MULTIDIMENSIONAL ANALYSIS OF GROUPED VARIABLES: AN INTEGRATED APPROACH

PROEFSCHRIFT

ter verkrijging van de graad van Doctor aan de Rijksuniversiteit te Leiden, op gezag van de Rector Magnificus Dr. L. Leertouwer, hoogleraar in de faculteit der Godgeleerdheid, volgens besluit van het college van dekanen te verdedigen op donderdag 18 maart 1993 te klokke 16.15 uur

door

Andreas Franciscus Maria Nierop

geboren te Leiden in 1954

Promotor: prof. dr. W.J. Heiser

Referent: prof. dr. J.P. van de Geer

Overige leden: prof. dr. J.M.F. ten Berge (Rijksuniversiteit Groningen)
dr. C.J.F. ter Braak (Dienst Landbouwkundig Onderzoek)
prof. dr. J.C. van Houwelingen
prof. dr. L.J.Th. van der Kamp

Dedicated to Gesshin Prabhasa Dharma Roshi

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Nierop, Andreas F.M.

Multidimensional analysis of grouped variables: An
integrated approach / Andreas F.M. Nierop. - [S.l. :
s.n.]. - Ill.
Thesis Leiden. - With ref.
ISBN 90-9005710-2
Subject headings: multivariate analysis / partial least
squares / generalized canonical correlation analysis.

© Cover photo: Annelies Hendriks: 'Lichen in Torcal'
Printed by 'Reprodienst Faculteit Sociale Wetenschappen,
Rijksuniversiteit Leiden' and by
Printing office 'Karstens drukkers b.v., Leiden'

ISBN 90-9005710-2

Acknowledgments

I like to express my gratitude to Geert de Soete for kindly supplying the Miller Nicely data. I am indebted to the Centre for Structure Elucidation and Instrumental Analysis of TNO Nutrition and Food Research for granting permission to analyse mass spectrometric data from barley tissue. Special thanks to Albert Tas and other colleagues from TNO for encouraging support during the last months.

Last but not least thanks to my wife Annelies Hendriks. Without her this monograph would not have been crystallized on paper.

Contents

Chapter 1

INTEGRATION OF DIVERGENT AIMS IN MULTIDIMENSIONAL ANALYSIS

The predictive value of multiset multivariate methods is related to the optimal integration of two criteria: stability and exactness. Based on strategies to combine stable with exact prediction, we introduce a classification of hybrid and adjusted multivariate methods. Some considerations on mathematical tools and presentation are added. An outline of the structure of this monograph is provided.

Introduction

Prediction is an important goal in multivariate analysis. We presume that the researcher has optimized measurements in such a way that they are representative and reliable for the samples to be studied. Within this scope we pursue the formulation of prediction methods that are *stable* and at the same time *exact*. Therefore we have to find some optimal integration of these two more or less divergent aims. In this monograph an approach is preferred in which stable and exact predictions are integrated in a theoretically simple way, because that is believed to offer possibilities for arriving at better predictions. *Theoretically simple* implies that the integrated methods must be characterized by the fact that they maximize (or minimize) *one* scalar function. (Without loss of generality the alternative of *minimizing* will subsequently be omitted.) In addition, the contributions of the building blocks of the fit function that define stability or exactness should not be balanced by user defined weights or weights determined by cross-validation. Such weighted fit functions introduce too many additional degrees of freedom to be called simple here.

Most existing methods, which will be classified as hybrid methods are not simple integrated. Sequential and cyclic hybrid methods do not maximize one fit function. *Sequential* hybrid methods maximize several fit functions successively, while utilizing optimal parameters of previously fitted models. An example is Principal Component-Discriminant Analysis (Hoogerbrugge, Willig and Kistemaker, 1983), where the predictor variables are first decomposed into principal components, after which only a few dominant components are used in a discriminant analysis. *Cyclic* hybrid methods maximize several fit functions cyclically, while utilizing optimal parameters of previously fitted models, until a stationary phase is reached. An example is "Soft Modelling" introduced and advocated by Herman Wold (1982). This type of modelling is often called "Partial Least Squares" or "PLS", which refers to the partitioning of parameters in estimable subsets. Apart from sequential and cyclic methods, many additive and multiplicative hybrid methods maximize one fit function, but incorporate (exponential) weights to balance stability and exactness of prediction. These weights are user defined or optimized by cross-validation. An example of an additive hybrid method with a balancing weight is Ridge Regression (Hoerl & Kennard, 1970).

In this monograph we pursue the construction of simple integrated methods. We apply two guide-lines to achieve our goal. The first guide-line is to specialize stability and exactness as much as possible in independent subfunctions, for example local versus global functions. A local function is non zero only at a limited range of its argument, while a global function maybe non zero everywhere. The second guide-line is to incorporate special constraints for improving stability or exactness. Methods with such constraints we call *adjusted methods*.

The working field of our simple integrated approach is the area of multiset analysis. In multiset analysis we consider information from different sources collected in two or more sets, where each set contains one or more variables. Sometimes the information is available as a number of matrices with similarities or dissimilarities. Apart from the construction of simple integrated methods, we aim at the description of multiset methods in a comprehensive system, using new theoretical concepts such as *filters*, *reflected variable*, *directed correlation* and *secondary prediction*.

In section 1.1 we relate stability and exactness of prediction with corresponding multivariate criteria, which are respectively *set variance* and *set correlation*. In the context of optimizing stability and exactness of prediction we provide in section 1.2 more examples of hybrid methods and elaborate on the difference with adjusted methods. In section 1.3 we introduce matrix algebra as a multidimensional tool for condensing large amounts of information. In section 1.4 we make some comments on the relation between models and scalar fit functions and section 1.5 offers an overview of the whole monograph.

1.1 Set variance and set correlation

In the multiset context we define multivariate prediction always between latent variables corresponding to different sets of observed variables. Generally each latent variable will be a linear combination of observed variables, so that they are not latent in the sense that we cannot actually compute them, as is the case in certain branches of factor analysis and LISREL modelling. The term latent only refers to the fact that these variables are not directly observed. This definition of multivariate prediction is without loss of generality, because prediction of one observed variable can be achieved by defining a set of variables with only one variable. If a latent variable is the weighted sum of one set of variables, we denote the latent variable by *set variate*. The stability of a predictive set variate is expected to be highest if the predictive set variate is as representative as possible for the corresponding set of observed predictor variables. A classical statistical measure for describing the dominant variation within a set of variables is *set variance*. Set variance in this predictive context refers to the total variance of all predictor variables accounted for by the predictive set variate. Set variance is the name used in this monograph for the "optimal weighting" criterion of Principal Component Analysis (Gifi, 1990, chapter 3).

The exactness of prediction of a criterion set variate by a predictive set variate can be quantified by the squared correlation. We refer to correlations between set variates of different sets of observed variables as *set correlations*. If the set correlation is 1 or -1 an exact prediction is possible. If the set correlation is 0, prediction is not possible. Examples of set correlation criteria are Canonical Correlation Analysis (CCA) and (Canonical) Discriminant Analysis (DA) (also called canonical variate analysis, Gittins, 1985).

In the context of this monograph set variance describes variation *within* sets of variables and set correlation describes relations *between* sets of variables. Stability of set variates and exactness of multivariate prediction can now be optimized by combining set variance and set correlation criteria in one analysis.

1.2 Hybrid and adjusted methods

We discern two main strategies for integrating the criteria of set variance and set correlation. One strategy leads to *hybrid* methods and the other to *adjusted* methods. In the next two sections we provide more examples of hybrid methods and elaborate on the difference with adjusted methods. Section 1.2.3 shows how *specialization* in hybrid methods changes the competitive parts in more complementary parts.

1.2.1 Hybrid methods use competitive subfunctions

Hybrid methods merge set variance and set correlation by joining their corresponding fit functions in an additive, multiplicative, sequential or cyclic way. In principle all participating fit functions are incorporated in an equivalent way. We give some more examples.

A not very obvious example of an additive hybrid method is Redundancy Analysis (RA) of Van den Wollenberg (1977). As will be shown in chapter 2, this method uses a set variance measure for the criterion set and a set correlation measure for the predictor set. The metric of sets gives an indication of the hybrid nature of RA (cf. Meulman, 1986). In terms of the metric of sets we analyze the criterion set in the Euclidean metric and the predictor set in the Mahalanobis metric. DeSarbo (1981) formulates a weighted additive hybrid model on top of this by adding up RA and CCA with weights specified by the user. Continuum Regression proposed by Stone & Brooks (1990) is an example of a exponentially weighted multiplicative hybrid model. As will be shown in chapter 5, a set correlation fit function is multiplied with an exponentially weighted set variance fit function. Many sequential hybrid models are formulated in two-step procedures by applying a set variance function in the first step followed by a set correlation function in the second step. For instance, the Principal Components regression method extracts some suitable number of dimensions in the first step and applies regression in a second step. Hoogerbrugge, Willig and Kistemaker (1983) describe such a procedure for Discriminant Analysis, which is commonly used in chemometrics.

In hybrid methods all participating fit functions are incorporated in an equivalent way. In this sense the contributing parts have to do the *same* job: optimizing a subfunction. We conceive hybrid methods therefore as *competitive* methods.

1.2.2 Adjusted methods for a complementary approach

Adjusted methods combine set variance and set correlation criteria in such a way that one main criterion is modified by constraints corresponding to other, secondary criteria for improving stability or exactness. The adjustments can be, for example, partialling out, improving or reflecting. Each adjustment will be carried out by enforcing some kind of constraint, named after the effect of the adjustment. The *partialling out constraint* implies forcing a secondary function to be equal to zero. It nullifies all relations with external information that is known to be irrelevant. The *improving constraint* locally improves a secondary *adjusting* function. The *reflecting constraint* filters out irrelevant information as much as possible by using known relevant external information as a mirror. Space restrictions, which force latent variables to be in the space of some designated set of variables, can always be simulated by extreme weighting in multiset hybrid models. Therefore these space restrictions are not incorporated in the list of constraints for adjusted methods.

In adjusted methods we do not want different subfunctions to do the *same* job in a *competitive* way, but to do *different* jobs in a *complementary* way. We achieve this goal by maximizing *one* fit function modified by constraints of other secondary models. Set variance and set correlation are simple integrated in adjusted methods. In chapter 3 on Set Correlation with Set Variance Constraints, the emphasis is primarily on maximizing squared correlations between set variates, and secondarily on locally improving the variance accounted for by the set variates by some fixed improvement step. So the main fit function is the sum of squared set correlations, and the secondary set variance constraint enables a local improvement on the variance accounted for. In chapter 4 on Set Variance with Set Correlation Constraints the emphasis is primarily on maximizing the variance accounted for by predictive set variates, and secondly on using the correlations with external set variates for modifying the importance of the predictive set variates. As an example consider these extremes: If the correlation is 1 or −1, the importance of the predictive set variates remains unchanged; if the correlation is 0 the importance is 0. In fact, we maximize the 'relevant variance accounted for' by filtering out irrelevant information as much as possible. The main fit function is to maximize variance accounted for and the secondary set correlation constraint is to assess the relevance of the information. For this purpose we formulate *reflecting filters*, which project variables on to a reference set and then reflect these variables back to their own set.

1.2.3 Specialization in hybrid methods

Competitive hybrid methods can approximate the complementary property of adjusted methods by an appropriate modification of the fit functions involved. In chapter 5 on Directed Correlations and Partial Least Squares we formulate a multiplicative hybrid method with a global and a local fit function. The global fit function is maximized over all sets, and the optimal set variate of all other sets can change if one of the optimal set variates is changed. The local fit function gives in principle an optimum for each set variate separately; the optimal set variates remain invariant under changes of other sets. By this approach the maximization of the local fit function is focused within sets and occupies a different niche from the global function, which maximizes the relations between sets. Another example of specialization is to model the projections of set variates in regions with moderate and high eigenvalues different from projections in spurious regions with low eigenvalues.

1.3 Multidimensional geometry by matrix algebra

To integrate concise descriptions of within set structure with concise descriptions of relations between sets we need some device for making concise descriptions. We have chosen to make these descriptions with matrix algebra, which offers an extension of the basic Euclidean geometry of two and three dimensions to *n* dimensions. Although it is possible to work with matrix algebra without drawing any geometric pictures, we like to keep in touch with basic Euclidean geometry. The visual illustration and explanation of matrix theory can provide valuable 'insights' and therefore several drawings of variable structures have been added. Variables or patterns can be geometrically represented by points or vectors (arrows) in a multidimensional space. We emphasize that *each* point or vector in this space stands for a whole variable or pattern. An appropriate low-dimensional representation of these points (for example in a plane) can reveal the structure of a set and/or the relations between sets. It is also possible to infer predictions from one set to another.

1.4 Models and fit functions

Throughout this monograph we will very often describe a model or multivariate technique by just giving the least squares function to fit the model. We opted for this approach with the following considerations.

- In most *loss functions* the model is fitted by minimizing the residual error and therefore the model can easily be derived from this function. If we do not confine ourselves to least squares functions, it is possible for many functions to derive a model and fit this model in another way. This viewpoint, however, will not be elaborated.
- For well-known techniques like Principal Component Analysis or Multiset Canonical Correlation Analysis it is possible to give two or maybe more models for the same least squares *fit function.* For such dual techniques a fit function gives a more precise description of its properties than just one model. A simple example of this phenomenon is given by the correlation between two variables. If we take two unit normalized variables, a fixed variable $\mathbf{h}$ and some space restricted latent variable $\mathbf{x}$ both having zero mean, then the fit function for the (Pearson) correlation is $\mathbf{h}'\mathbf{x}$. Maximizing $\mathbf{h}'\mathbf{x}$ gives the same result for $\mathbf{x}$ as minimizing the loss functions $\|\mathbf{h} - \mathbf{x}b\|^2$ or $\|\mathbf{x} - \mathbf{h}a\|^2$ with scalar weights b and a. The corresponding models derived from these loss functions are respectively $\mathbf{h}=\mathbf{x}b+\mathbf{e}_\mathbf{h}$ and $\mathbf{x}=\mathbf{h}a+\mathbf{e}_\mathbf{x}$. This makes our slight preference for fit functions plausible.

1.4.1 Notation

Without loss of generality all variables are assumed to be centred and to have unit sum of squares. For variable $\mathbf{h}$ this implies that $\mathbf{h}'\mathbf{1}=0$ and $\mathbf{h}'\mathbf{h}=1$, where $\mathbf{1}$ is a vector with elements 1. With this convenient normalization the (Pearson) correlation between two variables is denoted by the inner product, as we did in the previous section with $\mathbf{h}'\mathbf{x}$ for the correlation between $\mathbf{h}$ and $\mathbf{x}$. Geometrically the correlation $\mathbf{h}'\mathbf{x}$ gives the cosine of the angle between vector $\mathbf{h}$ and $\mathbf{x}$.

The optimal solution in p dimensions of several models is indeterminate, due to rotational freedom. Most fit functions of the methods involved can be maximized by an eigenvalue decomposition. Exceptions are Set Component Analysis and Nonlinear Reflected Discriminant Analysis, for which an iterative algorithm is described in chapter 6, and most of the PLS methods of chapter 5. We always implicitly assume that the optimal solution in p dimensions is defined by taking the first p eigenvectors with the eigenvalues arranged in descending order. By this choice, all solutions for different numbers of dimensions are nested.

1.5 Outline

In chapter 2 on Multiset Models we describe well-known multivariate methods by defining *filters* that transform the eigenvalues. The basic method for this filter approach is Multiset Filtered Component Analysis (MFCA). In chapter 2 and 3 the filters are simple functions of the eigenvalues. In chapter 4 transition matrices are incorporated in the filters, that modify the eigenvalues through projections. A preliminary version of the theory in chapter 2 is presented in Nierop, 1989. The structure of chapter 2 is guided by the construction of MVA methods with set variance and set correlation constituents. In a section about *one type of filter* we discuss methods like Multiset Principal Component Analysis (MPCA), which apply *only* set variance filters, and methods like Multiset Canonical Correlation Analysis (MCCA), which apply *only* set correlation filters. Two new MPCA methods are formulated for balancing the set variance between sets, based on *potential variance accounted for* and *information span*. In consecutive sections set variance and set correlation are integrated in hybrid methods. These methods are hybrid and not adjusted, because the integration is not with special constraints for improving stability or exactness. In a section about *different types of filters* we start with straightforward additive hybrid methods without weights. Some sequential hybrid methods are discussed in a section about *discrete compound filters*. Finally some weighted additive hybrid methods are described in the last section about *continuous compound filters*. All compound filters apply the specialization of hybrid methods mentioned at the end of section 1.2.3. Spurious regions with low eigenvalues are dominated by set variance modelling and regions with higher eigenvalues are dominated by set correlation modelling.

