frac{1}{2}})$ in terms of the variances $\{s_{ii}\}_{i=1}^p$ of all p variables and all the correlations $\{r_{ij}\}_{i=1,j=1}^{p\ \ p}$ between them. We have:

$$s_n(\mathbf{X}, \mathbf{X}\mathbf{D}_S^{-\frac{1}{2}}) = \frac{\sum_{i=1}^p \sum_{j=1}^p s_{ii} r_{ij}^2}{\sqrt{\left[\sum_{i=1}^p \sum_{j=1}^p s_{ii}s_{jj}r_{ij}^2\right]\cdot\left[\sum_{i=1}^p \sum_{j=1}^p r_{ij}^2\right]}} \qquad \text{(VI.2.9)}$$

VI.3 An example

The ideas expressed so far will now be illustrated with Kendall's soil data set (Table II.1).

The (Covariance Matrix) PCs of this data set were discussed in Section II.1. Their loadings were given in Table II.3. The comparable loadings (with the same original scaling of the variables) for MCCs are given in Table VI.2 (the corresponding values from Table II.3 being also given in brackets). These loadings for MCCs have been normalized so that the *j*-th column in Table VI.2 is the unit-norm vector $\frac{\mathbf{D}_S^{-\frac{1}{2}}\mathbf{b}_j}{||\mathbf{D}_S^{-\frac{1}{2}}\mathbf{b}_j||}$.

	ϕ_1	ϕ_2	ϕ_3	ϕ_4
$\mathbf{x}_1$	0.357 (0.956)	0.015 (−0.288)	−0.009 (−0.059)	−0.155 (0.006)
$\mathbf{x}_2$	0.847 (0.294)	−0.047 (0.945)	0.016 (0.142)	0.371 (−0.018)
$\mathbf{x}_3$	0.141 (0.015)	0.758 (−0.154)	−0.292 (0.979)	0.909 (−0.136)
$\mathbf{x}_4$	0.368 (0.001)	0.651 (−0.002)	0.956 (0.137)	−0.105 (0.991)

Table VI.2: Unit norm vectors of loadings of the original variables for each MCC (and for each PC between brackets) of Kendall's soil data

Table VI.2 does not allow us to interpret the MCCs, in the same way that Table II.3 did not provide sufficient information for an adequate interpretation of the PCs. But it is perfectly legitimate to make a *comparison* of both sets of coefficients and note the greater or lesser role played by each variable in both sets of components.

Thus, the loading for $\mathbf{x}_1$ on the first MCC (0.357) is considerably smaller than its counterpart for the first PC (0.956). It will be recalled that the first PC was very strongly correlated with $\mathbf{x}_1$ (a correlation of 0.9963) and we will not be surprised to find a smaller correlation of $\mathbf{x}_1$ with the first MCC. Conversely, the role of $\mathbf{x}_2$ is much enhanced when moving from the first PC to the first MCC, as are, to a lesser extent, those of $\mathbf{x}_3$ and $\mathbf{x}_4$. The loadings for the remaining MCCs are, generally speaking, small for $\mathbf{x}_1$ and $\mathbf{x}_2$, but large for $\mathbf{x}_3$ and/or $\mathbf{x}_4$. However, the discussion in Chapter II should warn us against expecting these MCCs to be 'dominated' by the last two variables, all the more so since $\mathbf{x}_3$ and $\mathbf{x}_4$ are the lowest variance variables (see Table II.2 — indeed it is arguably *because* they are low-variance variables that they tend to have large loadings , via $\mathbf{D}_S^{-\frac{1}{2}}$).

Of greater interest than loadings is to gauge the performance of MCCs and PCs in terms of the criteria underlying both methods, in an illustration of the results from the first two subsections of Section VI.2.

For greater ease in interpretation, Table VI.3 gives the *relative* importance of the traces of $\mathbf{S}_{[q]}$, $\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_{[q]}\mathbf{D}_S^{\frac{1}{2}}$, $\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_{[q]}\mathbf{D}_S^{-\frac{1}{2}}$ and $\mathbf{R}_{[q]}$, for $q = 1, ..., p$. In other words, it assesses the performance of the first q PCs and MCCs, for all values of q , under both criteria, relative to the total variance or SSC which are to be accounted for. For the quantities relating to sums of squared multiple correlations, this involves dividing by $p = 4$. For the quantities involving fitted variance, it involves dividing by $\mathrm{tr}(\mathbf{S})$.

Component	(Relative) Fitted variance		Mean SSC	
	PCs	MCCs	PCs	MCCs
no. q	$\mathrm{tr}(\mathbf{S}_{[q]})/\mathrm{tr}(\mathbf{S})$	$\mathrm{tr}(\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_{[q]}\mathbf{D}_S^{\frac{1}{2}})/\mathrm{tr}(\mathbf{S})$	$\mathrm{tr}(\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_{[q]}\mathbf{D}_S^{-\frac{1}{2}})/\mathrm{p}$	$\mathrm{tr}(\mathbf{R}_{[q]})/\mathrm{p}$
1	0.9172	0.8345	0.3910	0.4189
2	0.9923	0.8823	0.5727	0.7055
3	0.9973	0.9047	0.7568	0.9455
4	1.0000	1.0000	1.0000	1.0000

Table VI.3: Relative performance of the first q ($q = 1, ..., 4$) PCs and MCCs of Kendall's soil data

The first and last columns have the optimum values for each specific criterion and are therefore never smaller than those in the adjacent columns.

For this particular data set, the differences between criteria are much more pronounced than those between each method for a given criterion. To a large extent this reflects the sizable difference in the pattern of cumulative relative eigenvalues of the covariance matrix (the first column of Table VI.3) and of the correlation matrix (the last column of Table VI.3).

But considering the individual variables and their relation (in terms of both criteria) to the MCCs and PCs will provide a fuller understanding of the comparative behaviour of both sets of components. Tables VI.4 to VI.7 provide this information.

	MCC			
Variable	ϕ_1	ϕ_2	ϕ_3	ϕ_4
$\mathbf{x}_1$	0.8453	0.0378	0.0207	0.0962
$\mathbf{x}_2$	0.8264	0.0667	0.0117	0.0952
$\mathbf{x}_3$	0.0011	0.8004	0.1719	0.0266
$\mathbf{x}_4$	0.0030	0.2412	0.7557	0.0001

Table VI.4: Squares of correlations between each variable and each MCC of Kendall's soil data

Table VI.4 gives the diagonal elements of $\mathbf{R}_j$ $(j = 1, ..., 4)$. These are the squared correlations of each variable with each MCC. Table VI.5 gives the corresponding values for PCs, *i.e.*, the diagonal elements of $\mathbf{D}_S^{-\frac{1}{2}}\mathbf{S}_j\mathbf{D}_S^{-\frac{1}{2}}$, $(j = 1, ..., 4)$. These are merely the squares of the values in Table II.4.

In both tables, the cumulative sums on each row, for the first q columns give the

Variable	PC			
	ξ_1	ξ_2	ξ_3	ξ_4
$\mathbf{x}_1$	0.9926	0.0074	0.0000	0.0000
$\mathbf{x}_2$	0.5407	0.4586	0.0007	0.0000
$\mathbf{x}_3$	0.0302	0.2608	0.7016	0.0074
$\mathbf{x}_4$	0.0006	0.0001	0.0339	0.9655

Table VI.5: Squares of correlations between each variable and each PC in Kendall's soil data

squared multiple correlations of each variable with the first q MCCs or PCs. The sums of all rows must be one.