In chapter 3 on Set Correlation with Set Variance Constraints we formulate the adjusted method of Set Component Analysis (Nierop, 1989, 1993). The method is related with the hybrid methods of chapter 2 by describing quadratic filters. Various relations with other methods are discussed. For instance we show that maximizing the SCA fit function gives the same results as fitting the INDSCAL model and simultaneously penalizing non-orthogonality between the INDSCAL dimensions and the residuals. The properties of INDSCAL and SCA are compared in a simulation study.

In chapter 4 on Set Variance with Set Correlation Constraints or Reflected Variance the principle of reflecting variance (Nierop, 1991) is elaborated by defining Reflected Component Analysis (RCA) and Reflected Discriminant Analysis (RDA). It will be shown theoretically how and under which conditions RDA can improve group prediction compared to Discriminant Analysis (DA) and Principal Component - Discriminant Analysis (PC-DA). In a simulation study theoretical results are confirmed. Some multiset and nonlinear extensions are proposed.

In chapter 5 on Directed Correlations and Partial Least Squares a new multiplicative hybrid method is formulated that maximizes the product of two complementary fit functions, a local and a global MVA function. The local function gives a multiset alternative for maximizing variance accounted for. The global function maximizes the sum of squared correlations as formulated in chapter 3. These adjusted correlations are called *directed correlations* and are embedded in a multiset path analysis framework utilizing *primary* and *secondary* predictions. The product function that globally maximizes directed correlations and locally increases set variance as much as possible is called Lifted Directed Correlations (LDC). LDC is able to describe many existing MVA methods, hybrid and adjusted methods. It also reformulates some cyclic hybrid methods as multiplicative hybrid methods. Examples of these cyclic methods are basic Partial Least Squares (Wold, 1982), extended Partial Least Squares (Lohmöller, 1989), Consensus PLS (Geladi, Martens, Martens, Kalvenes & Esbensen, 1988), PLS1 regression (Stone & Brooks, 1990) and PLS Hierarchical

Components (Lohmöller, 1989). Hitherto the PLS system was basically defined by fitting several models cyclically. An overall maximization criterion was lacking. Therefore it was classified as 'soft modelling'. With the LDC fit function the appropriate maximization criterion is added to most of the PLS algorithms and therefore PLS is turned into 'hard modelling'. Furthermore PLS variants with theoretically better predictions can now be formulated, like in section 5.4.8 for the asymmetric PLS2 regression method (Manne, 1987). Continuum regression (Stone & Brooks, 1990) is reformulated as a weighted multiplicative hybrid method. At the end of chapter 5 adjusted methods like SCA of chapter 3 and reflected variance methods of chapter 4 are formulated as special cases of directed correlations theory.

In chapter 6 we present two algorithms for non eigenvalue-eigenvector problems. First a simultaneous and successive monotone convergent algorithm for Set Component Analysis (chapter 3) is developed, where an interesting general algorithmic subproblem is to maximize the set variance of different matrices by corresponding orthogonal latent variables. Secondly we elaborate a monotone convergent algorithm for Nonlinear Reflected Discriminant Analysis (chapter 4).

In chapter 7 we present analyses of real-life data using three methods developed in the preceding chapters. For a psychometric application of Set Component Analysis (chapter 3) we compare the SCA solution of the Miller-Nicely data with the corresponding INDSCAL solution. Reflected Discriminant Analysis from chapter 4 is applied on mass spectrometric barley tissue profiles and compared with results for PC-DA. The barley tissue profiles are also analysed with Nonlinear Reflected Discriminant Analysis.

Finally we have concluding remarks in chapter 8. Some methods and extensions in the line of this monograph are indicated that might give useful analytic tools in the future.

Chapter 2

A FILTER VIEW ON MULTISET MODELS

Many Multivariate Analysis (MVA) methods are build up with set variance and set correlation constituents. Our first aim is to show a variety of construction methods and not an exhaustive inventory of methods. Two new methods are proposed, based on *potential variance accounted for* and *information span*. The last three main sections show how set variance and set correlation can be integrated with competitive subfunctions and therefore illustrate the concept of hybrid methods.

Introduction

We describe well-known multiset multivariate methods in a comprehensive system by *filtering* the eigenvalues of sets of variables (Nierop, 1989). The approach is inspired by Van de Geer (1986). All filters in this chapter apply simple functions to the eigenvalues of the original data. In section 2.1 the basic method for filtering eigenvalues is described as Multiset Filtered Component Analysis (MFCA). We illustrate in the MFCA framework how space restrictions, which force latent variables to be in the space of some designated set of variables, can be used for specific prediction purposes. The global structure of this chapter is guided by the construction of MVA methods with set variance and set correlation constituents. Local attention is given to the balancing of set variance between sets.

In section 2.2 we discuss methods that define only *one type of filter* for all sets. In section 2.2.1 we introduce Multiset Principal Component Analysis (MPCA), which applies only *set variance* filters. Two new MPCA methods are formulated for balancing the set variance between sets, based on *potential variance accounted for* (section 2.2.4) and *information span* (section 2.2.5). A completely new approach of balancing sets is offered in chapter 5 and denoted by multiset reciprocal PCA. RPCA differs so much from the system of methods presented in this chapter that a separate treatment is justified. The last 'one type of filter' method we describe is Multiset Canonical Correlation Analysis (MCCA), defined only by *set correlation* filters. We show how to apply space restrictions in MCCA in order to obtain ordinary 2 sets

CCA. In the consecutive sections 2.3 to 2.5 set variance and set correlation are integrated in hybrid methods. These methods are hybrid and not adjusted, because the integration is without special constraints for improving stability or exactness. In section 2.3 we give some examples of straightforward additive hybrid methods that define a *different type of filter* for different sets. Redundancy Analysis (Van den Wollenberg, 1977) and multiset generalizations of RA are formulated by specifying different types of filters in MFCA. Sequential hybrid methods very often are two-step methods. The two-step methods are discussed in section 2.4 about *discrete compound filters*. In the last section about *continuous compound filters* we give weighted hybrid methods like Multiset Ridge Regression (MRR) and Fixed Set Component Analysis (FSCA) that approximate the complementary approach of adjusted methods mentioned in chapter 1 the most. All compound filters apply the specialization of hybrid methods mentioned at the end of section 1.2.3. Spurious regions with low eigenvalues are dominated by set variance modelling and regions with higher eigenvalues are dominated by set correlation modelling.

2.1 Multiset Filtered Component Analysis (MFCA)

We present Multiset Filtered Component Analysis as a tool box for constructing many MVA methods. The MFCA method is additive with respect to the contribution of the sets. Basic components are *filters* for each set which model the eigenvalue structure of these sets. In this chapter the filters only consist of simple functions applied to the eigenvalues.

Suppose the data to be analyzed are collected in the matrix $\mathbf{H}$, partitioned into K sets: $\mathbf{H}= (\mathbf{H}_1,..,\mathbf{H}_k,..,\mathbf{H}_K)$ with n rows (objects) and m_k columns (variables) for set $\mathbf{H}_k$. We assume without loss of generality that the variables are centred and have unit sum of squares, so the columns have sum of squares equal to 1. These assumptions imply that $\mathbf{H}_k'\mathbf{H}_k$ is a correlation matrix between the variables of set k. The Singular Value Decomposition (SVD) for set k is given by $\mathbf{H}_k = \mathbf{P}_k\mathbf{\Phi}_k\mathbf{Q}_k'$, where $\mathbf{P}_k$ $(n \times p_k)$ and $\mathbf{Q}_k$ $(m_k \times p_k)$ denote orthonormal singular vector matrices and $\mathbf{\Phi}_k$ denotes a diagonal matrix with p_k non-zero singular values in descending order $(p_k \leq m_k)$. So $\mathbf{H}_k\mathbf{H}_k'= \mathbf{P}_k\mathbf{\Phi}_k^2\mathbf{P}_k'$ and the eigenvectors $\mathbf{P}_k$ and the eigenvalues $\mathbf{\Phi}_k^2$ of the symmetric matrix $\mathbf{H}_k\mathbf{H}_k'$ are equal to respectively the singular vectors $\mathbf{P}_k$ and the squared non-zero singular values Φ_k of $\mathbf{H}_k$. For the analysis of $\mathbf{H}_k\mathbf{H}_k'$ we only need $\mathbf{P}_k$ and Φ_k^2, but not $\mathbf{Q}_k$. The eigenvectors $\mathbf{P}_k$ can be interpreted as independent basic *information patterns* derived from the variables and the eigenvalues Φ_k^2 as *information weights* assigned to these information patterns.

The additive fit function of Multiset Filtered Component Analysis is a function of common latent variables $\mathbf{X}$. It maximizes

$$\text{MFCA:} \quad \text{Fit}(\mathbf{X}) = \sum_{k=1}^{K} \text{tr } \mathbf{X}'\mathbf{P}_k\Omega_k(\Phi_k^2)\mathbf{P}_k'\mathbf{X}, \tag{2.1}$$

with $\mathbf{X}'\mathbf{X}=\mathbf{I}$, where $\mathbf{P}_k$ is given by the SVD for set k, $\mathbf{H}_k = \mathbf{P}_k\Phi_k\mathbf{Q}_k'$, and where $\Omega_k(\Phi_k^2)$ denotes a matrix with the filtered eigenvalues of set k. The filter Ω maps the values of some matrix $\mathbf{A}$ in a formal way into matrix $\Omega(\mathbf{A})$. The filter is indexed with k, so every set has its own filter Ω_k.

We now derive particular methods by specifying the filters. We give some simple examples and start with Multiset Principal Component Analysis (MPCA). The method will be discussed in section 2.2.1 and is defined by a set variance filter

$$\text{MPCA:} \qquad \Omega_k(\Phi_k^2) = \Phi_k^2 w_k^{-1}. \tag{2.2}$$

As will be shown later on in section 2.2.1, substitution of this filter in (2.1) results in (2.7). The loss function corresponding to (2.7) is called SUMPCA* by Kiers (1989, p.15). All different types of MPCA formulated in section 2.2.1 will be described by substituting different scalars w_k in (2.2). The names of the corresponding filters are: *identity* filter for w_k=1, *trace* filter for $w_k=\text{tr}\Phi_k^2$, *first eigenvalue* filter for $w_k=\phi_{1k}^2$ and *maxVAF* filter for $w_k=\sum_p \phi_{sk}^2$.

For Multiset Canonical Correlation Analysis (Carroll, 1968) to be discussed in section 2.2.6, we have a set correlation filter

$$\text{MCCA:} \qquad \Omega_k(\Phi_k^2) = \mathbf{I}, \tag{2.3}$$

where $\mathbf{I}$ is an identity matrix of appropriate size. Substitution of this *constant* filter in (2.1) results later in formula (2.22) of section 2.2.6 about MCCA. By applying the constant filter we replace the Pythagorean distances between the rows of a certain set by the Mahalanobis distances (Meulman, 1986).

For the description of Canonical Correlation Analysis (CCA) we have to add subspace restrictions for predictor set c to (2.3)

$$\text{CCA:} \qquad \Omega_k(\Phi_k^2) = \mathbb{I}, \text{ with } \mathbb{X} = \mathbb{P}_c\mathbb{P}_c'\mathbb{X}. \qquad \text{for } K{=}2 \qquad (2.4)$$

The canonical variates of set c are given by the optimal $\mathbb{X}$ for c =1 and c =2. This is explained in section 2.2.8.

Many other methods can be described with MFCA by specifying filters, whether or not in combination with subspace restrictions. In the next section we will first elaborate more extensively on the subject of subspace restrictions.

2.1.1 Subspace restrictions for prediction

In general, subspace restrictions introduce an asymmetry in the analysis concerning the location of the common latent variables in one particular set. The motivation for this asymmetry can be prediction; for instance, the prediction of one or more sets of criterion variables by a linear combination of predictor variables. The common latent variables of MFCA can only be expressed as a linear combination of predictor variables, if the space spanned by the predictor set also includes the common latent variables. For predictor set c this is achieved by requiring $\mathbb{X} = \mathbb{P}_c\mathbb{P}_c'\mathbb{X}$. The $\mathbb{X}$ are sometimes labeled with c in order to discriminate between the optimal solutions $\mathbb{X}_{(c)}$ that we obtain after imposing subspace restrictions on different sets c. We use the notation 'c' especially for subspace restrictions in order to avoid confusion with 'k' in subsequent sections. We give two examples of applying subspace restrictions for prediction purposes.

The introduction of prediction in MCCA (2.3) is formulated as

$$^{c}\text{MCCA:} \qquad \Omega_k(\Phi_k^2) = \mathbb{I}, \text{ with } \mathbb{X} = \mathbb{P}_c\mathbb{P}_c'\mathbb{X}, \qquad (2.5)$$

where the upper left superscript c of ^{c}MCCA indicates that we are dealing with a subspace restriction on set c in MCCA. By maximizing this fit function we find common latent variables that are a linear combination of set c and have the highest sum of squared canonical correlations with all other sets. In this case set c is a predictor set in a 'set variate' sense, because it maximizes the relations with the set variates and not with the set variables in terms of variance accounted for. The maximization of ^{c}MCCA for $K = 2$ and $c = 1$ or 2 gives the CCA solution, with the canonical variates of set c equal to $\mathbb{X}_{(c)}$. The two maximization problems lead in principle to the same eigenvalue problem (section 2.2.8). If we know the prediction for one set, we can easily derive from this solution the prediction for the other set. For more than two sets we cannot reduce the K maximization problems for $c = 1,\ldots,K$ to one single eigenvalue problem. Each prediction problem has to be solved by itself, unless some of the $\mathbb{P}_c$ matrices occupy exactly the same space. In Gifi (1990) a related version of ^{c}MCCA is used for multivariate analysis of variance.

In an analogous way we introduce prediction of set variables in MPCA (2.2) by

$$^c\text{MPCA:} \qquad \Omega_k(\Phi_k^2) = \Phi_k^2 w_k^{-1}, \text{ with } \mathbb{X} = \mathbb{P}_c\mathbb{P}_c{}'\mathbb{X}, \qquad (2.6)$$

with a subspace restriction on predictor set c in MPCA. By maximizing this fit function for w_k=1 $\forall k$ we find common latent variables that are a linear combination of set c and have the largest variance accounted for of all sets. For $w_k = (\text{tr}\Phi_k^2)$ $\forall k$ we maximize the mean proportion of variance of all sets accounted for by the predictor set. For two sets (2.6) is equal to Principal Covariates Regression as proposed by De Jong & Kiers (1992) with α=w_1 and $(1-\alpha)$=w_2.

It is important to notice that subspace restrictions in an additive multiset method like MFCA are also possible by introducing a very large weight w_c in filter Ω_c of predictor set c. In this way (2.6) can for instance be simulated with (2.2). Therefore subspace restrictions add no essential new feature to the MFCA method. Nevertheless they are convenient and in next sections we will find other examples of applying subspace restrictions for prediction purposes.

2.2 One type of filter

In the following subsections we give examples of methods that define only one type of filter for all sets, only set variance or only set correlation filters. First we present straightforward generalizations of Principal Component Analysis (PCA) for multiple sets (MPCA) by applying set variance filters and discuss in section 2.2.2 the upper bounds for the variance accounted for under several conditions. These upper bounds lead to the definition of *potential variance accounted for* and a corresponding balancing of sets in section 2.2.4. In section 2.2.5 the *information span* is proposed as a measure for assessing the efficiency of information transfer and the sets are balanced with regard to this information span. Next Multiset Canonical Correlation Analysis (MCCA) and ordinary 2 sets Canonical Correlation Analysis are defined briefly. These methods apply set correlation filters. The relation of CCA with MCCA and MFCA is established in section 2.2.8.

2.2.1 Multiset Principal Component Analysis (MPCA)

A generalization of PCA for multiple sets is described by maximizing the following function

$$\text{MPCA:}\quad \text{Fit}(\mathbf{X}) = \text{tr} \sum_{k=1}^{K} w_k^{-1}\, \mathbf{X}'\mathbf{H}_k\mathbf{H}_k'\mathbf{X} = \text{tr} \sum_{k=1}^{K} \mathbf{X}'\mathbf{P}_k(\Phi_k^2 w_k^{-1})\mathbf{P}_k'\mathbf{X}$$

$$= \text{tr}\ \mathbf{X}'\mathbf{H}\mathbf{D}_w^{-1}\mathbf{H}'\mathbf{X} = \text{tr}\ \mathbf{X}'\mathbf{P}_{\text{Part}}(\Phi_{\text{Part}}^2\mathbf{D}_w^{-1})\mathbf{P}_{\text{Part}}'\mathbf{X}, \qquad (2.7)$$

where $_n\mathbf{X}_p = (\mathbf{x}_1,\ldots,\mathbf{x}_s,\ldots,\mathbf{x}_p)$ denote the common latent variables with $\mathbf{X}'\mathbf{X}=\mathbf{I}$,

$w_1,\ldots,w_k,\ldots,w_K$ denote fixed *balancing constants* for set k,

$$\mathbf{D}_w = \begin{array}{|c|c|c|}\hline w_1\mathbf{I} & 0 & 0 \\ \hline 0 & w_k\mathbf{I} & 0 \\ \hline 0 & 0 & w_K\mathbf{I} \\ \hline \end{array}$$

,with $\mathbf{I}$ of appropriate size,

$\mathbf{P}_{\text{Part}}=(\mathbf{P}_1,\ldots,\mathbf{P}_k,\ldots,\mathbf{P}_K)$,where we use the notation 'Part' to indicate a partitioned matrix,

$$\Phi_{\text{Part}} = \begin{array}{|c|c|c|}\hline \Phi_1 & 0 & 0 \\ \hline 0 & \Phi_k & 0 \\ \hline 0 & 0 & \Phi_K \\ \hline \end{array}$$

The loss function corresponding to (2.7) is called SUMPCA* by Kiers (1989, p.15). We point out that $\mathbf{P}_{\text{Part}}\Phi_{\text{Part}}^2\mathbf{P}_{\text{Part}}'$ does not give the eigenvalue decomposition of matrix $\mathbf{H}\mathbf{H}'$, because $\mathbf{P}_{\text{Part}}=(\mathbf{P}_1,\ldots,\mathbf{P}_K)$ is usually not an orthonormal matrix. Only the $\mathbf{P}_k$ are orthonormal for each set k separately. As we saw in section 2.1 formula

(2.2) substitution of this MPCA filter in (2.1) results in (2.7). The last part of both lines in equation (2.7) is derived from the Singular Value Decomposition (SVD) of each $\mathbb{H}_k$, with $\mathbb{H}_k = \mathbb{P}_k\Phi_k\mathbb{Q}_k'$. This part of (2.7) is added to get used to the idea that the balancing of sets can also be described by a rescaling of the eigenvalues of these sets. It shows that instead of maximizing the variance accounted for by $\mathbb{X}$ of the partitioned matrix

$$\mathbb{H}\mathbb{D}_w^{-1/2} = (\mathbb{H}_1 w_1^{-1/2},..,\mathbb{H}_k w_k^{-1/2},..,\mathbb{H}_K w_K^{-1/2}),$$

we can equivalently maximize the variance accounted for of the partitioned matrix

$$\mathbb{P}_{\text{Part}}\Phi_{\text{Part}}\mathbb{D}_w^{-1/2} = (\mathbb{P}_1\Phi_1 w_1^{-1/2},..,\mathbb{P}_k\Phi_k w_k^{-1/2},..,\mathbb{P}_K\Phi_K w_K^{-1/2}).$$

As mentioned before the loss function corresponding to (2.7) is called SUMPCA* by Kiers (1989, p.15). The application of balancing constants can for instance be found in the sequential hybrid STATIS method developed by L'Hermier des Plantes (1976) (See also Kiers, 1989, p. 10). Because we are maximizing balanced variance accounted for, we can relate the balancing constants w_k with the variance accounted for. In the next sections we will first derive for each set separately the upper bounds of the variance accounted for, then we describe some examples of balancing in MPCA by defining several MPCA filters. All MPCA filters are conceived as *set variance* filters.

2.2.2 *Potential variance accounted for*

Our rationale for the balancing of sets in MPCA is to control the maximum influence of the sets on the solution with respect to their set variance. The influence of set k is measured by the Variance Accounted For (VAF) by p common latent variables $\mathbb{X}$. In matrix formulation this VAF is equal to

$$\text{VAF}(\mathbb{X},k) = \text{tr } \mathbb{X}'\mathbb{H}_k\mathbb{H}_k'\mathbb{X} = \text{tr } \mathbb{X}'\mathbb{P}_k\Phi_k^2\mathbb{P}_k'\mathbb{X}. \tag{2.8}$$

An upper bound for (2.8) can be derived by Theorem 2 of Ten Berge (1983). This theorem can be applied, because matrix $\mathbb{X}'\mathbb{P}_k$ is a suborthonormal matrix with rank $\leq p$ and Φ_k^2 is a fixed diagonal matrix. A matrix is suborthonormal if it is a submatrix of an orthonormal matrix. For $\mathbb{X}'\mathbb{P}_k$ this orthonormal matrix can be constructed by adding to the columns of $\mathbf{X}$ and $\mathbf{P}_k$ their (semi-)orthonormal complement and multiplying the two $(n \times n)$ orthonormal matrices in an appropriate way. The inequality resulting from Theorem 2 is in our case

$$\mathrm{tr}\ \mathbf{X}'\mathbf{P}_k\mathbf{\Phi}_k^2\mathbf{P}_k'\mathbf{X} \le \sum_{s=1}^{p} \phi_{sk}^2, \tag{2.9}$$

with the eigenvalues ϕ_{sk}^2 arranged in descending order. The upper bound for set k of the VAF in p dimensions is now given by the sum of the first p eigenvalues

$$\mathrm{maxVAF}(p,k) = \sum_{s=1}^{p} \phi_{sk}^2, \tag{2.10}$$

with the eigenvalues ϕ_{sk}^2 arranged in descending order. We will refer to the value of maxVAF(p,k) as the *potential variance accounted for* (*potential VAF*). It gives the highest possible variance accounted for that can be found for set k in a p dimensional solution. We now derive upper and lower bounds of maxVAF(p,k). The upper bound is reached, if $\mathbf{H}_k$ is of deficient rank in the sense that the m_k–p smallest eigenvalues are all equal to zero and maxVAF(p,k)=tr$\mathbf{\Phi}_k^2$=m_k. The lower bound is reached if $\mathbf{H}_k$ is orthonormal and therefore all eigenvalues are equal to 1. Under this condition the m_k–p smallest eigenvalues are as large as possible, and maxVAF(p,k)=p. Summarizing the results for set k with unit normalized variables the potential VAF varies between

$$p \le \mathrm{maxVAF}(p,k) \le m_k. \tag{2.11}$$

2.2.3 *Different ways for defining balance among sets*

We formulate some straightforward types of MPCA by fixing the balancing constants w_k in a simple way. In the next section we elaborate on the relation between balancing sets and variance accounted for. The first type of analysis is MPCAi with w_k=1 $\forall k$. So the rescaled eigenvalues are *identical* to the original eigenvalues. After substitution of w_k=1 $\forall k$ in (2.7) we have to maximize

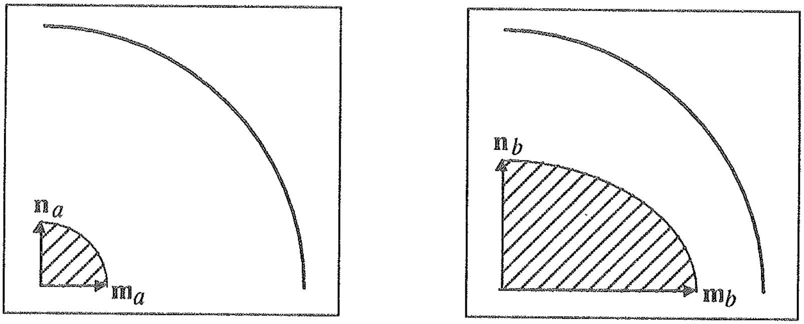
$$\text{MPCAi:}\quad \mathrm{Fit}(\mathbf{X}) = \mathrm{tr}\ \mathbf{X}'\mathbf{H}\mathbf{H}'\mathbf{X} = \mathrm{tr}\ \mathbf{X}'\mathbf{P}_{\mathrm{Part}}\mathbf{\Phi}_{\mathrm{Part}}^2\mathbf{P}_{\mathrm{Part}}'\mathbf{X}, \tag{2.12}$$

with $\mathbf{X}'\mathbf{X}=\mathbf{I}$ and notation according to (2.7). The type of set balancing presented above amounts to maximizing the variance accounted for by $\mathbf{X}$ of the partitioned matrix $\mathbf{P}_{\text{Part}}\Phi_{\text{Part}}$. The way the variables are grouped in sets has no influence on the solution, because $\mathbf{P}_{\text{Part}}\Phi^2_{\text{Part}}\mathbf{P}_{\text{Part}}'=\mathbf{H}\mathbf{H}'$.

The second type of analysis in this section we call MPCAt with $w_k = (\text{tr}\Phi_k^2)$ $\forall k$. In this case the balancing constant is the *trace* of $\mathbf{H}_k\mathbf{H}_k'$, which is equal to the sum of the eigenvalues Φ_k^2 and equal to the sum of squares of $\mathbf{H}_k$. With unit normalized variables we have $\text{tr}\Phi_k^2 = m_k$. In the format of (2.2) we maximize (2.1) with filter

$$\text{MPCAt:} \qquad \Omega_k(\Phi_k^2) = \Phi_k^2 \,/\, \text{tr}\Phi_k^2. \tag{2.13}$$

For MPCAt this amounts to maximizing the variance accounted for by $\mathbf{X}$ of the partitioned matrices $\mathbf{P}_{\text{Part}}\Phi_{\text{Part}}$ (or $\mathbf{H}$) after the sum of squares of each set is normalized to 1. Geometrically this normalization is achieved by rescaling all sumvectors $\mathbf{P}_k\Phi_k\mathbf{1}$ to sum of squares 1. Verifying this statement we have $\mathbf{1}'\Phi_k\mathbf{P}_k'\mathbf{P}_k\Phi_k\mathbf{1}=\mathbf{1}'\Phi_k^2\,\mathbf{1}=\text{tr}\Phi_k^2$, $\forall k$. In figure 2.1 we show the implications of trace balancing for set a and b with each 16 variables.



A set a B set b

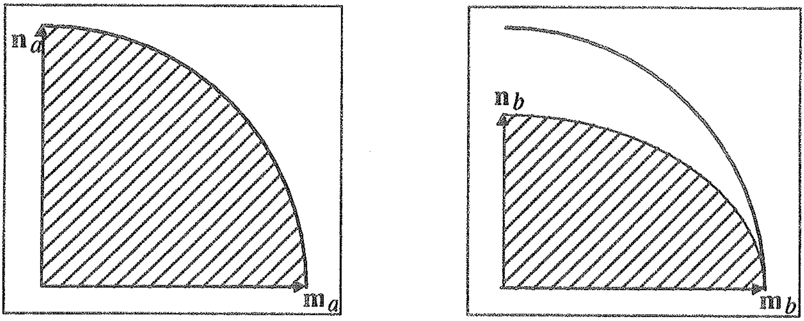
Figure 2.1 *Trace balancing in MPCAt.*

The eigenvalues of set a are all chosen equal to 1 and the first two eigenvalues of b are equal to 9 and 4. From the orthogonal rescaled matrix $\mathbf{P}_a\Phi_a(\text{tr}\Phi_a^2)^{-1/2}$ we have drawn the two largest columnvectors denoted here by $(\mathbf{m}_a,\mathbf{n}_a)$ in figure 2.1.A. The largest columnvectors $(\mathbf{m}_b,\mathbf{n}_b)$ from $\mathbf{P}_b\Phi_b(\text{tr}\Phi_b^2)^{-1/2}$ are drawn in figure 2.1.B. The contour of the shaded ellipses can be used to construct for any vectors $\mathbf{X}$ the variance accounted for. This will be elaborated in chapter 5 (see figure 5.1). The radius of the thick quarter-circles is equal to the length of the rescaled sumvectors $\mathbf{P}_a\Phi_a\mathbf{1}(\mathrm{tr}\Phi_a^2)^{-1/2}$ and $\mathbf{P}_b\Phi_b\mathbf{1}(\mathrm{tr}\Phi_b^2)^{-1/2}$. All sumvectors rescaled by trace balancing lie on a hypersphere with radius 1. The potential VAF for p=2 is .13 for set a and .81 for set b. Generally we state that the trace balancing allows for great relative differences in potential VAF, especially when the number of dimensions p is small. Of course the MPCAt solutions will generally be dominated by sets with large potential VAF. We choose MPCAt if we want the mean proportion of variance accounted for by $\mathbf{X}$ as large as possible.

The third type of analysis is MPCAf with $w_k = \phi_{1k}^2 \; \forall k$. The balancing constant in (2.7) is now the *first eigenvalue* ϕ_{1k}^2 of $\mathbf{H}_k\mathbf{H}_k'$, with the eigenvalues arranged in descending order. We maximize (2.1) with filter

MPCAf: $$\Omega_k(\Phi_k^2) = \Phi_k^2 \,/\, \phi_{1k}^2. \tag{2.14}$$

In figure 2.2 we show the implications of the first eigenvalue balancing by rescaling set a and b appropriately.



A set a B set b

Figure 2.2 *First eigenvalue balancing in MPCAf.*

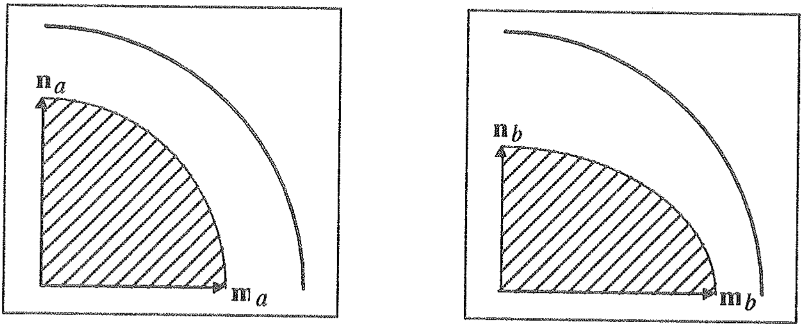
Figure 2.2 has the same design as figure 2.1 only the scaling is changed. The radius of the thick quarter-circles is equal to the length of the largest columnvectors $\mathbf{m}_a$ and $\mathbf{m}_b$ of respectively $\mathbf{P}_a\Phi_a\phi_{1a}^{-1}$ and $\mathbf{P}_b\Phi_b\phi_{1b}^{-1}$. The potential VAF for p=2 is 2.00 for set a and 1.44 for set b. Generally we state that the first eigenvalue balancing allows for small relative differences in potential VAF, especially when the number of dimensions is very small. The relative differences between sets in potential VAF can increase quickly, if the number of dimensions p increases.

2.2.4 *Balancing of sets related to variance accounted for*

In the previous section we defined several types of set balancing. For the identical balancing in MPCAi with w_k=1 $\forall k$, the potential VAF of set k is between p and m_k, dependent on the structure of the rescaled variables $\mathbf{H}_k w_k^{-1/2}$ as derived in (2.11). For MPCAt this range is between p/m_k and 1. We can find considerable differences in potential VAF between sets, if the number of variables of sets is large compared to the number of dimensions. From the potential variance point of view the appropriate balancing for set k in (2.7) would be to take w_k = maxVAF(p,k) $\forall k$, as formulated in (2.10), in order to obtain an equally balanced maximum influence of the sets on the solution with respect to their set variance. The *maxVAF* balancing gives another type of MPCA. We maximize (2.1) with filter

$$\text{MPCAm:} \qquad \Omega_k(\Phi_k^2) = \Phi_k^2 / \sum_{s=1}^{p} \phi_{sk}^2, \qquad (2.15)$$

with p equal to the number of columns of **X**. In figure 2.3 with the same design as for figure 2.1 we illustrate the maxVAF balancing by rescaling set a and b for p=2 dimensions (2.15).



A set a **B** set b

Figure 2.3 *MaxVAF balancing in MPCAm.*

The radius of the thick quarter-circles is equal to the length of the rescaled sumvectors $\mathbf{P}_a\Phi_a\mathbf{1}(\phi_{1a}^2+\phi_{2a}^2)^{-1/2}$ and $\mathbf{P}_b\Phi_b\mathbf{1}(\phi_{1b}^2+\phi_{2b}^2)^{-1/2}$. We have drawn the largest columnvectors ($\mathbf{m}_a$,$\mathbf{n}_a$) and ($\mathbf{m}_b$,$\mathbf{n}_b$) from respectively $\mathbf{P}_a\Phi_a(\phi_{1a}^2+\phi_{2a}^2)^{-1/2}$ and $\mathbf{P}_b\Phi_b(\phi_{1b}^2+\phi_{2b}^2)^{-1/2}$ in figure 2.3.A and 2.3.B. The potential VAF for p=2 is 1.00 for set a and 1.00 for set b as could be expected.