The first MCC is almost orthogonal to $\mathbf{x}_3$ and $\mathbf{x}_4$ and forms an almost equal angle with both $\mathbf{x}_1$ and $\mathbf{x}_2$. It is, roughly speaking, an average of the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ when they are normalized to have equal length, although merely looking at the loadings in Table VI.2 would not tell us this. This provides yet another example (this time in the context of MCCA) of the problems discussed in Chapter II.

The 'strategy' which seems to have been followed in obtaining the first MCC is to ignore $\mathbf{x}_3$ and $\mathbf{x}_4$ and maximize the SSC by evenly sharing the correlations with both remaining variables. The first PC (which was obtained by a different criterion) resulted in a similar focus on $\mathbf{x}_1$ and $\mathbf{x}_2$, but by concentrating on variances rather than correlations ended up almost coinciding with the larger of the two vectors representing those variables in $\mathbb{R}^n$.

That both methods choose to practically ignore $\mathbf{x}_3$ and $\mathbf{x}_4$ in the first component is to some extent a coincidence. PCA ignores them because of their very low variances

(0.61 and 0.25) as compared with $\mathbf{x}_1$ and $\mathbf{x}_2$ (75.75 and 13.13). MCCA ignores them because they have very low squared correlations with $\mathbf{x}_1$ and $\mathbf{x}_2$, as well as among themselves (none of these squared correlations exceeds 0.05 - see Table II.2).

The latter point can best be elaborated upon by keeping in mind the geometry of what is happening. If all four variables were uncorrelated among themselves, *i.e.*, orthogonal in $\mathbb{R}^n$, then any linear combination would do as a first MCC since the sum of squares of its direction cosines on the orthonormal basis provided by the four variables would always be one. Let us assume, for argument's sake, that $\mathbf{x}_3$ and $\mathbf{x}_4$ remain exactly uncorrelated among themselves and with the first two variables. As $\mathbf{x}_1$ and $\mathbf{x}_2$ become correlated, it will pay to have the first MCC close to them, since its SSC will thus have two 'larger' terms in one go. In particular, say the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^n$ is θ, and consider a linear combination of those two vectors, whose angle with $\mathbf{x}_1$ is ψ (hence $\theta - \psi$ with $\mathbf{x}_2$). The sum of squared correlations of this linear combination with all four variables will be $f(\psi) = \cos^2(\psi) + \cos^2(\theta - \psi)$, since the linear combination will also be orthogonal to $\mathbf{x}_3$ and $\mathbf{x}_4$. A trivial maximization of $f(\psi)$ shows that the optimum choice for ψ is $\theta/2$. Thus, we should choose the bisector of the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$.

The above discussion gives a rough idea of how the first MCC for this data set was arrived at and also a better understanding of the rationale behind MCCs. Since MCCs are scale invariant, any re-scaling of the variables will leave them unchanged.

Not so with PCs. By suitably increasing the variances of $\mathbf{x}_3$ and $\mathbf{x}_4$ it will always be possible to make the first PC a vector which practically ignored $\mathbf{x}_1$ and $\mathbf{x}_2$, rather than $\mathbf{x}_3$ and $\mathbf{x}_4$.

Moving on to the second MCC we see that attention has been shifted to $\mathbf{x}_4$ and, in particular, $\mathbf{x}_3$. The third MCC is similar, but reversing the emphasis on each variable. The sum of each variable's squared correlations with each of the first three MCCs (*i.e.*, those variables' squared multiple correlations with the first three MCCs) exceeds, in all cases, 0.90.

On the other hand, the second PC is still accounting for some 'left-over correlation' with $\mathbf{x}_2$, as well as focusing on $\mathbf{x}_3$. The third PC is accounting for almost all the 'remaining correlation' with $\mathbf{x}_3$. But these first three PCs are all practically orthogonal to $\mathbf{x}_4$ which, in turn, is very strongly correlated to the fourth PC. This is a rather striking contrast to what happened with MCCs.

Let us now consider the behaviour of both sets of components with regards to the PCA criterion of fitted variance. Table VI.6 gives the diagonal elements of $\mathbf{D}_S^{\frac{1}{2}}\mathbf{R}_j\mathbf{D}_S^{\frac{1}{2}}$ and Table VI.7 those of $\mathbf{S}_j$, $(j = 1, ..., 4)$. The sums of each row are the variances of the corresponding variables. The *proportions* of each variable's variance which are accounted for by each component were given in Tables VI.4 and VI.5 (see Theorem VI.6).

It is clear from these Tables that the 'strategy' followed in determining the first

Variable	MCC			
	ϕ_1	ϕ_2	ϕ_3	ϕ_4
$\mathbf{x}_1$	64.0339	2.8654	1.5673	7.2846
$\mathbf{x}_2$	10.8493	0.8756	0.1542	1.2494
$\mathbf{x}_3$	0.0007	0.4893	0.1051	0.0163
$\mathbf{x}_4$	0.0007	0.0602	0.1885	0.0000

Table VI.6: Variance of each variable in Kendall's soil data that is accounted for by each MCC

Variable	PC			
	ξ_1	ξ_2	ξ_3	ξ_4
$\mathbf{x}_1$	75.1907	0.5590	0.0016	0.0000
$\mathbf{x}_2$	7.0990	6.0204	0.0090	0.0001
$\mathbf{x}_3$	0.0184	0.1594	0.4289	0.0045
$\mathbf{x}_4$	0.0001	0.0000	0.0084	0.2409

Table VI.7: Variance of each variable in Kendall's soil data that is accounted for by each PC

PC was to focus on fitting the variance of $\mathbf{x}_1$, where most of the variance is concentrated and, along the way, also fitting most of the variance of $\mathbf{x}_2$ which is somewhat correlated with $\mathbf{x}_1$. The fourth variable is all but ignored until the last PC since its variance is negligible in global terms. Roughly speaking, the second PC focused on accounting for the residual variance of $\mathbf{x}_2$ and in so doing also accounting for some of $\mathbf{x}_3$'s variance; the third PC accounts for $\mathbf{x}_3$'s residual variance and ξ_4 accounts for practically all – and only for – the variance of $\mathbf{x}_4$. The influence of each variable's variance on the PCs is therefore clearly visible.

The situation with MCCs is, in this respect, less clear, which comes as no surprise. But the way in which MCCs globally maximize the proportion of each variable's variance that is accounted for (Theorem VI.6 and Table VI.4) helps in understanding the logic behind the numbers in Table VI.6.

Despite the differences discussed above, the global similarity of PCs and MCCs, as measured by the similarity index in $\mathbb{R}^n$ of the original and standardized data matrices (equation (VI.2.3)) is a respectable 0.9530.

VI.4 An overview of MCCA

MCCA is a method akin to PCA in that we seek to replace p variables with a new set of p uncorrelated variables. The difference lies in the criteria used to obtain the new set of variables.