The balancing of both MPCAt (2.13) and MPCAf (2.14) appears to be special cases of MPCAm balancing. It only depends on the number of dimensions we want to compute. For p equal to the maximum number of dimensions we find the equality MPCAm = MPCAt and for $p = 1$ the equality MPCAm = MPCAf. The solutions of MPCAm are not nested. (Solutions are nested if successive computation of p dimensions always gives the same results as simultaneous computation of p dimensions for all possible p.) This property can be seen as a drawback compared to MPCAt and MPCAf. If one definitely wants to choose only between *trace* balancing or *first eigenvalue* balancing, the number of dimensions of the solution compared to the maximum number of variables in a set has to be decisive.

An even more strict equality concept can be formulated by requiring an equal balance of the influence of sets on *each* dimension with respect to set variance. In the case of successive one dimensional solutions the MPCAm fit function can be adapted to this concept by using *deflation*. After each successive step the original matrices $\mathbf{H}_k$ are replaced by their antiprojections on the previous dimensions of $\mathbf{x}_r$. The resulting $\mathbf{X}$ will be orthogonal. In fact we apply first eigenvalue balancing, because in each step there is only one dimension to find the maximum of

$$\text{dMPCAf: Fit}(\mathbf{X}) = \sum_{s=1}^{p}\sum_{k=1}^{K} \mathbf{x}_s'\mathbf{P}_{(k)s}(\mathbf{\Phi}^2_{(k)s}/\phi^2_{1(k)s})\mathbf{P}_{(k)s}'\mathbf{x}_s, \qquad (2.16)$$

where deflation enters this fit function by defining

$$\mathbf{H}_{(k)s}\mathbf{H}_{(k)s}' = \mathbf{P}_{(k)s}\mathbf{\Phi}^2_{(k)s}\mathbf{P}_{(k)s}' = \mathbf{H}_k\mathbf{H}_k' = \mathbf{P}_k\mathbf{\Phi}^2_k\mathbf{P}_k' \qquad \text{for } s = 1 \ \forall k$$

$$\begin{aligned}\mathbf{H}_{(k)s}\mathbf{H}_{(k)s}' &= \mathbf{P}_{(k)s}\mathbf{\Phi}^2_{(k)s}\mathbf{P}_{(k)s}' = (\mathbf{I}-\mathbf{x}_{s-1}\mathbf{x}_{s-1}')\mathbf{H}_{(k)s-1}\mathbf{H}_{(k)s-1}'(\mathbf{I}-\mathbf{x}_{s-1}\mathbf{x}_{s-1}')\\ &= (\mathbf{I}-\mathbf{x}_{s-1}\mathbf{x}_{s-1}')\mathbf{P}_{(k)s-1}\mathbf{\Phi}^2_{(k)s-1}\mathbf{P}_{(k)s-1}'(\mathbf{I}-\mathbf{x}_{s-1}\mathbf{x}_{s-1}'). \qquad \text{for } s = 2,\ldots,p \ \forall k\end{aligned}$$

In each successive dimension the datamatrices $\mathbf{H}_{k(r)}$ change and therefore the singular vectors and singular values of the sets also change.

2.2.5 The information span of matrices

Instead of variance accounted for we can formulate another reasonable principle for the balancing of sets by assessing the amount of superfluous information. The efficiency of information transfer can be called the information span.

The information span of a *set of variables* is high if there is no superfluous or spurious information. Geometrically this means that the variables are all orthogonal. The information span of a set is low if all the variables of the set contain the same information. We then have just m_k replications of the same variable.

We gain an insight in the information span of a matrix by studying the eigenvalue structure of a matrix. We presume that the researcher has collected the data in such a way that all the variables of one set are possible candidates to describe some relation with other sets or with some latent variable. All the variables contain reliable information of equal importance and therefore we set the information weight of each *variable* equal to one, equal to the scaling of unit normalized variables. With these assumptions and normalizations the eigenvalues indicate the *information weight* of the eigenvectors (*information patterns*). If an eigenvalue (information weight) is greater than 1, the information of the corresponding eigenvector (information pattern) is supported too much by the variables with respect to the efficiency of information transfer. It indicates that there is a replication or gradual resemblance between some variables. The part of the eigenvalue that is greater than 1 is actually referring to superfluous information. If an eigenvalue is smaller than 1, the information of the corresponding eigenvector is not supported enough by the variables. In the extreme case the eigenvalue (information weight) is close to zero and the information of the corresponding eigenvector (information pattern) is very spurious. Summarizing we can construct a measure for information span of a matrix by adding up the non-superfluous part of the eigenvalues by taking 1 as an upper cut off value and dividing this by the sum of all eigenvalues. We describe the information span I_k of set $\mathbb{H}_k$ with m_k unit normalized variables by

$$I_k = \sum_{t=1}^{m_k} I(\phi_{tk}^2), \qquad \forall k \qquad (2.17)$$

with

$$\left\{ \begin{array}{ll} I(\phi_{tk}^2) = Q_k \phi_{tk}^2 m_k^{-1} & \text{for } Q_k \phi_{tk}^2 \leq 1 \\ I(\phi_{tk}^2) = Q_k m_k^{-1} & \text{for } Q_k \phi_{tk}^2 > 1 \end{array} \right\}, \qquad \forall k,t$$

where the rank quotient Q_k is defined by

$$Q_k = \sum_{t=1}^{m_k} Q(\phi_{tk}^2), \qquad \forall k \quad (2.18)$$

with

$$\left\{ \begin{array}{ll} Q(\phi_{tk}^2) = 0 & \text{for } \phi_{tk}^2 = 0 \\ Q(\phi_{tk}^2) = m_k^{-1} & \text{for } \phi_{tk}^2 > 0 \end{array} \right\}, \qquad \forall k,t$$

where ϕ_{tk} denote the diagonal elements of Φ_k defined by the complete full rank SVD $\mathbf{H}_k = \mathbf{P}_k \Phi_k \mathbf{Q}_k'$. The rank quotient Q_k is incorporated in the information span I_k in order to correct for matrices $\mathbf{H}_k$ of deficient rank that have for instance less rows than columns. By adding many variables to a set one can artificially blow up spurious information patterns to have a 'stable' information weight ≥1. Usually this kind of stabilization is misleading, but if one is sure that all the added variables are very reliable, the correction can of course be omitted by setting Q_k=1, $\forall k$ (2.18).

The definition of information span for unit normalized variables results in an upper bound of $I_k = 1$, if there is no superfluous information, and a lower bound of $I_k = m_k^{-1}$, if we have m_k replications of the same variable. The rank of $\mathbf{H}_k$ is given by $Q_k m_k$. The *efficient* rank of $\mathbf{H}_k$ is defined by $I_k m_k$, if desired rounded off to the nearest integer. In chapter 7 the efficient rank will be computed for real-life examples.

By now we can formulate a balancing of MPCA based on the principle of information span. We formulate MPCAs by specifying the appropriate *information span* filter

$$\text{MPCAs:} \qquad \Omega_I(\Phi_k^2) = \Phi_k^2 I_k, \qquad \forall k \quad (2.19)$$

which has to be substituted in (2.1). The same balancing is achieved by taking the balancing constants of MPCA in (2.7) equal to $w_k = I_k^{-1}$. If the variables of all sets are unit normalized, we can say that in MPCAs the sets are weighted by their efficient rank.

The effects of the balancing in MPCAs for the extreme cases of I_k appeal to common sense. If there is no superfluous information, there is no change in the weight of the set. If all the variables of the set are exactly equal, the set contributes to the solution as if it has only one unit normalized variable. In this way the sets are balanced by the diversity of their information patterns. In the preceding sections the emphasis was more on the information weights and less on the diversity of the information patterns. The balancing of sets by the trace of the eigenvalues in formula (2.13) is an example of focussing only on the quantity of information. The first eigenvalue and maxVAF balancing give an intermediate approach, because they partly include the variation of the eigenvalue structure. In some simple eigenvalue structure cases they give the same results as the information span balancing.

The techniques formulated by moulding MPCA emphasize an equally balanced influence on the solution with respect to set variance, but they do not affect the correlations of set variates. With set variates we denote linear combinations of set variables. In the next section we will discuss a technique that balances the influence on the solution only with respect to set correlation.

2.2.6 *Multiset Canonical Correlation Analysis (MCCA)*

In MCCA as formulated by Carroll (1968) we maximize squared correlations between canonical variates and p common latent variables

$$\text{MCCA:}\quad \text{Fit}(\mathbf{X},\mathbf{Z}_1,..,\mathbf{Z}_k,..,\mathbf{Z}_K) = \sum_{s=1}^{p}\sum_{k=1}^{K} (\mathbf{x}_s'\mathbf{z}_{(k)s})^2, \tag{2.20}$$

where $\quad {}_n\mathbf{X}_p = (\mathbf{x}_1,\ldots,\mathbf{x}_s,\ldots,\mathbf{x}_p)$ denote the common latent variables with $\mathbf{X}'\mathbf{X}=\mathbf{I}$,

and $\quad {}_n(\mathbf{Z}_k)_p = (\mathbf{z}_{(k)1},\ldots,\mathbf{z}_{(k)s},\ldots,\mathbf{z}_{(k)p})$ denote the unit normalized canonical variates for set k and dimension s,

with $\quad \mathbf{z}_{(k)s} = \mathbf{H}_k\mathbf{t}_{(k)s} = \mathbf{P}_k\Phi_k\mathbf{Q}_k'\mathbf{t}_{(k)s} = \mathbf{P}_k\mathbf{v}_{(k)s}$,

and $\quad \mathbf{z}_{(k)s}'\mathbf{z}_{(k)s} = \mathbf{t}_{(k)s}'\mathbf{H}_k'\mathbf{H}_k\mathbf{t}_{(k)s} = \mathbf{v}_{(k)s}'\mathbf{P}_k'\mathbf{P}_k\mathbf{v}_{(k)s} = \mathbf{v}_{(k)s}'\mathbf{v}_{(k)s} = 1. \qquad \forall k,s$

As in the preceding sections the SVD for set k is given by $\mathbf{H}_k = \mathbf{P}_k\Phi_k\mathbf{Q}_k'$, where $\mathbf{P}_k$ $(n \times p_k)$ and $\mathbf{Q}_k$ $(m_k \times p_k)$ denote orthonormal singular vector matrices and Φ_k

denotes a diagonal matrix with p_k non-zero singular values in descending order. The canonical weights $\mathbf{t}_{(k)s}$ can be derived from the weights $\mathbf{v}_{(k)s}$ by $\mathbf{t}_{(k)s} = \mathbf{Q}_k\Phi_k^{-1}\mathbf{v}_{(k)s}$ $\forall k,s$. Originally Carroll (1968) also incorporated weights for the sets as we have done in MPCA, but we have omitted them in MCCA.

To show the relation of MCCA with MFCA we need a description of MCCA with only the $\mathbf{X}$ as unknown parameters. Therefore we want to find suboptimal values for the canonical variates $\mathbf{z}_{(k)s}$, which are a function of $\mathbf{x}$. We substitute $\mathbf{z}_{(k)s} = \mathbf{P}_k\mathbf{v}_{(k)s}$ with $\mathbf{v}_{(k)s}'\mathbf{v}_{(k)s} = 1 \;\; \forall k,s$ in (2.20) and compute a conditional maximum for (2.20) with $\mathbf{X}$ fixed by maximizing

$$\text{Fit}(\mathbf{x}_s, c_{(k)s}, \mathbf{v}_{(k)s}) = (\mathbf{x}_s'\mathbf{P}_k\mathbf{v}_{(k)s})^2 + c_{(k)s}, \qquad \forall k,s \qquad (2.21)$$

where $\mathbf{x}_s$ denotes the fixed common latent variable $\mathbf{x}_s$,
and $c_{(k)s}$ denotes the sum for all other fixed parameters. $\forall k,s$

By applying the Cauchy-Schwarz inequality on the non fixed parameters of (2.21) we know that $(\mathbf{x}_s'\mathbf{P}_k\mathbf{v}_{(k)s})^2 \leq (\mathbf{v}_{(k)s}'\mathbf{v}_{(k)s})(\mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s) = (\mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s)$. A maximum for (2.21) is reached if $(\mathbf{x}_s'\mathbf{P}_k\mathbf{v}_{(k)s})^2 = (\mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s)$ and therefore $\mathbf{v}_{(k)s} = \mathbf{P}_k'\mathbf{x}_s(\mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s)^{-1/2}$. With this equality we simplify (2.20) by inserting the suboptimal values $\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s(\mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s)^{-1/2}$ for $\mathbf{z}_{(k)s}$

$$\text{MCCA:} \quad \text{Fit}(\mathbf{X}) = \sum_{s=1}^{p}\sum_{k=1}^{K} (\mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s(\mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s)^{-1/2})^2$$

$$= \text{tr} \sum_{k=1}^{K} \mathbf{X}'\mathbf{P}_k\mathbf{P}_k'\mathbf{X} = \text{tr}\, \mathbf{X}'\mathbf{P}_{\text{Part}}\mathbf{P}_{\text{Part}}'\mathbf{X}, \qquad (2.22)$$

with $\mathbf{X}'\mathbf{X} = \mathbf{I}$ and the matrices $\mathbf{P}_k$ collected in one partitioned matrix $\mathbf{P}_{\text{Part}} = (\mathbf{P}_1, \ldots, \mathbf{P}_K)$. As we saw in section 2.1 formula (2.3) substitution of this *constant* filter in (2.1) results in (2.22). The constant filter is a *set correlation* filter.

2.2.7 *Canonical Correlation Analysis (CCA)*

Ordinary 2-sets CCA is usually not defined in terms of the common latent variables $\mathbf{X}$. Rather we simply maximize the sum of the canonical correlations between the canonical variates of two sets. Using exactly the same notation as in the previous section this can be expressed as the maximization of

$$\text{CCA:} \qquad \text{Fit}(\mathbf{Z}_1,\mathbf{Z}_2) = \sum_{s=1}^{p} \mathbf{z}_{1(r)}'\mathbf{z}_{2(r)} = \text{tr } \mathbf{Z}_1'\mathbf{Z}_2 = \text{tr } \mathbf{V}_1'\mathbf{P}_1'\mathbf{P}_2\mathbf{V}_2, \qquad (2.23)$$

with $\mathbf{Z}_k = \mathbf{H}_k\mathbf{T}_k = \mathbf{P}_k\mathbf{\Phi}_k\mathbf{Q}_k'\mathbf{T}_k = \mathbf{P}_k\mathbf{V}_k$

and $\mathbf{Z}_k'\mathbf{Z}_k = \mathbf{T}_k'\mathbf{H}_k'\mathbf{H}_k\mathbf{T}_k = \mathbf{V}_k'\mathbf{P}_k'\mathbf{P}_k\mathbf{V}_k = \mathbf{V}_k'\mathbf{V}_k = \mathbf{I}.$ for $k = 1,2$

The solution for $\mathbf{V}_1$ and $\mathbf{V}_2$ can be found by taking respectively the p principal left and right singular vectors of matrix $\mathbf{P}_1'\mathbf{P}_2$. The singular values give the canonical correlations.

2.2.8 *Relation between CCA and MCCA*

We can derive CCA from MCCA (and MFCA) by imposing *subspace restrictions* on the common latent variables $\mathbf{X}$. We want the solution $\mathbf{X}$ to be in the subspace associated with $\mathbf{H}_c$, spanned by the orthonormal basis $\mathbf{P}_c$ (see section 2.1.1). In other words we require

$$\mathbf{X} = \mathbf{P}_c\mathbf{P}_c'\mathbf{X}, \qquad (2.24)$$

The same restriction is obtained if we require $\mathbf{X} = \mathbf{P}_c\mathbf{V}_c$ and therefore $\mathbf{V}_c = \mathbf{P}_c'\mathbf{X}$, because $\mathbf{P}_c'\mathbf{P}_c = \mathbf{I}$.

The latent variables $\mathbf{X}$, restricted to be in some specific set c, are denoted by $\mathbf{X}_{(c)}$. Insertion for $\mathbf{X}$ of respectively $\mathbf{X}_{(1)} = \mathbf{P}_1\mathbf{V}_1$ and $\mathbf{X}_{(2)} = \mathbf{P}_2\mathbf{V}_2$ in (2.22) for $K = 2$ results in the maximization of two different functions:

$$^1\text{MCCA: Fit}(\mathbf{V}_1) = \text{tr } \mathbf{V}_1'\mathbf{V}_1 + \mathbf{V}_1'\mathbf{P}_1'\mathbf{P}_2\mathbf{P}_2'\mathbf{P}_1\mathbf{V}_1$$
$$^2\text{MCCA: Fit}(\mathbf{V}_2) = \text{tr } \mathbf{V}_2'\mathbf{V}_2 + \mathbf{V}_2'\mathbf{P}_2'\mathbf{P}_1\mathbf{P}_1'\mathbf{P}_2\mathbf{V}_2, \qquad (2.25)$$

with respectively $\mathbf{X}_{(1)}'\mathbf{X}_{(1)}=\mathbf{V}_1'\mathbf{V}_1=\mathbf{I}$ for ^{1}MCCA, and $\mathbf{X}_{(2)}'\mathbf{X}_{(2)}=\mathbf{V}_2'\mathbf{V}_2=\mathbf{I}$ for ^{2}MCCA. The formulation of (2.25) is consistent with the formulation of ^{c}MCCA in (2.5). The optimal solutions for $\mathbf{V}_1$ and $\mathbf{V}_2$ can be found by taking the p principal eigenvectors of respectively matrix $\mathbf{P}_1'\mathbf{P}_2\mathbf{P}_2'\mathbf{P}_1$ and $\mathbf{P}_2'\mathbf{P}_1\mathbf{P}_1'\mathbf{P}_2$. These eigenvectors are equal to respectively the p principal left and right singular vectors of matrix $\mathbf{P}_1'\mathbf{P}_2$, which we recognize from (2.23). The eigenvalues of $\mathbf{P}_1'\mathbf{P}_2\mathbf{P}_2'\mathbf{P}_1$ and

$\mathbf{P}_2'\mathbf{P}_1\mathbf{P}_1'\mathbf{P}_2$ are equal to the squared singular values of $\mathbf{P}_1'\mathbf{P}_2$ and therefore equal to the squared canonical correlations. For K=2 the solutions of both ^{c}MCCA and CCA are nested and the relation between the fit functions of ^{c}MCCA and CCA is given dimensionwise by $^c\text{MCCA}_{\text{Fit}}=1+(\text{CCA}_{\text{Fit}})^2$. The canonical variates of set 1 are given by the optimal $\mathbf{X}_{(1)}$ and the canonical variates of set 2 by the optimal $\mathbf{X}_{(2)}$, because $\mathbf{X}_{(1)}=\mathbf{Z}_1=\mathbf{P}_1\mathbf{V}_1$ and $\mathbf{X}_{(2)}=\mathbf{Z}_2=\mathbf{P}_2\mathbf{V}_2$, see (2.23).

2.3 Different types of filter

In the preceding sections we applied always the same filter for all sets. Only subspace restrictions introduced some asymmetry in the analysis. We now describe another kind of asymmetry in the analysis by combining different types of filters in one analysis. A set correlation filter is assigned to one set or group of sets and a set variance filter is assigned to another set or group of sets. In this way we introduce several additive hybrid methods. First we formulate Redundancy Analysis as an example with two sets. Next we give some generalizations of Redundancy Analysis for multiple sets.

2.3.1 Redundancy Analysis (RA)

The technical formulation of Redundancy Analysis (RA) can be found in Anderson (1951), who defined the model by imposing linear restrictions on regression coefficients or in other words by reducing the rank of the regression matrix. The model is also called the Reduced Rank Regression model. For a recent overview see Van der Leeden (1990). We begin with a description of Redundancy Analysis and subsequently we show how RA can be conceived of as a two set MFCA with different filters.

In RA as defined by Anderson (1951) we maximize the variance of the criterion set that is accounted for by the canonical variates of the predictor set. Some authors (e.g.,Van den Wollenberg, 1977) divide the variance accounted for by the number of criterion variables, which are also referred to as criteria. For each dimension r of the solution the variance accounted for is called the *redundancy* of the criteria. The sum of the redundancies is called the *overall redundancy*. We indicate the predictor set with c

and the criterion set with k and maximize the overall redundancy of the criteria as follows

$$\text{RA:} \qquad \text{Fit}(\mathbf{Z}_c) = \sum_{s=1}^{p} \eta^2_{(k)s} = \text{tr } \mathbf{Z}_c'\mathbf{H}_k\mathbf{H}_k'\mathbf{Z}_c$$

$$= \text{tr } \mathbf{V}_c\mathbf{P}_c'\mathbf{P}_k\mathbf{\Phi}^2_k\mathbf{P}_k'\mathbf{P}_c\mathbf{V}_c, \qquad (2.26)$$

where $\mathbf{Z}_c = \mathbf{H}_c\mathbf{T}_c = \mathbf{P}_c\mathbf{\Phi}_c\mathbf{Q}_c'\mathbf{T}_c = \mathbf{P}_c\mathbf{V}_c$
denote the canonical variates of the predictor set c with $\mathbf{Z}_c'\mathbf{Z}_c = \mathbf{V}_c'\mathbf{V}_c = \mathbf{I}$,
$\mathbf{H}_k = \mathbf{P}_k\mathbf{\Phi}_k\mathbf{Q}_k'$ denotes the SVD of the criterion set k
and $\eta^2_{(k)s}$ is the redundancy of the m_k criteria for dimension s.

By specifying the filters and subspace restrictions we formulate RA as a two sets MFCA and maximize for predictor set c and criterion set k

$$^c\text{RA:} \qquad \Omega_k(\mathbf{\Phi}^2_k) = \mathbf{\Phi}^2_k$$

$$\Omega_c(\mathbf{\Phi}^2_c) = \mathbf{I}, \text{ with } \mathbf{X} = \mathbf{P}_c\mathbf{P}_c'\mathbf{X}. \qquad \text{for } K=2 \quad (2.27)$$

In this way we denote the use of different filters in one analysis accompanied by subspace restrictions.

We still have to show that (2.27) does the same job as the ordinary formulation of RA. We substitute (2.27) with $\mathbf{P}_c'\mathbf{X} = \mathbf{V}_c$ in (2.1) and compare it with (2.26). We obtain the equality $^c\text{RA}_{\text{Fit}}=p+\text{RA}_{\text{Fit}}$. Therefore maximization of these two functions gives the same optimal canonical variates for the predictor set.

The top filter in (2.27) is an identity filter and represents the set variance part of this additive hybrid method. The bottom filter is a constant filter and represents the set correlation part. As indicated in section 2.1.1 the subspace restriction in ^{c}RA (2.27) can be simulated by introducing a very large weight in the filter of predictor set c.

2.3.2 Multiset Redundancy Analysis (MRA)

Generalizations of RA for multiple sets can be formulated by combining two different types of filters in the following way

$$^{c}\text{MRA:} \qquad \Omega_k(\Phi_k^2) = \Phi_k^2 w_k^{-1}$$

$$\Omega_c(\Phi_c^2) = \mathbb{I}, \text{ with } \mathbb{X}_c = \mathbb{P}_c\mathbb{P}_c{'}\mathbb{X}_c, \qquad \forall k \neq c \qquad (2.28)$$

where c denotes the predictor set and
$w_1,\ldots,w_k,\ldots,w_K$ denote fixed balancing constants for criterion set k $\forall k \neq c$

With this function we maximize the variance of the criterion sets that is accounted for by linear combinations of the predictor set. The choice of the balancing constants is discussed extensively in section 2.2.1.

2.3.3 Multiset MIMIC method (MMIMIC)

The Multiple effect Indicators for Multiple Causes (MIMIC) model (Hauser & Goldberger, 1971) is basically a two sets model, where one set of variables, the input set, influences another set of variables, the output set. Other names for the input variables are exogenous or independent variables and for the output variables endogenous or dependent variables. The influence of the input set on the output set is mediated by unobserved latent variables.

We define the MIMIC method generalized for multiple sets. In addition we give the fit function for the ordinary MIMIC method, which is a special two sets case. The Multiset MIMIC (MMIMIC) method resembles the MRA method. The subspace restrictions are omitted compared to MRA and there are several input sets instead of one predictor set. We obtain

$$\text{MMIMIC:} \qquad \Omega_k(\Phi_k^2) = \Phi_k^2 w_k^{-1} \qquad \text{for} \quad l = 1,\ldots,L \,,$$

$$\Omega_l(\Phi_l^2) = \mathbb{I}, \qquad k = (L+1),\ldots,K \; L < K \qquad (2.29)$$

where $1,\ldots,l,\ldots,L$ denote the input sets and
$w_{L+1},\ldots,w_k,\ldots,w_K$ denote fixed balancing constants for output set k.

With this method we mediate the influence of the input sets on the output sets by common latent variables $\mathbb{X}$. For w_k=1 $\forall k$ the MMIMIC method is one of the generalizations of RA for multiple sets suggested by Van de Geer (1984).

The ordinary MIMIC solution is found by inserting filter (2.29) with $L = 1, K = 2$ and $w_2 = 1$ in (2.1). We maximize for input set c and output set k

$$\text{MIMIC:}\quad \text{Fit}(\mathbf{X}) = \text{tr } \mathbf{X}'\mathbf{P}_c\mathbf{P}_c'\mathbf{X} + \text{tr } \mathbf{X}'\mathbf{P}_k\Phi_k^2\mathbf{P}_k'\mathbf{X}. \tag{2.30}$$

The MIMIC fit function is the same as the reformulated reduced rank regression function described by De Leeuw & Bijleveld (1987) and Bijleveld (1989). In fact they create a family of solutions by introducing a weight α^2 for input set c, tr α^2 $\mathbf{X}'\mathbf{P}_c\mathbf{P}_c'\mathbf{X}$. For the limiting case α=0 they prove that (2.30) is equal to principal component analysis of the output variables, which can be easily verified by omitting the left part in (2.30). In this way the set correlation part disappears and only the set variance part remains. For $\alpha \rightarrow \infty$ they prove that (2.30) is equal to RA, which can be understood by realizing that the left part has an absolute maximum of p if the common latent variables are in the space of the input variables, so if $\mathbf{X}=\mathbf{P}_c\mathbf{P}_c'\mathbf{X}$. After insertion of $\mathbf{X}=\mathbf{P}_c\mathbf{P}_c'\mathbf{X}=\mathbf{P}_c\mathbf{V}_c$ for $\mathbf{X}$ we recognize in the right part of (2.30) the formulation of RA in (2.26). In Van der Burg (1988) method (2.30) is described in a comparable way as a two sets generalization of RA by releasing the subspace restrictions of the RA predictor set.

2.4 Discrete compound filters

In this section we discuss the possibility of constructing compound filters by combining two filters in one, separated by a threshold value. We show how *reduced rank preprocessing* steps can be incorporated in the analysis by applying this kind of filters. The concept is illustrated by elaborating the practice of replacing a set of variables by an approximation of lower rank in a first step, followed by an analysis of this reduced rank approximation in a second step. Usually these methods are *two-step hybrid methods* fitting a *set variance* function in the first step and a *set correlation* function in the second step. In chapter 1 we classified these methods as sequential hybrid methods.

2.4.1 Two-step hybrid methods

Two-step hybrid methods usually combine reduced rank preprocessing with CCA or from CCA derived methods, like Discriminant Analysis (See Gittins, 1985). The purpose is to eliminate the possibility of finding CCA solutions with very small variance accounted for by the canonical variates. This is achieved by literally eliminating from each set the part of the information that projects on singular vectors with small singular values. In other words, we are reducing the rank of the set k by replacing $\mathbf{H}_k = \mathbf{P}_k\Phi_k\mathbf{Q}_k'$ by $\underline{\mathbf{H}}_k = \mathbf{P}_k\underline{\Phi}_k\mathbf{Q}_k'$, where $\underline{\Phi}_k$ denotes a diagonal matrix equal to Φ_k ,but with singular values below a certain threshold value made equal to zero. After this preprocessing step CCA is performed in a second step on the matrices $\underline{\mathbf{H}}_k$.

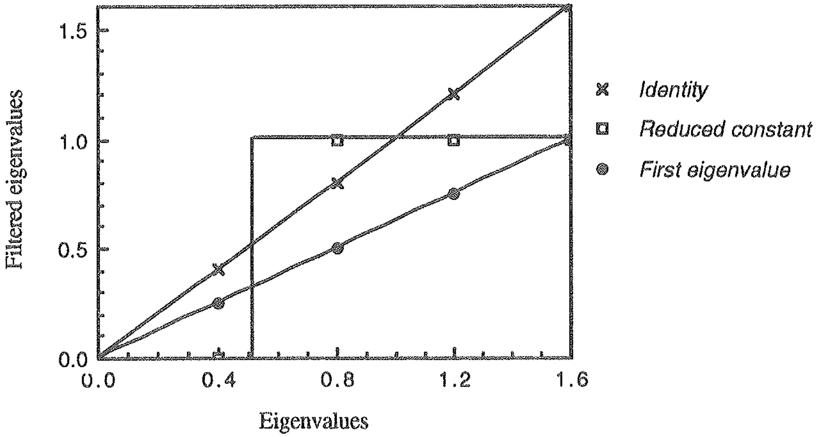


Figure 2.4 *Reduced constant filter.*

The two step hybrid method for CCA described above can be compressed in one MFCA step by filtering the eigenvalues of set k in such a way that all eigenvalues below a certain threshold become equal to zero and above this threshold become equal to one. The resulting filter is represented in figure 2.4 as the *reduced constant* filter, together with the *identity* filter and the *first eigenvalue* filter. For the definitions of the last two filters see (2.2) and (2.3). The *eigenvalues* are on the horizontal axis and the filtered eigenvalues are on the vertical axis. The identity filter is given by a sloping line with an arbitrary chosen largest eigenvalue of 1.6. The reduced constant filter consists of two parts, separated by a threshold. In figure 2.4 we took a value for the threshold of $0.33\times\phi_1^2$. Eigenvalues beneath this threshold are transformed to 0 and above this threshold to 1. The left part in the filter of this hybrid method approximates the identity filter which is a set variance filter, and the right part is equal to the constant filter which is a set correlation filter. In this way the left part eliminates the small variances and the right part gives the relevant spatial information. From this point of view we can make a less drastically *pseudo reduced constant* version of the reduced constant filter by not eliminating the small variances, but by taking the left part of the filter equal to the identity, trace or first eigenvalue filter. The non eliminating approach illustrates the hybrid nature of the reduced constant filter more clearly, because the set variance part is not approximated roughly, but presented exactly. This kind of filter is called *pseudo* reduced constant, because a reduced constant filter always involves a dimension reduction of the data, which is not the case for a pseudo reduced constant filter. We have already applied the principles of the pseudo reduced constant filter in the definition of the information span I_k (2.17).

2.5 Continuous compound filters

A major drawback in the application of two-step hybrid methods is the arbitrariness of the threshold for *selecting* principal components. There are many different methods to find a reasonable value for the threshold. This creates the problem of choosing the appropriate selection method, maybe even different methods for different sets. It is possible to approach this problem in another way by replacing the discrete two-step reduced constant filter by an one-step continuous filter that approximates the (pseudo) reduced constant filter without a threshold. For a good approximation we need some nonlinear continuous function, that is close to one for high eigenvalues and rapidly decreases to zero for very small eigenvalues. In other words high eigenvalues must approximate the set correlation property and low eigenvalues the set variance property of the hybrid method. Two such continuous compound filters are described in the next sections. In section 2.5.1 we propose a multiset generalization of RR and derive an appropriate *ridge* filter. This continuous compound ridge filter shows that Ridge Regression (RR) is a weighted hybrid method. In section 2.5.2 we define Fixed Set Component Analysis (FSCA) by specifying a quadratic first eigenvalue filter. This method brings us close to the next chapter, because there we discuss the adjusted method of Set Component Analysis (SCA) by applying a free quadratic filter that results in the maximization of the sum of squared correlations of adjusted set variates.

2.5.1 *Multiset Ridge Regression (MRR)*

In Ridge Regression (Hoerl & Kennard, 1970, Golub & Van Loan, 1990, p.565) a loss function minimizes for predictor set $\mathbf{H}_c$ and criterion variable $\mathbf{h}_k$

$$\text{RR:} \qquad \text{Loss}(\mathbf{t}_c) = (\mathbf{h}_k - \mathbf{H}_c\mathbf{t}_c)'(\mathbf{h}_k - \mathbf{H}_c\mathbf{t}_c) + \nu_c\mathbf{t}_c'\mathbf{t}_c \tag{2.31}$$

with $\nu_c \geq 0$.