The Most Correlated Components (MCCs) are uncorrelated variables which optimize the following equivalent criteria:

i). successively maximize the sum of squared correlations between an MCC and all p original variables.

ii). successively maximize the sum of squared correlations between each of the p variables and its orthogonal projection onto the subspace spanned by the first q $(q = 1, ..., p)$ MCCs.

iii). successively maximize the sum of squares of the proportion of each variable's variance which is preserved by a regression onto the MCC.

In particular, the first MCC of a data set is the variable which is 'globally most correlated' with the p original variables.

It turns out that MCCs are merely the Correlation Matrix PCs of the original variables. But the above criteria have no need for a standardization of the variables. It is possible to work with, and interpret the MCCs in terms of, the original variables. Or any re-scaling of those original variables, since MCCs are scale-invariant.

The scale-invariance of MCCA is probably its greatest merit. In discussing Correlation Matrix PCA as a means of overcoming the scale dependence of PCs, Chatfield and Collins (1980) [8, (pg.71)] said that it "avoids rather than solves the scaling problem". We can now turn this around and say that MCCs (*i.e.*, Correlation Matrix PCs viewed as the optimal solutions to the above problems) solve rather than avoid the scaling problem. In Subsection I.5.3 we saw how the set of all $n \times p$ data matrices can be partitioned into equivalence classes by the equivalence relation $\mathbf{X}_1 \sim \mathbf{X}_2 \iff \mathbf{X}_2 = \mathbf{X}_1\mathbf{D}$ for some positive diagonal matrix $\mathbf{D}$. All matrices in any such equivalence class share common MCCs (but not common PCs).

MCCA also has a role to play in dimensionality reduction, although it will, in general, require more dimensions than PCA to achieve the same values for its own 'index of interestingness'.

The main draw-back of MCCA is that it is intrinsically an '$\mathbb{R}^n$ technique', whose parallel effects in $\mathbb{R}^p$ do not appear to have any optimal properties of their own.

Closely related to this drawback is the need to exclude data sets which are by nature n functions sampled at p points of their independent variable. For such data sets, linear combinations of the p 'variables' are usually meaningless.

But if MCCA is, unlike Covariance Matrix PCA, inappropriate for 'sampled functions' data sets, it is, again unlike Covariance Matrix PCA, very appropriate for data sets whose p variables are in different, or locally arbitrary, units of measurement.

A third type of n individuals *vs.* p variables data sets is that where the units of measurement for the p variables are the same and are only really meaningful when they are the same. For such data sets (an example of which was given in Chapter II – the French families' foodstuffs expenditures data set) either a Covariance Matrix PCA or an MCCA are conceivable. The choice between a Covariance or Correlation Matrix PCA for such data sets has often been based on whether all p variables should have an equal weight in the analysis or not, but it is now clear that a more fundamental decision on the criteria which are to optimized is implicitly being made. Both methods can be used simultaneously, producing two sets of uncorrelated components which are optimum for different criteria.

All things considered, MCCA deserves a place of its own in multivariate data analysis.

Chapter VII

Some ideas for future work

Previous Chapters have covered a host of different topics within the general context of descriptive PCA. The requirements of writing a PhD thesis inevitably imply that a line must be drawn somewhere, leaving many open questions and unexplored ideas. In this final Chapter we mention some of these topics for future research.

Interpreting PCs

Chapter II has many loose ends which require further consideration. In the first place, it would be helpful to carry out a much more extensive study of real or simulated data sets in order to assess the frequency of misleading interpretations which result from the standard way of reading PC loadings. The fact that the examples given in Chapter II (and a few others which were not presented here) were among the first to be scrutinized suggests that such erroneous interpretations (in particular of second and subsequent Covariance Matrix PCs, for reasons discussed in Section II.2) are fairly widespread. But a more precise assessment of the scale of the problem is called for. The less problematic case of Correlation Matrix PCs also deserves a more detailed analysis. A more thorough study of this kind should also focus on the role of corrected loadings. In Section II.3 it was noted that corrected loadings enabled us to write PCs as linear combinations of p vectors of equal size, thus removing the major source of trouble with the standard loadings. But problems remained and it was suggested that corrected loadings could prove most valuable as a starting point for algorithms seeking the 'best' subset of variables to use as regressors for the PC. Just how efficient this idea might prove to be is another open issue.

The use of such algorithms also leaves a number of unanswered questions. Will the specific nature of this regression problem (where the regressed variables – the PCs – are *exact* linear combinations of all p variables; where appropriate starting points – the variables with high magnitude corrected loadings – can be suggested; and where complementary information is available – in the spectral decompositions of the covariance and correlation matrices, whose significance was discussed in Chapter VI) mean that one or more of the existing algorithms are particularly well suited for our purposes? Or can it suggest new algorithms? In addition, can more be said concerning the thresholds for 'acceptability' of a given subset of variables? It is quite likely that the geometry of $\mathbb{R}^n$ will guide us towards answers to these questions.

Matrix Spaces

The geometry of the inner product linear spaces of matrices associated with PCA is surely one of the lines of future research which deserves greatest attention. In Section III.3 the geometry of the **PSD** cone in $\mathbb{S}_m$ was considered in some detail. It was shown how the eigenvalues and eigenvectors of **p.s.d.** matrices gave rise to interesting characterizations of the matrices' location in the cone. It is highly likely that more can be said in this respect and, with some luck, further results may prove more directly useful for practical applications of PCA and for multivariate analysis in general.

One question of interest is whether there is some meaningful geometric characterization of the locus of covariance (**p.s.d.**) matrices which share a common correlation matrix. A positive answer to this question could prove particularly useful for a fuller understanding of the relations between Covariance and Correlation Matrix PCAs.

Another relevant aspect is to what degree the characterizations of **p.s.d.** matrices can be paralleled in the more general matrix space $\mathbb{M}_{n\times p}$. The role played by the rank one symmetric matrices $\mathbf{xx}'$ in Section III.3 is, to some extent, reflected in the role of the rank one matrices $\mathbf{xy}' \in \mathbb{M}_{n\times p}$, for $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^p$ (see Eaton (1983) [12, (Section 1.5)] or Horn and Johnson (1991) [26, (Chapter 3)]) and the role of the eigenvalues of **p.s.d.** matrices is reflected in that of the singular values of any matrix. But it is only natural that the more general nature of the space $\mathbb{M}_{n\times p}$ will restrict the parallels between both sets of results.

Data transformations and projections

There is room for improvement in studying the effects of linear transformations of a data set. The effects of any particular transformation or class of transformations can be further scrutinized by elaborating on the results in Chapter IV. The role of the 'matrix of interactions' defined in Subsection IV.3.3 and/or of the matrix $\mathbf{Q}$ defined in Theorem IV.7 also deserves more attention, in particular, the comments following Theorem IV.7 concerning the bounds on the eigenvalues of the matrix $\mathbf{Q}$.

The special nature of projective transformations led to many pleasing results in Chapter V. Some of these results still need fine tuning. For example, the robustness of the result in point iv) of Theorem V.4 – discussed in the example of Subsection V.3.3 – should be further explored. The discussion which followed Theorem V.5, concerning the relation between eigenvectors of the original and the projected matrices whose associated eigenvalues were of very different sizes, is another example of an issue requiring a numerical analysis.