We propose the following multiset generalization of RR for predictor sets c and one unknown common latent criterion variable $\mathbf{x}$ by minimizing

$$\text{MRR}_{p=1}\text{: Loss}(\mathbf{x},\mathbf{t}_c) = \sum_{c=1}^{K} (\mathbf{x} - \mathbf{H}_c\mathbf{t}_c)'(\mathbf{x} - \mathbf{H}_c\mathbf{t}_c) + \nu_c\mathbf{t}_c'\mathbf{t}_c, \tag{2.32}$$

with $\mathbf{x}'\mathbf{x}=1$ and $\nu_c \geq 0 \; \forall c$.

To show the relation of MRR with MFCA we need a description of MRR with only the $\mathbf{x}$ as unknown parameters. Therefore we want to find suboptimal values for the MRR weights $\mathbf{t}_c$, which are a function of $\mathbf{x}$. Analogous to the preceding sections the SVD for set c is given by $\mathbf{H}_c = \mathbf{P}_c\Phi_c\mathbf{Q}_c'$, where $\mathbf{P}_c$ $(n \times p_c)$ and $\mathbf{Q}_c$ $(m_c \times p_c)$ denote orthonormal singular vector matrices and Φ_c denotes a diagonal matrix with p_c non-zero singular values in descending order. The MRR weights $\mathbf{t}_c$ can be derived from weights $\mathbf{v}_c$ by $\mathbf{t}_c = \mathbf{Q}_c\Phi_c^{-1}\mathbf{v}_c \;\; \forall c$. We insert $\mathbf{Q}_c\Phi_c^{-1}\mathbf{v}_c$ for $\mathbf{t}_c$ in (2.32) and compute a conditional minimum for $\text{MRR}_{p=1}$ with $\mathbf{x}$ fixed by minimizing

$$\begin{aligned}\text{Loss}(\mathbf{x},c_c,\mathbf{v}_c) &= (\mathbf{x} - \mathbf{P}_c\mathbf{v}_c)'(\mathbf{x} - \mathbf{P}_c\mathbf{v}_c) + \nu_c\mathbf{v}_c'\Phi_c^{-2}\mathbf{v}_c + c_c \\ &= \mathbf{x}'\mathbf{x} - 2\mathbf{x}'\mathbf{P}_c\mathbf{v}_c + \mathbf{v}_c'\mathbf{v}_c + \nu_c\mathbf{v}_c'\Phi_c^{-2}\mathbf{v}_c + c_c \\ &= \mathbf{x}'\mathbf{x} - 2\mathbf{x}'\mathbf{P}_c\mathbf{v}_c + \mathbf{v}_c'(\mathbf{I} + \nu_c\Phi_c^{-2})\mathbf{v}_c + c_c \\ &= \text{SSQ}((\mathbf{I} + \nu_c\Phi_c^{-2})^{-1/2}\mathbf{P}_c'\mathbf{x} - (\mathbf{I} + \nu_c\Phi_c^{-2})^{1/2}\mathbf{v}_c) + \underline{c}_c \qquad \forall c\end{aligned} \tag{2.33}$$

where SSQ(M) denotes the sum of squares of the elements of M,
$\mathbf{x}$ denotes the fixed common latent variable $\mathbf{x}$,
and c_c,$\underline{c}_c$ denotes the sum for all other fixed parameters.

The minimum of (2.33) is reached for

$$\mathbf{v}_c \quad = (\mathbf{I} + \nu_c \Phi_c^{-2})^{-1/2} \mathbf{P}_c{}'\mathbf{x}. \qquad \forall c \quad (2.34)$$

After insertion of $\mathbf{Q}_c \Phi_c^{-1} \mathbf{v}_c$ for the MRR weights $\mathbf{t}_c$ in (2.32) with $\mathbf{v}_c$ according to (2.34) we minimize

$$\text{MRR}_{p=1}\text{: Loss}(\mathbf{x}) = \sum_{c=1}^{K} \mathbf{x}\, '\mathbf{P}_c (\mathbf{I} + \nu_c \Phi_c^{-2})^{-1} \mathbf{P}_c{}'\mathbf{x}, \qquad (2.35)$$

with $\mathbf{x}'\mathbf{x}=1$ and $\nu_c \geq 0\ \forall c$.

After minimization of (2.35) the optimal MRR variates are

$$\mathbf{H}_c \mathbf{t}_c = \mathbf{P}_c \mathbf{v}_c = \mathbf{P}_c (\mathbf{I} + \nu_c \Phi_c^{-2})^{-1/2} \mathbf{P}_c{}'\mathbf{x}. \qquad \forall c \quad (2.36)$$

From (2.35) we extract the appropriate MFCA *ridge* filter for specifying Multiset Ridge Regression,

$$\text{MRR:} \qquad \Omega_c(\Phi_c^2) = (\mathbf{I} + \nu_c \Phi_c^{-2})^{-1} \qquad \forall c \quad (2.37)$$

with $\nu_c \geq 0\ \forall c$.

Because $\nu_c \geq 0$, we always have $\mathbf{0}\mathbf{0}' \leq (\mathbf{I} + \nu_c \Phi_c^{-2})^{-1} \leq \mathbf{I}$, with $\mathbf{0}$ a column vector of appropriate size with elements 0. For K=p=1 and $\mathbf{x}$ equal to criterion variable $\mathbf{h}_k$, MRR is equal to ordinary ridge regression and the optimal RR variate is computed with (2.36) after inserting $\mathbf{h}_k$ for $\mathbf{x}$. The ridge filter in (2.37) contains a set correlation filter $\mathbf{I}$ added to a weighted set variance filter $\nu_c \Phi_c^{-2}$. By substituting extreme values for ν_c we know the corresponding extreme fit functions. If ν_c is very large the diagonal elements of $(\mathbf{I} + \nu_c \Phi_c^{-2})^{-1}$ are almost equal to $\Phi_c^2 \nu_c^{-1}$ and therefore MRR is almost equal to MPCA, with balancing constants w_c equal to ν_c. If ν_c=0, MRR is equal to MCCA. In figure 2.5 we give the ridge filter for ν_c=0, $\phi_{1c}^2/9$, $\phi_{1c}^2/3$, ϕ_{1c}^2 and $3\phi_{1c}^2$ in order to demonstrate the weighted hybrid nature of MRR graphically. The representation of the horizontal axis is more general than in figure 2.4, because the *eigenvalue quotient* gives the eigenvalues divided by the largest eigenvalue. For ν_c=0 the ridge filter is equal to the constant filter.

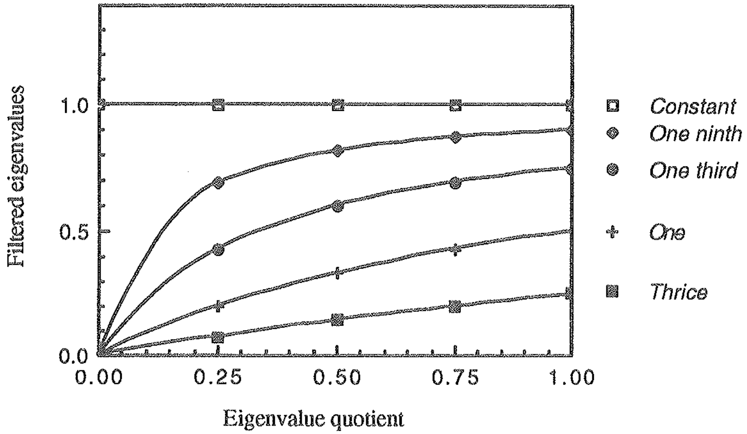


Figure 2.5 *Ridge filter.*

As v_c increases the curve transforms more and more into an oblique line that represents a rescaled identity filter. In this way the ridge filter defines a whole series of subfilters that ranges from set correlation to set variance. The extremes of this range reveal the hybrid nature of the ridge regression method. The curve of the one ninth ridge subfilter bears much resemblance with the curve of the filter formulated in the next section for Fixed Set Component Analysis (FSCA). The main difference in the specifications of the filters is that for the ridge filter we have a selection problem. We have to choose the constants v_c a priori or by some of the manifold data based methods. This problem does not occur for the FSCA method.

2.5.2 Fixed Set Component Analysis (FSCA)

In FSCA the term 'Fixed' indicates the fact that we actually use a fixed form of the SCA filter discussed in the next chapter. We maximize (2.1) with a *quadratic first eigenvalue* filter

$$\text{FSCA:} \qquad \Omega_k(\Phi_k^2) = \mathbb{I} - (\mathbb{I} - \Phi_k^2\phi_{1k}^{-2})^2. \qquad \forall k \quad (2.38)$$

Because ϕ_{1k}^2 is the largest eigenvalue of Φ_k^2, we always have $\mathbb{O}\mathbb{O}' \leq \Phi_k^2\phi_{1k}^{-2} \leq \mathbb{I}$. The quadratic first eigenvalue filter in (2.38) is a special (ν=1) subfilter of the weighted hybrid filter $\mathbb{I} - \nu(\mathbb{I} - \Phi_k^2\phi_{1k}^{-2})^2$. This weighted FSCA filter contains a weighted set variance filter $\nu(\mathbb{I} - \Phi_k^2\phi_{1k}^{-2})^2$ subtracted from a set correlation filter $\mathbb{I}$. By substituting extreme values for v_c we know the corresponding extreme fit functions. If v is very large the weighted FSCA filter transforms all diagonal elements of Φ_k^2 with values ϕ_{1k}^2 to the value 1. All diagonal elements of Φ_k^2 smaller than ϕ_{1k}^2 will be highly negative and the optimal solution of X for weighted FSCA will avoid eigenvectors with smaller eigenvalues than ϕ_{1k}^2. If v=0, weighted FSCA is equal to MCCA. In figure 2.6 we represent the quadratic first eigenvalue filter to show how this filter approximates the reduced constant filter by a very simple polynomial filter.

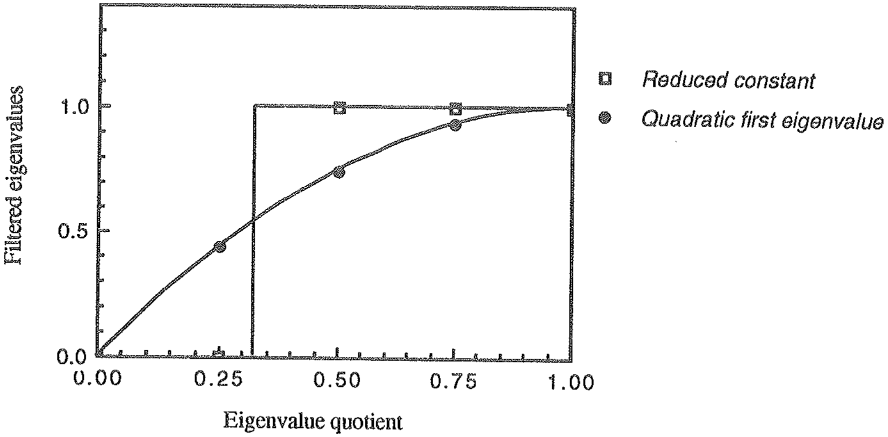


Figure 2.6 *Quadratic first eigenvalue filter.*

It is clear that the approximation of the reduced constant filter is rather crude, but we have to bear in mind that the location of the threshold is variable. We require for a general approximation that the filtered eigenvalues are near one at the right side and go down steeply at the left side, like the curve of the *one ninth* ridge subfilter. Other filters could be defined, like growth curve or S-shape filters, but the simplicity of the quadratic first eigenvalue filter makes it attractive.

The continuous compound filters discussed previously approximate a set correlation part for high eigenvalues and a set variance part for low eigenvalues. We are integrating two corresponding separate fit functions in a continuous way and therefore dealing with a hybrid method. In the next two chapters we will discuss adjusted methods that maximize one fit function modified by set variance or set correlation constraints.

Chapter 3

SET CORRELATION WITH SET VARIANCE CONSTRAINTS

Set Component Analysis is described from several points of view. (1) The method integrates a set correlation and a set variance part by maximizing the sum of squared set correlations and adjusting the set variates with set variance constraints. (2) SCA is identical to Multiset CCA with proportionality restrictions on the variable weights. (3) By defining a free quadratic filter, SCA is related with the filter theory formulated in the previous chapter. We conclude this chapter by indicating relations with other methods and presenting a simulation study of INDSCAL compared with SCA. The relation between INDSCAL and SCA is established by proposing and fitting a new model, the *INDRES model.*

Introduction

Set Component Analysis (Nierop, 1989, 1993) is an adjusted method. It maximizes exactness of prediction with special constraints to improve stability. The main fit function is the sum of squared set correlations, and the secondary set variance constraint enables a local improvement on the variance accounted for.

Maximization and improvement are well-known in the context of multivariate optimization problems. Very often there exists no analytical method to find an optimal solution. In that case a monotone convergent algorithm is constructed that improves the value of some target function in successive steps until a local or global maximum is reached. The improvement steps can be derived by several methods like partitioning the function in several quadratic parts (Huygens principle), determination of the first derivative, or majorization (de Leeuw & Heiser, 1980). By applying this knowledge it is possible to integrate two different functions by combining the maximization of a main fit function with the improvement constraint of an *adjusting* function. How this combination can be made is illustrated in section 3.1 for two functions: squared canonical correlations and variance accounted for. The resulting method is Set Component Analysis (SCA).

In subsequent sections we will discuss some characteristics of SCA. In section 3.2 we explicitly show the relation between the variable weights and the structure correlations. In section 3.3 we relate the SCA method to the filter framework outlined in chapter 2. In section 3.4 we explain the relation of SCA with FSCA, MPCA, MCCA and INDSCAL. The relation between INDSCAL and SCA is established by proposing and fitting the INDRES model. In the closing section we compare the properties of INDSCAL and SCA in a simulation study.

3.1 Set Component Analysis

For the construction of the SCA method we integrate the *maximization* of the sum of squared canonical correlations with the *improvement* of variance accounted for. For the maximization of the sum of squared canonical correlations we use the multiset MCCA method described in (2.20)

$$\text{MCCA:}\quad \mathrm{Fit}(\mathbf{X},\mathbf{Z}_1,..,\mathbf{Z}_k,..,\mathbf{Z}_K) = \sum_{s=1}^{p}\sum_{k=1}^{K} (\mathbf{x}_s'\mathbf{z}_{(k)s})^2,$$

with orthonormal latent variables $\mathbf{x}_s$ and unit normalized canonical variates $\mathbf{z}_{(k)s}$. The adjusting function for the improvement of variance accounted for is

$$\mathrm{VAF}(\mathbf{z}_{(k)s}) = \mathbf{z}_{(k)s}'\mathbf{H}_k\mathbf{H}_k'\mathbf{z}_{(k)s} = \mathbf{z}_{(k)s}'\mathbf{S}_k\mathbf{z}_{(k)s}, \qquad \forall k,s \quad (3.1)$$

where $\mathbf{S}_k$=$\mathbf{H}_k\mathbf{H}_k'$. For the improvement of (3.1) we take one step of the Power Method (Wilkinson, 1965) and normalize to unit sum of squares

$$\overset{t+1}{\mathbf{z}_{(k)s}} = \mathbf{S}_k\overset{t}{\mathbf{z}_{(k)s}}(\overset{t}{\mathbf{z}_{(k)s}}'\mathbf{S}_k\mathbf{S}_k\overset{t}{\mathbf{z}_{(k)s}})^{-1/2}, \qquad \forall k,s \quad (3.2)$$

and define $\overset{t}{\mathbf{z}_{(k)s}}$ as the MCCA *canonical variate* and the adjusted canonical variate $\overset{t+1}{\mathbf{z}_{(k)s}}$ as the (SCA) *set variate*. In section 2.2.6 we showed that the canonical variate must be equal to

$$\overset{t}{\mathbf{z}_{(k)s}} = \mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s(\mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s)^{-1/2}. \qquad \forall k,s \quad (3.3)$$

Substitution of (3.3) in (3.2) gives the SCA set variate

$$\mathbf{z}_{(k)s} = \mathbf{S}_k\mathbf{x}_s(\mathbf{x}_s'\mathbf{S}_k\mathbf{S}_k\mathbf{x}_s)^{-1/2}. \qquad \forall k,s \quad (3.4)$$

The set variates can be conceived as a weighted projection of $\mathbf{x}_s$ on to $\mathbf{P}_k$ and the canonical variates in (3.3) as an unweighted projection. Substitution of $\mathbf{S}_k=\mathbf{P}_k\Phi_k^2\mathbf{P}_k'$ in (3.4) shows that the eigenvalues Φ_k^2 are the projection weights. Rather than having (3.4) as a side product, we can constrain the variates $\mathbf{Z}_k$ in the MCCA fit function (2.20) so that they satisfy (3.4). This is achieved by inserting (3.4) in (2.20) and results in the SCA fit function that maximizes the sum of the squared *set correlations*

$$\text{SCA:}\quad \text{Fit}(\mathbf{x}_s) = \sum_{s=1}^{p}\sum_{k=1}^{K} \rho^2_{(\mathbf{x}_s;\mathbf{S}_k\mathbf{x}_s)} = \sum_{s=1}^{p}\sum_{k=1}^{K} (\mathbf{x}_s'\mathbf{z}_{(k)s})^2 = \sum_{s=1}^{p}\sum_{k=1}^{K} \frac{(\mathbf{x}_s'\mathbf{S}_k\mathbf{x}_s)^2}{\mathbf{x}_s'\mathbf{S}_k\mathbf{S}_k\mathbf{x}_s}\,, \tag{3.5}$$

where $\rho_{(\mathbf{x}_s;\mathbf{S}_k\mathbf{x}_s)}$ denotes the correlation between $\mathbf{x}_s$ and $\mathbf{S}_k\mathbf{x}_s$,

$_n\mathbf{X}_p = (\mathbf{x}_1,\ldots,\mathbf{x}_s,\ldots,\mathbf{x}_p)$ denote common latent variables with $\mathbf{X}'\mathbf{X}=\mathbf{I}$

and $\mathbf{z}_{(k)s} = \mathbf{S}_k\mathbf{x}_s(\mathbf{x}_s'\mathbf{S}_k\mathbf{S}_k\mathbf{x}_s)^{-1/2}$ denote the set variates with $\mathbf{S}_k=\mathbf{H}_k\mathbf{H}_k'\quad \forall\, k,s$.

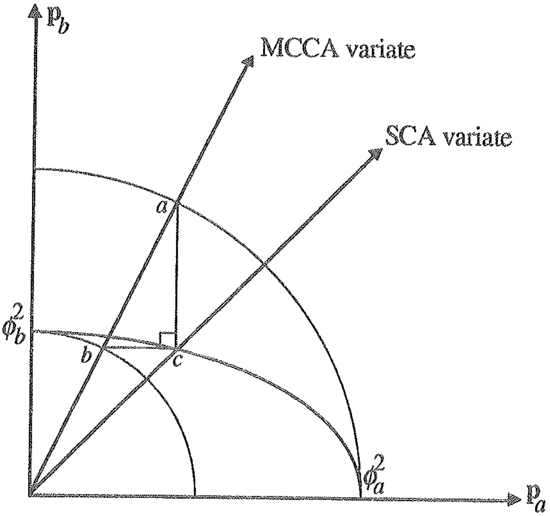


Figure 3.1 *Improvement of variance accounted for.*

In figure 3.1 we give a geometric construction of the SCA set variate derived from the position of the MCCA canonical variate. The 2 dimensional case is sufficient to illustrate the Power Method, because more dimensional cases can be described by analogous successive plane rotations. All vectors in figure 3.1 are unit normalized. We start with some known MCCA canonical variate located in the plane of the eigenvectors $\mathbf{p}_a$ and $\mathbf{p}_b$ with eigenvalues $\phi_a^2 > \phi_b^2$. The variance accounted for by some SCA variate has a maximum ϕ_a^2, when the variate is $\mathbf{p}_a$ and a minimum ϕ_b^2, when the variate is $\mathbf{p}_b$. The position of the SCA variate in figure 3.1 is constructed with the intersection points a and b. They are the intersection points of the MCCA variate and the circles described by the radii ϕ_a^2 and ϕ_b^2, respectively. The intersection point c of a vertical line through a and a horizontal line through b is located on the SCA variate and therefore fixes its direction. It is clear that the variance accounted for by the SCA variate is higher than the variance accounted for by the MCCA variate, because it is closer to the maximum direction in this plane: $\mathbf{p}_a$. The construction method we have applied is just one of the many methods for constructing an ellipse. The points c of this ellipse can be found by varying the starting position of the MCCA variate. It should be noticed that the direction of the SCA variate is independent of the normalization of the MCCA variate and the total sum of the eigenvalues. In other words the SCA solution is scale free with respect to the normalization of sets.

3.2 Variable weights proportional to structure correlations

In this section we want to emphasize an interesting property of SCA compared to MCCA. In both fit functions we maximize the sum of the squared correlations of the unit normalized variates $\mathbf{z}_{(k)s}$ with the common latent variables $\mathbf{x}_s$. If we compare SCA in (3.5) with MCCA in (2.20), the only difference we observe is the definition of the weighted sum of variables $\mathbf{z}_{(k)s}$. For SCA we have

$$\begin{aligned} \mathbf{z}_{(k)s} &= \mathbf{S}_k\mathbf{x}_s(\mathbf{x}_s'\mathbf{S}_k\mathbf{S}_k\mathbf{x}_s)^{-1/2} \\ &= \mathbf{H}_k\{\mathbf{H}_k'\mathbf{x}_s(\mathbf{x}_s'\mathbf{S}_k\mathbf{S}_k\mathbf{x}_s)^{-1/2}\} = \mathbf{H}_k\mathbf{t}_{(k)s}. \qquad \forall k,s \end{aligned}$$

The SCA weights $\mathbf{t}_{(k)s}=\mathbf{H}_k'\mathbf{x}_s(\mathbf{x}_s'\mathbf{S}_k\mathbf{S}_k\mathbf{x}_s)^{-1/2}$ of the variables $\mathbf{H}_k$ are for each set k proportional to the structure correlations $\mathbf{H}_k'\mathbf{x}_s$. In MCCA we do not have these restrictions for the weights $\mathbf{t}_{(k)s}$. From this point of view SCA can be defined as MCCA with proportionality restrictions on the variable weights.This definition of SCA is simpler than the first definition of SCA in the previous section.

Nevertheless we preferred to define SCA first as MCCA with local variance improvement constraints, because we understand the predictive properties of SCA better with this definition.

The proportionality restrictions on the variable weights facilitate the interpretation of the weights and structure correlations in SCA. In MCCA the weights and structure correlations of the same variables can diverge to a large extent, which makes a consistent interpretation difficult. In this case interpretation is usually confined to the structure correlations.

3.3 A filter view on SCA

In order to relate the SCA method with the filter theory discussed in chapter 2, we reformulate the fit function (3.5) by first introducing regression weights and secondly balancing factors. The concept of balancing is introduced because we also want to show in section 3.4.2 that SCA can be conceived as maximizing weighted variance accounted for. Furthermore, the reformulated SCA fit function has a computational advantage. It is simpler to derive an algorithm to find the SCA solution with this alternative loss function than with the formulation of (3.5), because the complicated function of $\mathbf{x}_s$ in the denominator will disappear. In chapter 6 we describe a monotone convergent algorithm for SCA.

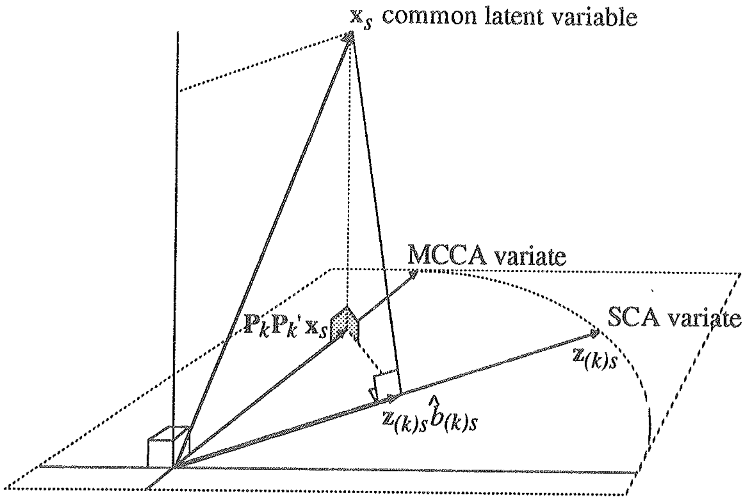

$\mathbf{x}_s$ common latent variable
MCCA variate
SCA variate
$\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s$
$\mathbf{z}_{(k)s}$
$\mathbf{z}_{(k)s}\hat{b}_{(k)s}$

Figure 3.2 *Introduction of regression weights.*

For each set k the squared correlation between (unit normalized) $\mathbf{x}_s$ and $\mathbf{z}_{(k)s}$ is equal to the squared length of the projection of $\mathbf{x}_s$ on to $\mathbf{z}_{(k)s}$. In figure 3.2 this projected vector is given by $\mathbf{z}_{(k)s}\hat{b}_{(k)s}$, with $\hat{b}_{(k)s}=\mathbf{x}_s'\mathbf{z}_{(k)s}$. This implies that instead of maximizing for each set the squared projection length $(\mathbf{x}_s'\mathbf{z}_{(k)s})^2$, we can also maximize

$\mathbf{x}_s'\mathbf{x}_s$ – (the squared distance of $\mathbf{x}_s$ to the SCA variate),

or we can maximize

$\mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s$ – (the squared distance of $\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s$ to the SCA variate).

In other words we have

$$\text{SCA:}\quad \text{Fit}(\mathbf{x}_s,b_{(k)s}) = \sum_{s=1}^{p}\sum_{k=1}^{K} \mathbf{x}_s'\mathbf{x}_s - (\mathbf{x}_s - \mathbf{z}_{(k)s}b_{(k)s})'(\mathbf{x}_s - \mathbf{z}_{(k)s}b_{(k)s})$$

$$= \sum_{s=1}^{p}\sum_{k=1}^{K} \mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s - (\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s - \mathbf{z}_{(k)s}b_{(k)s})'(\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s - \mathbf{z}_{(k)s}b_{(k)s}). \qquad (3.6)$$

In figure 3.2 this Pythagorean property can be verified. By fixing the $\mathbf{X}$ and setting the first derivative equal to zero we find suboptimal regression weights $\hat{b}_{(k)s}$. As we would expect we find $\hat{b}_{(k)s}=\mathbf{x}_s'\mathbf{z}_{(k)s}$ and after substitution in (3.6) we obtain again the sum of all squared projection lengths $(\mathbf{x}_s'\mathbf{z}_{(k)s})^2$.

In formula (3.4) we saw that $\mathbf{z}_{(k)s}$ is proportional to $\mathbf{S}_k\mathbf{x}_s$. Therefore by definition we can replace $\mathbf{z}_{(k)s}$ in (3.6) by $\mathbf{S}_k\mathbf{x}_s$ and the weights $b_{(k)s}$ by the reciprocal values of balancing factors $w_{(k)s}$. Recapitulating, the reformulated SCA method maximizes

$$\text{SCA:}\quad \text{Fit}(\mathbf{x}_s,w_{(k)s}) = \sum_{s=1}^{p}\sum_{k=1}^{K} \mathbf{x}_s'\mathbf{x}_s - (\mathbf{x}_s - \mathbf{S}_k\mathbf{x}_s w_{(k)s}^{-1})'(\mathbf{x}_s - \mathbf{S}_k\mathbf{x}_s w_{(k)s}^{-1})$$

$$= \sum_{s=1}^{p}\sum_{k=1}^{K} \mathbf{x}_s'\mathbf{P}_k\{\mathbf{I} - (\mathbf{I} - \Phi_k^2 w_{(k)s}^{-1})^2\}\mathbf{P}_k'\mathbf{x}_s, \qquad (3.7)$$

where $w_{(1)s},\ldots,w_{(k)s},\ldots,w_{(K)s}$ denote free balancing factors for set k and dimension s, and the remaining parameters are defined and normalized as usual. The balancing factors $w_{(k)s}$ in (3.7) are free in the sense that the *optimal* $w_{(k)s}$ have to be found by maximizing the SCA fit function.

The formulation of (3.7) shows that SCA finds some optimal transformation of the eigenvalues of each set k and therefore can be defined as a MFCA method by specifying the appropriate filter. The MFCA filter for SCA is a *free quadratic* filter, which is defined for each set k and each dimension s as

$$\text{SCA:} \qquad \Omega_{(k)s}(\Phi_k^2) = \mathbf{I} - (\mathbf{I} - \Phi_k^2 w_{(k)s}^{-1})^2, \tag{3.8}$$

where $w_{(1)s},\ldots,w_{(k)s},\ldots,w_{(K)s}$ denote free balancing factors for set k and dimension s.

After substitution of (3.8) in (2.1) we must realize that we have introduced in the MFCA(X) function extra unknown parameters by incorporating the free balancing factors in the filters.

3.4 Relations of SCA with other methods

SCA has many connections with other methods. The following sections elaborate on relations with FSCA, MPCA, MCCA and INDSCAL. The reformulation of SCA with the Directed Correlations method and relations with some PLS methods are given in the last two sections of chapter 5.

3.4.1 Relation with FSCA

In section 2.5.2 on FSCA we fixed the balancing factors of SCA equal to ϕ_{1k}^2. By fixing the $\mathbf{X}$ in (3.7) and setting the first derivative equal to zero we find *suboptimal* balancing factors, which are a function of $\mathbf{X}$. We denote these suboptimal balancing factors with $\hat{w}_{(k)s}$. The suboptimal balancing factors are equal to the reciprocal regression weights

$$\hat{w}_{(k)s} = \frac{\mathbf{x}_s'\mathbf{S}_k\mathbf{S}_k\mathbf{x}_s}{\mathbf{x}_s'\mathbf{S}_k\mathbf{x}_s} = \frac{\mathbf{t}_{(k)s}'\mathbf{H}_k'\mathbf{H}_k\mathbf{t}_{(k)s}}{\mathbf{t}_{(k)s}'\mathbf{t}_{(k)s}} = \frac{\mathbf{t}_{(k)s}'\mathbf{Q}_k\Phi_k^2\mathbf{Q}_k'\mathbf{t}_{(k)s}}{\mathbf{t}_{(k)s}'\mathbf{t}_{(k)s}}, \quad \forall k,s \tag{3.9}$$

with $\mathbf{t}_{(k)s} = \mathbf{H}_k'\mathbf{x}_s \quad \forall k,s.$

In other words the suboptimal balancing factors are equal to the variance of $\mathbf{H}_k$ accounted for by the proportional regression weights $\mathbf{t}_{(k)s}=\mathbf{H}_k'\mathbf{x}_s$. The upper bound of $\hat{w}_{(k)s}$ is equal to the largest eigenvalue of matrix Φ_k^2, which is the first eigenvalue ϕ_{1k}^2. In section 2.5.2 on FSCA we fixed the balancing factors equal to this upper bound. The lower bound of $\hat{w}_{(k)s}$ is almost zero, because we have defined Φ_k^2 with only non-zero eigenvalues of $\mathbf{H}_k$. In summary, we have

$$0 < \hat{w}_{(k)s} \leq \phi_{1k}^2. \qquad \forall k,s \qquad (3.10)$$

Analogous to the presentation of the quadratic first eigenvalue filter in figure 2.6 we represent in figure 3.3 the *free quadratic* filter for $\hat{w}_{(k)s}=0.33\times\phi_{1k}^2$ and for the upper bound $\hat{w}_{(k)s}=\phi_{1k}^2$.

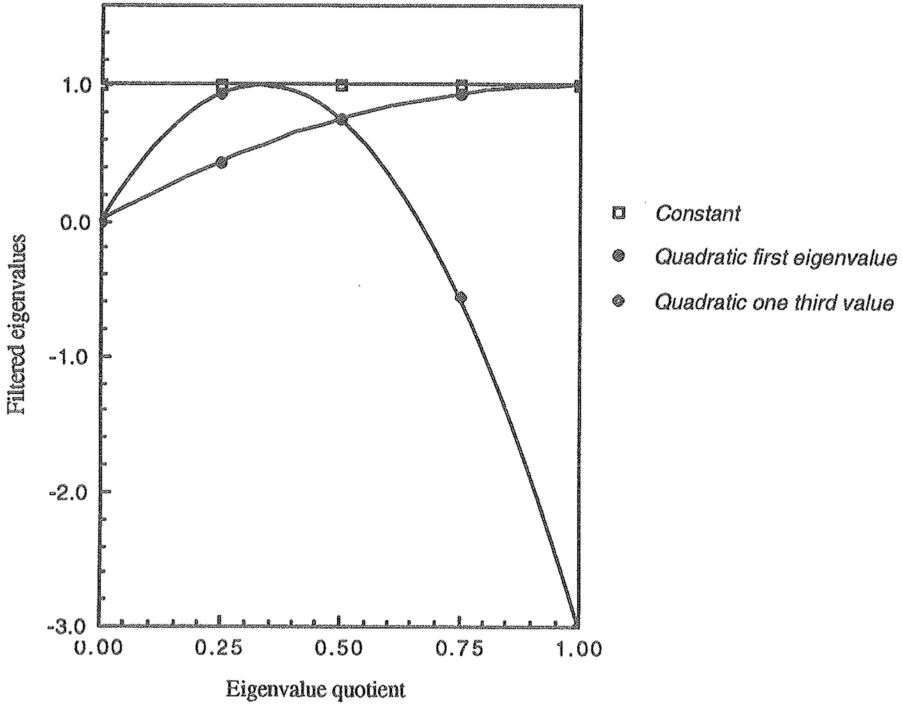


Figure 3.3 *Free quadratic filter of SCA.*

Generally the filtered eigenvalue in figure 3.3 is maximal, when the eigenvalue quotient is equal to $\hat{w}_{(k)s}/\phi_{1k}^2$.