The new solution to the problem of removing isometric size from morphometric data will obviously require passing the test of useful practical applications.

One concept not mentioned so far but which might be useful in the study of linear transformations of the data is the **Kronecker product** of any two matrices. If $\mathbf{A} \in \mathbb{M}_{n\times p}$ and $\mathbf{B} \in \mathbb{M}_{m\times q}$, the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ is defined as the

$mn \times pq$ matrix which is best described as an '$n \times p$ block matrix', with each block of size $m \times q$ and such that the (i, j)-th block is given by $a_{ij}\mathbf{B}$ (Graham (1981) [20]).

It will be recalled that the *vec* operator (Section III.1) defined an isomorphism from $\mathbb{M}_{n \times p}$ to $\mathbb{R}^{np}$. If we vectorize the matrix resulting from a general linear transformation of both the columns and rows of a data matrix $\mathbf{Y} \in \mathbb{M}_{n \times p}$, we have $vec(\mathbf{TYU}') = (\mathbf{T} \otimes \mathbf{U})vec(\mathbf{Y})$. Thus, the effects of a general linear transformation of a data matrix may also be studied by looking at the effects in $\mathbb{R}^{np}$ of a matrix which is the Kronecker product of the row- and column-transforming matrices.

MCCA

The fact that the MCCs of a data set are merely its Correlation Matrix PCs implies, in a sense, that there is already an enormous mass of results and applications relating to MCCA. But these must be looked at with new eyes.

It should be noted that much in the theoretical results of Chapter VI emerged from the fact that the new criterion underlying MCCA implied a maximization of the trace of matrix (VI.1.8). Any other criterion which leads to an optimization of the trace of a matrix of the kind $\mathbf{A}'\mathbf{SW}(\mathbf{W}'\mathbf{SW})^{-1}\mathbf{W}'\mathbf{SA}$ (for some matrix $\mathbf{A}$) will lead to a similar solution, if we take $\mathbf{W} = \mathbf{AU}$ where $\mathbf{U}$ is the matrix of eigenvectors of matrix $\mathbf{A}'\mathbf{SA}$. The characterizations of this solution would, to a large extent, be analogues of those given in Theorems VI.2, VI.3 and VI.5. In other words (and as Meredith and Millsap stressed [43]), there are ready-made Component Analyses in waiting which 'only' need a meaningful criterion to be found for some choice of **A**.

Inference

The fact that inferential considerations sometimes obscure fundamental underlying properties of PCA does not imply that, once such properties are uncovered, we cannot return to an inferential context. On several occasions it was noted that some key results in this thesis (the use of multiple correlations and regressions when interpreting PCs; the essence of MCCA; most results involving vectors in $\mathbb{R}^p$, for example) are valid in an inferential context. It is likely that the geometry of inner product spaces defined by random variables (see Eaton (1983) [12]) may also enable us to obtain analogues of the results in Chapter III. Generalizing these results to an inferential context is therefore another major avenue of future research.

Meta-analyses

The geometry of matrix spaces implies that analogues of the Component Analyses, with matrices taking the part of variables in $\mathbb{R}^n$, are conceivable. For example, since the cosines of angles between matrices in$\mathbb{C}_{n\times p}$ are the global similarity indices for their PCAs, an 'MCCA criterion' of 'globally minimizing angles' between t such matrices and any linear combination of them will be optimized by the linear combination whose PCA is 'globally most similar' to those of the t matrices. The details and potential interest of such 'Meta-analyses' of sets of matrices also deserves attention.

Appendix A

Basic concepts in linear spaces

Linear (vector) spaces play a key role throughout this thesis. In this Appendix we briefly recall some well-known concepts defined for general linear spaces or in the specific context of the vector spaces $\mathbb{R}^m$. The general results are freely adapted from Halmos (1958) [22, (section 41)] and Goffman and Pedrick (1965) [18, (sections 1.1;2.11;2.21;4.1;4.2)], where the relevant proofs may be found. There is an underlying assumption that we are dealing with *real* and *finite-dimensional* linear spaces.

Definition A.1 *Let* $\mathcal{L}$ *be a linear space. A* **distance** *or* **metric** *in* $\mathcal{L}$ *is a function* $d : \mathcal{L} \times \mathcal{L} \rightarrow \mathbb{R}$ *such that:*

i). $d(x, y) \geq 0 \qquad \forall x, y \in \mathcal{L}$

with $d(x, y) = 0 \iff x = y$ [Positive definiteness]

ii). $d(x, y) = d(y, x) \qquad \forall x, y \in \mathcal{L}$ [Symmetry] *iii).* $d(x, z) \leq d(x, y) + d(y, z) \qquad \forall x, y, z \in \mathcal{L}$ [Triangle inequality]

Definition A.2 *Let* $\mathcal{L}$ *be a linear space. A* **norm** *in* $\mathcal{L}$ *is a function* $||\cdot|| : \mathcal{L} \rightarrow \mathrm{I\!R}$ *such that:*

i). $||\mathbf{x}|| \geq 0 \qquad \forall \mathbf{x} \in \mathcal{L}$

with $||\mathbf{x}|| = 0 \iff \mathbf{x} = \mathbf{0}$

ii). $||\alpha\mathbf{x}|| = |\alpha| \cdot ||\mathbf{x}|| \qquad \forall \mathbf{x} \in \mathcal{L}, \alpha \in \mathrm{I\!R}$

iii). $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}|| \qquad \mathbf{x}, \mathbf{y} \in \mathcal{L}$

Definition A.3 *Let* $\mathcal{L}$ *be a linear space. An* **inner product** *in* $\mathcal{L}$ *is a function* $< \cdot, \cdot >$: $\mathcal{L} \times \mathcal{L} \rightarrow \mathrm{I\!R}$ *such that:*

i). $< \mathbf{x}, \mathbf{y} >=< \mathbf{y}, \mathbf{x} > \qquad \forall \mathbf{x}, \mathbf{y} \in \mathcal{L}$ [Symmetry]

ii). $< \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} >= \alpha < \mathbf{x}, \mathbf{z} > +\beta < \mathbf{y}, \mathbf{z} >$

$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{L}, \forall \alpha, \beta \in \mathrm{I\!R}$ [Bilinearity]

iii). $< \mathbf{x}, \mathbf{x} >\geq 0 \qquad \forall \mathbf{x} \in \mathcal{L}$

with $< \mathbf{x}, \mathbf{x} >= 0 \iff \mathbf{x} = 0$ [Positive definiteness]

Any inner product in $\mathcal{L}$ **induces** a norm by taking:

$$||\mathbf{x}|| = \sqrt{< \mathbf{x}, \mathbf{x} >} \qquad \forall \mathbf{x} \in \mathcal{L}$$

In turn, any norm in $\mathcal{L}$ **induces** a distance by taking:

$$d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$$

Proposition A.4 (Cauchy-Schwarz inequality) *Let $\mathcal{L}$ be an inner product linear space. For every $\mathbf{x}, \mathbf{y} \in \mathcal{L}$ we have:*

$$|< \mathbf{x}, \mathbf{y} >| \leq ||\mathbf{x}|| \cdot ||\mathbf{y}||$$

where $|| \cdot ||$ is the norm induced by the inner product. The equality is attained if and only if $\mathbf{y} = \alpha \mathbf{x}$ for some scalar α.

Definition A.5 *The* **angle** *between two elements* $\mathbf{x}$,$\mathbf{y}$ *of an inner product linear space $\mathcal{L}$ is defined as:*

$$\sphericalangle(\mathbf{x}, \mathbf{y}) = \arccos\left(\frac{< \mathbf{x}, \mathbf{y} >}{||\mathbf{x}|| \cdot ||\mathbf{y}||}\right)$$

Definition A.6 *Let $\mathcal{L}$ be an inner product linear space. Two* **elements** $\mathbf{x}, \mathbf{y} \in \mathcal{L}$ *are said to be* **orthogonal** *if* $< \mathbf{x}, \mathbf{y} >= 0$. *Two* **subspaces** $\mathcal{M}, \mathcal{N} \in \mathcal{L}$ *are said to be* **orthogonal** *if any vector in one subspace is orthogonal to every vector in the other subspace, i.e., if* $< \mathbf{x}, \mathbf{y} >= 0$, $\forall \mathbf{x} \in \mathcal{M}$ *and* $\forall \mathbf{y} \in \mathcal{N}$.

Definition A.7 *Let $\mathcal{L}$ be an inner product linear space, and $\mathcal{M}$ a subspace of $\mathcal{L}$. We define the* **orthogonal complement** *(or* ortho-complement*) of $\mathcal{M}$ to be the set $\mathcal{M}^{\perp}$ of all elements of $\mathcal{L}$ which are orthogonal to every element of $\mathcal{M}$. Formally:*

$$\mathcal{M}^{\perp} = \{\mathbf{y} \in \mathcal{L} :< \mathbf{y}, \mathbf{x} >= 0 \quad \forall \mathbf{x} \in \mathcal{M}\}$$

Proposition A.8 *The orthogonal complement of a subspace is itself a subspace.*

Definition A.9 *Let $\mathcal{L}$ be a linear space and $\mathbf{P} : \mathcal{L} \rightarrow \mathcal{L}$ a linear transformation on $\mathcal{L}$.* $\mathbf{P}$*is said to be* **idempotent** *if* $\mathbf{P}^2 = \mathbf{P}$.

Definition A.10 *A linear space $\mathcal{L}$ is the* **direct sum** *of two of its subspaces $\mathcal{M}$ and $\mathcal{N}$ if every element $\mathbf{x} \in \mathcal{L}$ has a unique decomposition $\mathbf{x} = \mathbf{y} + \mathbf{z}$ with $\mathbf{y} \in \mathcal{M}$ and $\mathbf{z} \in \mathcal{N}$. We write $\mathcal{L} = \mathcal{M} \oplus \mathcal{N}$.*

Definition A.11 *If a linear space $\mathcal{L}$ is the direct sum of subspaces $\mathcal{M}$ and $\mathcal{N}$, the* **projection on $\mathcal{M}$ along $\mathcal{N}$** *is defined to be the transformation $\mathbf{P} : \mathcal{L} \rightarrow \mathcal{M}$ such that for every element $\mathbf{x} \in \mathcal{L}$ with decomposition $\mathbf{x} = \mathbf{y} + \mathbf{z}$ $(\mathbf{y} \in \mathcal{M}, \mathbf{z} \in \mathcal{N})$, we have $\mathbf{Px} = \mathbf{y}$.*

Proposition A.12 *Let $\mathcal{L}$ be a linear space and $\mathcal{M}$ and $\mathcal{N}$ two subspaces such that $\mathcal{L} = \mathcal{M} \oplus \mathcal{N}$. Let $\mathbf{P}$ be the projection on $\mathcal{M}$ along $\mathcal{N}$. Let $\mathbf{I}_d : \mathcal{L} \rightarrow \mathcal{L}$ be the identity transformation on $\mathcal{L}$, which maps every element of $\mathcal{L}$ onto itself. Then:*

i). $\mathcal{M} \cap \mathcal{N} = \{\mathbf{0}\}$

ii). The **range** *of $\mathbf{P}$ (that is, the set of all images under $\mathbf{P}$) is $\mathcal{M}$.*

iii). The **nullspace** *of $\mathbf{P}$ (that is, the set of all solutions to the equation $\mathbf{Px} = \mathbf{0}$) is $\mathcal{N}$.*

iv). $\mathbf{I}_d - \mathbf{P}$ is the projection on $\mathcal{N}$ along $\mathcal{M}$.

Proposition A.13 *A linear transformation $\mathbf{P} : \mathcal{L} \rightarrow \mathcal{L}$ is a projection on some subspace of the linear space $\mathcal{L}$ if and only if it is idempotent.*

Specifying one of the subspaces, say $\mathcal{M}$, in Proposition A.12 is not enough to completely characterize the other subspace, *i.e.*, $\mathcal{N}$, in the direct sum. But for inner product linear spaces, there is one special case among these alternative direct summands which deserves special attention.

Proposition A.14 (Orthogonal Decomposition Theorem) *Let $\mathcal{L}$ be an inner product linear space and $\mathcal{M}$ a subspace of $\mathcal{L}$. $\mathcal{L}$ is the direct sum of $\mathcal{M}$ and its orthogonal complement $\mathcal{M}^{\perp}$, that is, $\mathcal{L} = \mathcal{M} \oplus \mathcal{M}^{\perp}$.*

Definition A.15 *Let $\mathcal{L}$ be an inner product linear space and $\mathcal{M}$ a subspace of $\mathcal{L}$. The projection on $\mathcal{M}$ along $\mathcal{M}^{\perp}$ is called the* **orthogonal projection on** *$\mathcal{M}$.*

Proposition A.16 (Orthogonal Projection Theorem) *Let $\mathcal{L}$ be an inner product linear space, $\mathcal{M}$ a subspace of $\mathcal{L}$, $\mathbf{I}_d$ the identity transformation on $\mathcal{L}$ and $\mathbf{P}$ the orthogonal projection on $\mathcal{M}$. Every element $\mathbf{x} \in \mathcal{L}$ is uniquely decomposable as $\mathbf{x} = \mathbf{Px} + (\mathbf{I}_d - \mathbf{P})\mathbf{x}$. $\mathbf{Px}$ is the element of $\mathcal{M}$ which is at a minimum distance (in the metric induced by the inner product on the linear space) to $\mathbf{x}$.*

The above results apply to general linear spaces or general inner product linear spaces. For the specific case when the linear spaces are the usual vector spaces $\mathbb{R}^m$, with the conventional Euclidean inner product ($<\mathbf{x},\mathbf{y}>= \mathbf{x}'\mathbf{y}, \ \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^m$), additional results can be obtained. Linear transformations on such spaces are represented by $m \times m$ matrices. Linear transformations from one space, $\mathbb{R}^k$, to another, $\mathbb{R}^m$, are represented by $m \times k$ matrices.

Definition A.17 *Let $\mathbf{A}$ be an* m×k *matrix. The subspace of $\mathbb{R}^m$ which is spanned by the* k *columns of $\mathbf{A}$ is called the* **rangespace** *of the matrix $\mathbf{A}$, and is denoted by $\Re(\mathbf{A})$. The set of all vectors of $\mathbf{x} \in \mathbb{R}^k$ such that $\mathbf{Ax} = \mathbf{0}$ is called the* **nullspace** *of the matrix $\mathbf{A}$, and denoted by $\mathcal{N}(\mathbf{A})$. The maximum number of linearly independent columns of $\mathbf{A}$ is called the* **rank** *of the matrix $\mathbf{A}$.*

Proposition A.18 *Let* $\mathbf{A}$ *be an* m×k *matrix of rank* r . *Then:*

i). $\dim(\Re(\mathbf{A})) = r$

ii). $\dim(\mathcal{N}(\mathbf{A})) = k\text{-}r$

iii). $\mathcal{N}(\mathbf{A})$ *is a linear subspace of* $\mathrm{I\!R}^k$.

iv). $\Re(\mathbf{A}) = \mathcal{N}(\mathbf{A}')^{\perp}$

v). $\mathcal{N}(\mathbf{A}) = \Re(\mathbf{A}')^{\perp}$

Strang (1988) [63] refers to the last two results in the previous Proposition as the **Fundamental Theorem of Linear Algebra.**

Idempotent linear transformations, described in Definition A.9, are thus represented by $m \times m$ matrices $\mathbf{P}$, for which $\mathbf{P}^2 = \mathbf{P}$. Such matrices are called **idempotent matrices**. Proposition A.13 therefore tells us that $\mathbf{P}$ is an idempotent matrix if and only if it represents a projection onto some subspace of $\mathrm{I\!R}^m$.

The following results for the linear spaces $\mathrm{I\!R}^m$ are loosely adapted from Basilevski (1983) [4, (sections 4.8 and 5.3.4)].

Proposition A.19 *Let* $\mathbf{P}$ *be an* m×m *idempotent matrix. Then* $\mathbf{P}$ *can be written as* $\mathbf{P} = \mathbf{A}(\mathbf{B}'\mathbf{A})^{-1}\mathbf{B}'$ *for some* m×k *matrices* $\mathbf{A}$ *and* $\mathbf{B}$, *whose* k *columns are linearly independent. Furthermore:*

i). The nullspace of matrix $\mathbf{P}$ *is the orthogonal complement of the subspace spanned by the* k *columns of* $\mathbf{B}$, *that is,* $\Re(\mathbf{B})^{\perp}$. *Its dimension is* $m - k$.

ii). The range of matrix $\mathbf{P}$ *is the subspace spanned by the* k *columns of matrix* $\mathbf{A}$, *i.e.,* $\Re(\mathbf{A})$. *Its dimension is* k.

iii). $\mathbf{P}$ *is a projector on* $\Re(\mathbf{A})$ *along* $\Re(\mathbf{B})^{\perp} = \mathcal{N}(\mathbf{B}')$. *This projection matrix is unique, although matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *in its decomposition can be replaced by other matrices whose* k *columns are bases for the subspaces* $\Re(\mathbf{A})$ *and* $\mathcal{N}(\mathbf{B}')$.

iv). $\mathbf{P}'$ *is a projector on* $\Re(\mathbf{B})$ *along* $\mathcal{N}(\mathbf{A}')$. $\mathbf{I}_m - \mathbf{P}$ *is a projector on* $\mathcal{N}(\mathbf{B}')$ *along* $\Re(\mathbf{A})$.

v). $\mathbf{P}$ *has* k *eigenvalues equal to one and* $m - k$ *eigenvalues equal to zero. Any vector in* $\Re(\mathbf{A})$ *is an eigenvector of* $\mathbf{P}$ *with associated eigenvalue 1. Any vector in* $\Re(\mathbf{B})^{\perp}$ *is an eigenvector of* $\mathbf{P}$ *with associated eigenvalue zero.*

vi). The trace and rank of $\mathbf{P}$ *are equal, both being* k .

vii). $\mathbf{P}$ *is an orthogonal projection matrix if and only if* $\mathbf{P}$ *is symmetric.*

viii). Any m×m *orthogonal projection matrix* $\mathbf{P}$ *can be written as* $\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ *for some* m×k *matrix* $\mathbf{A}$ *whose columns are linearly independent. The columns of* $\mathbf{A}$ *are a basis for the subspace (of dimension* k *) onto which* $\mathbf{P}$ *projects.*

The results in Proposition A.19 are used extensively throughout the thesis.

Appendix B

100 km race data

This Appendix lists the full data set used in Subsection V.3.3. The 80×10 data set gives the times taken, by each of the eighty runners who completed the 1984 Lincoln 100 km race, to cover the ten successive 10-km stretches in the race. All times are in minutes. The runners are numbered according to their ranking at arrival.

No.	10k	20k	30k	40k	50k	60k	70k	80k	90k	100k
1	37.0	37.8	36.6	39.6	41.0	41.0	41.3	45.7	45.1	43.1
2	39.5	42.2	40.0	42.3	40.6	40.8	42.0	43.7	41.0	43.9
3	37.0	37.8	36.6	39.6	41.0	44.8	44.5	49.4	44.6	47.7
4	37.1	38.0	37.7	42.4	41.6	43.5	48.7	49.7	44.8	47.0
5	42.2	44.5	41.9	43.4	43.0	47.2	49.1	49.9	46.8	52.3
6	43.0	44.6	41.2	42.1	42.5	46.8	47.5	55.8	56.6	58.6
7	43.2	44.4	41.0	43.4	43.0	47.2	52.4	57.3	54.4	53.5
8	43.2	46.7	44.8	47.5	47.4	47.7	49.9	52.1	50.7	50.0
9	38.5	41.4	40.1	43.2	43.2	51.5	56.7	71.5	56.2	48.2
10	42.5	43.1	40.6	44.5	45.4	52.3	59.7	59.3	55.0	49.6

No.	10k	20k	30k	40k	50k	60k	70k	80k	90k	100k
11	38.0	40.1	39.1	43.8	46.6	51.9	59.2	63.5	57.6	58.4
12	46.0	50.4	46.8	47.4	44.1	43.4	46.3	55.0	64.9	56.2
13	44.8	46.0	43.1	46.5	46.3	49.0	52.5	58.4	60.9	55.2
14	44.8	46.0	43.1	46.5	46.3	49.0	52.5	58.4	60.9	55.2
15	47.0	49.4	46.8	48.6	47.8	50.8	50.3	54.0	54.4	53.6
16	45.0	46.7	45.3	49.9	47.8	51.2	54.1	58.7	53.3	50.7
17	45.0	46.7	43.8	48.0	47.2	47.5	51.7	57.3	60.4	55.6
18	43.1	44.5	41.0	42.5	40.6	42.8	46.5	73.2	70.8	63.4
19	45.2	46.9	45.5	48.8	50.1	51.2	56.4	55.2	56.6	53.5
20	43.0	46.1	44.7	47.4	47.1	46.8	54.6	60.4	68.0	51.6
21	38.3	41.6	39.6	40.7	41.6	41.6	47.2	62.4	82.7	77.1
22	45.0	47.1	45.3	49.1	46.8	47.4	50.3	55.1	66.4	64.6
23	43.2	46.1	45.2	48.4	49.9	49.6	52.7	58.1	62.8	62.6
24	41.2	44.6	43.8	48.4	48.8	53.4	58.9	68.6	59.1	53.4
25	49.2	48.8	48.7	51.8	48.2	52.8	50.2	58.0	58.7	57.5
26	48.0	52.9	49.6	50.1	48.1	48.1	49.1	54.6	62.7	64.0
27	46.0	49.9	47.7	50.4	52.9	51.4	55.6	57.8	59.7	55.8
28	46.1	46.0	42.2	44.4	46.0	49.0	53.3	66.7	72.9	67.6
29	48.0	52.9	49.6	50.1	48.4	50.0	58.5	62.9	60.1	60.1
30	45.1	49.7	46.5	46.5	49.3	58.8	58.7	64.7	64.0	63.4
31	49.2	54.5	51.3	56.1	53.9	53.2	53.4	58.8	62.4	59.4
32	47.0	49.4	46.8	49.7	50.3	55.5	59.8	67.1	64.2	70.4
33	48.2	54.1	51.2	53.5	54.8	55.7	55.2	65.7	62.3	62.3
34	46.5	50.8	48.0	51.4	50.0	58.6	61.6	61.5	61.9	75.4
35	47.3	51.2	49.5	52.6	57.9	58.6	66.4	70.6	56.4	55.6
36	48.2	53.9	50.9	54.0	52.4	59.3	77.5	60.6	55.8	61.4
37	43.3	45.2	42.7	44.9	47.3	52.9	69.3	92.2	57.3	79.1
38	52.0	53.0	50.0	51.6	55.4	56.3	56.7	68.4	66.9	65.4
39	49.2	54.5	50.8	53.6	53.4	56.0	62.3	65.8	66.1	65.6
40	49.3	52.8	51.1	53.8	52.4	59.3	63.2	73.7	62.3	62.3
41	47.2	51.3	49.5	52.6	51.6	60.1	64.4	66.5	66.6	76.9
42	49.2	48.8	49.2	54.2	60.8	60.4	64.0	69.9	66.1	65.1
43	45.2	50.2	48.7	53.6	53.5	60.3	59.2	71.4	75.9	71.8
44	45.3	51.0	46.9	50.0	51.0	59.7	78.2	68.9	69.7	72.8
45	49.2	48.3	46.2	51.6	51.9	61.1	71.8	74.6	70.3	69.9

No.	10k	20k	30k	40k	50k	60k	70k	80k	90k	100k
46	49.2	48.7	48.8	52.6	57.9	65.3	71.7	64.1	70.7	68.2
47	46.0	49.2	47.5	51.5	54.1	61.4	66.5	76.5	77.0	68.6
48	46.5	51.1	49.9	56.1	53.6	58.2	66.3	76.2	70.6	76.3
49	51.6	54.0	52.1	55.1	57.4	61.0	63.4	70.2	73.4	67.9
51	45.0	50.3	48.8	53.6	54.4	58.9	67.6	77.7	79.9	81.1
52	48.0	52.9	49.6	50.1	53.5	65.6	72.8	74.1	72.6	78.1
53	53.2	55.1	55.0	59.3	59.4	63.2	66.1	66.7	73.7	68.3
54	62.5	67.5	73.1	68.2	47.1	51.9	58.3	68.5	64.4	62.1
55	49.2	48.8	53.4	56.1	59.8	65.2	72.8	71.4	79.8	70.5
56	49.2	48.6	47.5	51.8	57.7	63.5	63.5	69.5	92.6	83.4
57	51.6	53.7	49.2	58.3	56.4	65.3	74.8	75.4	75.8	69.2
58	58.7	62.7	56.3	58.6	66.3	62.9	67.4	71.4	69.6	60.8
59	49.2	53.3	53.7	54.8	59.3	73.9	70.8	86.3	61.8	78.7
60	59.0	64.6	61.4	64.0	60.2	64.0	66.2	69.5	69.3	64.9
61	50.1	53.9	52.7	59.8	58.2	71.4	72.3	78.4	77.5	74.9
62	55.0	58.5	59.4	63.4	57.0	66.4	67.7	68.7	75.9	77.7
63	47.2	52.1	51.7	61.0	73.2	74.5	69.2	76.5	75.6	70.9
64	51.7	54.0	53.0	55.6	56.0	62.9	76.2	81.1	85.5	87.8
65	50.0	47.3	44.1	51.7	62.8	75.3	78.1	81.2	85.5	87.8
66	56.2	59.7	55.6	58.2	64.4	76.1	68.4	75.3	84.8	70.5
67	56.2	59.7	55.6	58.2	64.4	76.1	68.4	75.3	84.8	70.5
68	46.5	51.7	52.3	61.7	66.8	68.1	76.9	74.9	83.7	96.1
69	56.2	60.0	60.4	67.7	64.7	73.1	68.7	72.1	70.3	86.9
70	49.2	53.0	52.5	55.5	57.1	77.7	86.6	71.2	82.5	104.6
71	51.6	54.2	58.7	59.8	65.4	76.2	73.1	93.0	83.7	74.3
72	58.1	62.0	60.2	63.7	65.7	78.8	69.4	81.7	79.2	81.9
73	48.2	54.0	52.5	55.5	60.1	73.9	71.8	86.8	91.9	110.0
74	55.0	60.9	55.0	63.5	68.8	84.6	64.8	95.0	81.8	75.4
75	56.2	60.4	63.1	65.0	71.7	78.8	77.0	89.0	83.5	61.0
76	48.0	52.9	52.8	70.5	77.1	85.3	76.9	93.9	88.4	68.2
77	46.5	51.0	63.6	66.7	75.0	81.0	76.0	95.4	80.9	79.3
78	46.5	51.0	63.6	66.7	75.0	81.0	76.0	95.4	80.9	79.3
79	52.2	55.5	55.9	70.6	77.7	86.6	71.6	86.9	87.8	71.2
80	50.5	55.4	64.1	66.3	75.6	86.6	71.6	87.3	89.2	73.4

Bibliography

[1] T.W. Anderson. *Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, 1958.

[2] A. Antoniadis, J. Charre, S. Degerine, G. Gregoire, A. LeBreton, and S. Martin. Modeles d'analyse de la variance à deux facteurs pour l'etude des precipitations sur un reseau de stations. Research document no. 584, Informatique et Mathématiques Appliquées de Grenoble (IMAG), 1986.

[3] G.P. Barker. Theory of cones. *Linear Algebra and its Applications*, 39:263–291, 1981.

[4] A. Basilevski. *Applied Matrix Algebra in the Statistical Sciences.* North-Holland, 1983.

[5] P. Besse and J.O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51(2):285–311, 1986.

[6] R.G. Brereton. *Chemometrics: applications of Mathematics and Statistics to Laboratory Systems.* Ellis Horwood Limited, 1990.

[7] F. Cailliez and J.-P. Pages. *Introduction à l'Analyse des Données.* Société de Mathématiques Appliquées et de Sciences Humaines, Paris, 1976.

[8] C. Chatfield and A.J. Collins. *Introduction to Multivariate Analysis.* Chapman and Hall, 1980.

[9] F. Critchley. Optimal norm characterisations of multidimensional scaling methods and some related data analysis problems. In E. Diday et. al., editor, *Data Analysis and Informatics*, pages 209–229. North-Holland, 1980.

[10] W.E. Deming. *Some theory of sampling.* John Wiley & Sons, 1950.

[11] N.R. Draper and H. Smith. *Applied Regression Analysis.* John Wiley & Sons, second edition, 1981.

[12] M.L. Eaton. *Multivariate Statistics - A vector space approach.* John Wiley & Sons, 1983.

[13] Y. Escoufier. The duality diagram: a means for better practical applications. In P. and L. Legendre, editors, *Developments in Numerical Ecology.* Springer-Verlag, Berlin-Heideberg, 1987.

[14] B. Everitt. *Cluster Analysis.* Halsted Press, second edition, 1980.

[15] B. Flury. *Common Principal Components and related Multivariate Models.* John Wiley & Sons, 1988.

[16] I.M. Gelfand. *Lectures on Linear Algebra.* Interscience Publisher, N.Y., 1961.

[17] R. Gnanadesikan. *Methods for Statistical Data Analysis of Multivariate Observations.* John Wiley & Sons, 1977.

[18] C. Goffman and G. Pedrick. *First course in Functional Analysis.* Prentice-Hall, 1965.

[19] J.C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3 and 4):325–338, 1966.

[20] A. Graham. *Kronecker products and Matrix Calculus with Applications.* Ellis Horwood Ltd, Chichester, 1981.

[21] M.J. Greenacre. *Theory and Applications of Correspondence Analysis.* Academic Press, 1984.

[22] P.R. Halmos. *Finite-dimensional vector spaces.* D. Van Nostrand Company, 1958.

[23] R.D. Hill and S.R. Waters. On the cone of positive semidefinite matrices. *Linear Algebra and its Applications*, 90:81–88, 1987.

[24] R. Hindley. Pacing. Personal communication, 1984.

[25] R. Horn and C. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[26] R. Horn and C. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.

[27] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441 and 498–520, 1933.

[28] J. Huxley. *Problems of relative growth.* Dover Publications, Inc., second edition, 1972.

[29] J.E. Jackson. *A user's guide to principal components.* John Wiley & Sons, 1991.

[30] J.N.R. Jeffers. Two case studies in the application of principal component analysis. *Applied Statistics*, 16:225–236, 1967.

[31] P. Jolicoeur. The multivariate generalization of the allometry equation. *Biometrics*, 19:497–499, 1963.

[32] I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, 1986.

[33] M.C. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society*, 150:1–36, 1987.

[34] J.-B. Kazmierczak. Analyse logarithmique: deux examples d'application. *Revue de Statistique Appliquée*, 33:13–24, 1985.

[35] M. Kendall. *Multivariate Analysis.* Charles Griffin & Co., second edition, 1980.

[36] M. Kendall, A. Stuart, and J.K. Ord. *Kendall's Advanced Theory of Statistics,* volume 1. Charles Griffin & Co., fifth edition, 1987.

[37] M.G. Kendall. *A course in the geometry of n dimensions.* Charles Griffin & Company, 1961.

[38] W. Krzanowski. *Principles of multivariate analysis: a user's perspective.* Clarendon Press, Oxford, 1988.

[39] S.R. Lay. *Convex sets and their applications.* John Wiley & Sons, 1982.

[40] L. Lebart, A. Morineau, and J.-P. Fénelon. *Traitement des données statistiques.* Dunod, 1982.

[41] L. Lebart, A. Morineau, and K.M. Warwick. *Multivariate Descriptive Statistical Analysis.* John Wiley & Sons, N.Y., 1984.

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

[43] W. Meredith and R.E. Millsap. On component analyses. *Psychometrika,* 50:495–507, 1985.

[44] P.L. Meyer. *Introductory Probability and Statistical Applications.* Addison-Wesley, 1970.

[45] H. Minc. *Nonnegative matrices.* John Wiley & Sons, N.Y., 1988.

[46] I. Noy-Meir. Data transformations in ecological ordination. (i).some advantages of non-centering. *The Journal of Ecology*, 61:329–341, 1973.

[47] M. Okamoto. Optimality of principal components. In P.R. Krishnaiah, editor, *Multivariate Analysis II*, pages 673–685. Academic Press, N.Y. and London, 1969.

[48] K. Pearson. On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, 6(11):559–572, 1901.

[49] M.C. Pease. *Methods of Matrix Algebra.* Academic Press, 1965.

[50] R.W. Preisendorfer. *Principal Component Analysis in Meteorology and Oceanography.* Elsevier, Amsterdam, 1988.

[51] J.O. Ramsay. When the data are functions. *Psychometrika*, 47:379–396, 1982.

[52] J.O. Ramsay, J. ten Berge, and G.P.H. Styan. Matrix correlation. *Psychometrika*, 49(3):403–423, 1984.

[53] C. Ranatunga. Methods of removing 'size' from a data set. Msc thesis in statistics, University of Kent at Canterbury, 1989.

[54] C.R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhyā*, 26:329–358, 1964.

[55] C.R. Rao. *Linear Statistical Inference and its Applications.* John Wiley & Sons, second edition, 1973.

[56] C.R. Rao. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In P.R. Krishnaiah, editor, *Multivariate Analysis V*, pages 3–22. North-Holland, 1980.

[57] P. Robert and Y. Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Applied Statistics*, 25(3):257–265, 1976.

[58] R. Sabatier, J.-D. Lebreton, and D. Chessel. Principal component analysis with instrumental variables as a tool for modelling compositional data. In R. Coppi and S. Bolasco, editors, *Multiway Data Analysis*, pages 341–352. Elsevier Science Publishers B.V. (North-Holland), 1989.

[59] S.R. Searle. *Matrix Algebra useful for Statistics.* John Wiley & Sons, 1982.

[60] G.A.F. Seber. *Multivariate Observations.* John Wiley & Sons, 1984.

[61] G.E. Shilov. *An Introduction to the Theory of Linear Spaces.* Prentice-Hall, 1961.

[62] K.M. Somers. Allometry, isometry and shape in principal component analysis. *Systematic Zoology*, 38:169–173, 1989.

[63] G. Strang. *Linear Algebra and its Applications*. Harcourt Brace Jovanovich, third edition, 1988.

[64] Y. Takane and T. Shibayama. Principal component analysis with external information on both subjects and variables. *Psychometrika*, 56:97–120, 1991.

[65] P. Tarazaga. Eigenvalue estimates for symmetric matrices. *Linear Algebra and its Applications*, 135:171–179, 1990.

[66] G. Teissier. Relative growth. In T.H. Waterman, editor, *The Physiology of Crustacea, Vol. 1: Metabolism and Growth*. Academic Press, London, 1960.

[67] M. Volle. *Analyse des donnèes*. Economica, 1981.

[68] S. Winsberg. Two techniques: Monotone spline transformations for dimensional reduction in pca and easy-to-generate metrics for pca of sampled functions. In J.A. Van Rijckevorsel and J. De Leeuw, editors, *Component and Correspondence Analysis*, pages 129–135. John Wiley & Sons, 1988.

[69] S. Winsberg and J. Kruskal. Easy to generate metrics for use with sampled functions. In *COMPSTAT 1986*, pages 55–60. Physica-Verlag, Heidelberg for IASC (International Association for Statistical Computing), 1986.