3.4.2 *Relation with variance accounted for and MPCA*

As a general rule SCA gradually prevents the occurrence of small variance accounted for. This is achieved by *differential weighting* of $\mathbf{P}_k'\mathbf{x}_s$, the projection weights of $\mathbf{x}_s$ projected on to the subspaces $\mathbf{P}_k$. For each set k and dimension s we express the variance accounted for $VAF(\mathbf{X},k)$ as a product of the correlations of set variates and the suboptimal balancing factors $\hat{w}_{(k)s}$ (3.9). By substituting (3.9) in (3.5), we obtain

$$VAF(\mathbf{X},k) = \sum_{s=1}^{p} \rho^2_{(k)s}\hat{w}_{(k)s}, \qquad \forall k,s \quad (3.11)$$

where $\rho^2_{(k)s} = (\mathbf{x}_s'\mathbf{z}_{(k)s})^2$ and $VAF(\mathbf{X},k) = \text{tr } \mathbf{X}'\mathbf{S}_k\mathbf{X}$.

If we want to prevent small variances accounted for, we must not only maximize the correlations of set variates $\rho^2_{(k)s}$, but we must also prevent small values for the suboptimal balancing factors $\hat{w}_{(k)s}$. The latter goal is pursued by differential weighting of $\mathbf{P}_k'\mathbf{x}_s$, which are the projection weights of the latent variable $\mathbf{x}_s$ projected on to the subspaces $\mathbf{P}_k$ spanned by the sets. The differential weights are equal to the filtered eigenvalues given in (3.7). The available projection space of the common latent variable $\mathbf{x}_s$ for high $\hat{w}_{(k)s}$ values is gradually reduced if $\hat{w}_{(k)s}$ gets smaller, because the penalty for projecting on singular vectors with large singular values is increasing fast. As we see in figure 3.3 for $\hat{w}_{(k)s} = 0.33\times\phi^2_{1k}$ the differential weight for projection on the first singular vector is already −3. For smaller values of $\hat{w}_{(k)s}$ this differential weight decreases fast.

We cannot only produce (3.11) by combining (3.9) and (3.5), but also the equality $\rho^2_{(k)s} = \mathbf{x}_s'\mathbf{S}_k\mathbf{x}_s\hat{w}^{-1}_{(k)s}$ $\forall k,s$. It shows how SCA maximizes weighted variance accounted for. SCA can be formulated as a MPCA method by taking the balancing constants of MPCA in (2.7) for set k and dimension s equal to the suboptimal balancing factors in (3.9). The balancing of SCA is closely related to the balancing of MPCAs in (2.19) with balancing constants equal to $w_k = \bar{I}_k^{-1}$. If the efficient rank of a matrix is equal to the number of variables m_k, the suboptimal balancing factors are equal to 1 just as $\bar{I}_k^{-1}$. If the efficient rank of a matrix goes to its minimum value of 1, the suboptimal balancing factors go to their maximum value m_k, which is equal to the maximum of $\bar{I}_k^{-1}$. The difference between SCA and MPCAs is that SCA assesses the amount of superfluous information for each set and dimension separately, whereas MPCAs is doing this for each set independent of dimensions.

3.4.3 Relation with MCCA

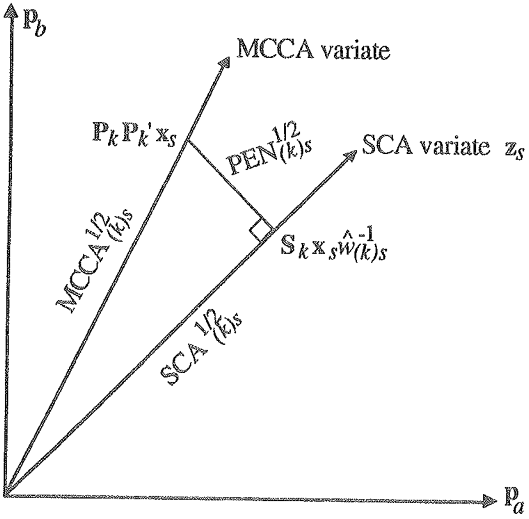
The MCCA fit function gives the upper bounds for the SCA fit function, as we can show by rewriting the second line of (3.7) for each set and dimension as

$$\mathrm{SCA}_{(k)s} = \mathrm{MCCA}_{(k)s} - \mathrm{PEN}_{(k)s}, \qquad \forall k,s \qquad (3.12)$$

where $\mathrm{MCCA}_{(k)s} = \mathbf{x}_s'\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s$ denotes the MCCA fit as defined in (2.22)
and $\mathrm{PEN}_{(k)s} = \mathbf{x}_s'\mathbf{P}_k(\mathbf{I} - \Phi_k^2 w_{(k)s}^{-1})^2\mathbf{P}_k'\mathbf{x}_s$ defines a penalty function.

In fact we reformulated in (3.12) the SCA method as a hybrid method. Because $\mathrm{PEN}_{(k)s} \geq 0$, we always have $\mathbf{0}\mathbf{0}' \leq \mathbf{I} - (\mathbf{I} - \Phi_k^2 w_{(k)s}^{-1})^2 \leq \mathbf{I}$. The free quadratic filter in (3.8) is a special (v=1) subfilter of the weighted hybrid filter $\mathbf{I} - v(\mathbf{I} - \Phi_k^2 w_{(k)s}^{-1})^2$. This weighted filter contains a weighted set variance filter $v(\mathbf{I} - \Phi_k^2 w_{(k)s}^{-1})^2$ subtracted from a set correlation filter $\mathbf{I}$. By inserting the weighted hybrid SCA filter in (2.1) we define the weighted hybrid SCA fit function. By substituting extreme values for v we know the corresponding extreme fit functions. If v is very large the weighted hybrid SCA filter transforms all diagonal elements of Φ_k^2 with values $\hat{w}_{(k)s}$ to the value 1 (see section 3.4.1). All diagonal elements of Φ_k^2 smaller than $\hat{w}_{(k)s}$ will be highly negative and the optimal solution of $\mathbf{X}$ for weighted hybrid SCA will avoid eigenvectors with eigenvalues non equal to $\hat{w}_{(k)s}$. If v=0, weighted hybrid SCA is equal to MCCA.

Geometrically the penalty is related with the size of the improvement step of the MCCA variate needed to obtain a larger variance accounted for. This relation is valid for each dimension s and set k separately, and is illustrated in figure 3.4 for suboptimal balancing factors $\hat{w}_{(k)s}$. Figure 3.4 is based upon figure 3.1. As in figure 3.1 we assume without loss of generality that the MCCA canonical variate is located in the plane of the eigenvectors $\mathbf{p}_a$ and $\mathbf{p}_b$ with eigenvalues $\phi_a^2 > \phi_b^2$. The projection of the common latent variate $\mathbf{x}_s$ on the space of $\mathbf{H}_k$ and therefore on the MCCA variate is given by $\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s$ The projection of $\mathbf{x}_s$ on the SCA variate $\mathbf{z}_{(k)s}$ is given by $\mathbf{S}_k\mathbf{x}_s\hat{w}_{(k)s}$, where $\hat{w}_{(k)s}$ is defined in (3.9). The lengths of the projected vectors are respectively $\mathrm{MCCA}^{1/2}_{(k)s}$ and $\mathrm{SCA}^{1/2}_{(k)s}$, and correspond to the correlations of $\mathbf{x}_s$ with the MCCA and the SCA variate.



$\mathbf{p}_b$
MCCA variate
$\mathbf{P}_k\mathbf{P}_k'\mathbf{x}_s$
$\mathrm{PEN}^{1/2}_{(k)s}$
SCA variate $\mathbf{z}_s$
$\mathrm{MCCA}^{1/2}_{(k)s}$
$\mathbf{S}_k\mathbf{x}_s\hat{w}^{-1}_{(k)s}$
$\mathrm{SCA}^{1/2}_{(k)s}$
$\mathbf{p}_a$

Figure 3.4 *Geometric illustration of penalty function.*

Recapitulating, the adjusted method SCA is a special (v=1) subfilter of the weighted hybrid SCA method. For v=0, weighted hybrid SCA is equal to MCCA. On the other hand we must bear in mind that the weighted hybrid SCA method is generally *not* an adjusted method, because the original goal of maximizing the sum of squared set correlations is only preserved in one special case.

3.4.4 Relation with INDSCAL

First we describe the INDSCAL model and fit function and elaborate some of the INDSCAL properties. Secondly the relation between INDSCAL and SCA is established by proposing and fitting a new model, the *INDRES model.*

The weighted Euclidian three-way scaling model referred to as the INDSCAL model was proposed independently by Bloxom (1968), Horan (1969) and Carroll & Chang (1970). The INDSCAL model is formulated by weighting squared estimated distances between objects. The INDSCAL model in scalar product form (Arabie, Carroll & DeSarbo, 1987) is described by

$$\mathbb{S}_k = \mathbb{X}\mathbb{W}_k\mathbb{X}' + \mathbb{E}_k, \qquad \forall k \quad (3.13)$$

where $\mathbb{S}_k$ denotes a $m \times m$ scalar product matrix,
$\mathbb{X}\mathbb{W}_k\mathbb{X}'=\mathbb{M}_k$ denotes fitted model parameters,
$\mathbb{X}$ denotes unit orthonormalized dimensions ($m \times p$),
$\mathbb{W}_k$ denotes a $p \times p$ diagonal matrix with dimension weights $w_{(k)s}$,
and $\mathbb{E}_k$ denotes a $m \times m$ matrix with residuals.

The interpretation of the matrices used in (3.13) can easily be embedded in the notation of the preceding sections:

$\mathbb{S}_k$: In the beginning of this chapter we defined $\mathbb{S}_k=\mathbb{H}_k\mathbb{H}_k'$, but in fact the matrix $\mathbb{S}_k$ does not necessarily have to be equal to $\mathbb{H}_k\mathbb{H}_k'$. Without loss of generality it can be any positive semi-definite matrix of suitable converted dissimilarity or similarity measures.

$\mathbb{X}$: The dimensions in the INDSCAL model do not have to be orthogonal. Despite some loss of generality we use in this section the orthogonal version of INDSCAL to show the relation with SCA. Kroonenberg (1983, p.118) denotes this method as 'orthonormal INDSCAL', Kiers (1989, p.14) refers to it by the acronym INDORT and gives an elaborate discussion on the subject. In most practical applications the optimal INDSCAL dimensions will be near to orthogonality. See Arabie, Carroll & DeSarbo, 1987, page 36: "The axes provided by INDSCAL generally turn out to be orthogonal or nearly so". Therefore our shift from INDSCAL to INDORT will have no major implications. For the one dimensional solution there are certainly no implications, because in that case the solutions are exactly equal.

$\mathbb{W}_k$: The optimal dimension weights $w_{(k)s}$ in the INDSCAL model are a measure of relative importance, just as the balancing factors of SCA in section 3.4.2.

The INDSCAL model is fitted in least squares sense by minimizing the loss function:

$$\text{INDSCAL:} \qquad \text{Loss}(\mathbb{X},\mathbb{W}_k) = \text{tr} \sum_{k=1}^{K} \mathbb{E}_k\mathbb{E}_k, \qquad (3.14)$$

and fitting the orthonormal INDSCAL model in the least squares sense comes down to minimizing the loss function

$$\text{INDORT:}\qquad \text{Loss}(\mathbf{X},\mathbf{W}_k) = \text{tr}\sum_{k=1}^{K}\mathbf{E}_k\mathbf{E}_k = \text{tr}\sum_{k=1}^{K}(\mathbf{X}\mathbf{W}_k\mathbf{X}'-\mathbf{S}_k)(\mathbf{X}\mathbf{W}_k\mathbf{X}'-\mathbf{S}_k)$$

$$= \text{tr}\sum_{s=1}^{p}\sum_{k=1}^{K}(\mathbf{x}_s w_{(k)s}\mathbf{x}_s'-\mathbf{S}_k)(\mathbf{x}_s w_{(k)s}\mathbf{x}_s'-\mathbf{S}_k) + c, \qquad (3.15)$$

where c denotes a constant with $c = \sum_{k=1}^{K}(1\text{-}p)\text{tr}\,\mathbf{S}_k^2$,

with $\mathbf{X}'\mathbf{X}=\mathbf{I}$. The transition to the last part of (3.15) with constant c added is due to the orthogonality of $\mathbf{X}$. Analogous to the procedure in section 3.4.1 we find ***suboptimal*** dimension weights, which are a function of $\mathbf{X}$. We denote these suboptimal dimension weights with $\tilde{w}_{(k)s}$. The suboptimal dimension weights are equal to

$$\tilde{w}_{(k)s} = \mathbf{x}_s'\mathbf{S}_k\mathbf{x}_s. \qquad \forall k,s \qquad (3.16)$$

Substitution of the suboptimal dimension weights (3.16) in (3.15) reveals after minor elaboration a very simple fit function. (See Kiers 1989, p.43). It turns out that minimization of (3.15) produces the same optimal $\mathbf{X}$ as maximization of

$$\text{INDORT:}\qquad \text{Fit}(\mathbf{X}) = \sum_{s=1}^{p}\sum_{k=1}^{K}\tilde{w}_{(k)s}^2 = \sum_{s=1}^{p}\sum_{k=1}^{K}\mathbf{x}_s'\mathbf{S}_k\mathbf{x}_s\tilde{w}_{(k)s}$$

$$= \sum_{s=1}^{p}\sum_{k=1}^{K}(\mathbf{x}_s'\mathbf{S}_k\mathbf{x}_s)^2. \qquad (3.17)$$

Domination by sets with low information span

We inserted the formulation $\mathbf{x}_s'\mathbf{S}_k\mathbf{x}_s\tilde{w}_{(k)s}$ in (3.17) to clarify a relation between orthonormal INDSCAL and SCA. In section 3.4.2 the equality $\rho_{(k)s}^2 = \mathbf{x}_s'\mathbf{S}_k\mathbf{x}_s\hat{w}_{(k)s}^{-1}$ $\forall k,s$, showed how SCA maximizes weighted variance accounted for. In the same way INDORT can be formulated as a MPCA method by taking the balancing constants $w_{(k)s}$ of MPCA in (2.7) for set k and dimension s equal to the reciprocal suboptimal dimension weights in (3.16), $w_{(k)s}=\tilde{w}_{(k)s}^{-1}$. The balancing of the INDORT sets indicates that (orthonormal) INDSCAL solutions will be dominated by sets with a *low information span* and a *low efficient rank*, if the sets are normalized to the same total sum of squares.

The information span I_k and the efficient rank are defined in section 2.2.5 beginning with formula (2.17) and further.

A matrix $\mathbf{H}_k$ with a low information span has much variance concentrated on only a limited subspace. This subspace therefore attracts any solution space $\mathbf{X}$, that seeks to maximize the VAF and therefore even more the orthonormal INDSCAL $\mathbf{X}$, that seeks to maximize the squared VAF for each dimension. The SCA solution is not dominated by sets with much redundant information and will be better balanced in this respect.

Simple structure with equal information span

There is another important aspect in regard to the weighting of VAF. If the information span of all sets, normalized to the same total sum of squares, is equal to I_k=1, we have $\mathbf{S}_k$=$\mathbf{P}_k\mathbf{P}_k'$ $\forall k$. Even in this case the orthonormal INDSCAL solution will still emphasize some sets as much as possible in order to obtain a simple structure. This tendency to exaggerate the differences between the sets is analogous to the simple structure rotation of *variables* instead of *sets*. The quartimax fit function is in this respect a special case of the INDORT$_{\text{fit}}$ function in (3.17). The functions are equal if each set $\mathbf{h}_k$ consists only of one variable. This property explains why INDSCAL tends to find a unique orientation of dimensions, even if K=1. The SCA solution for I_k=1 $\forall k$, is exactly equal balanced in the sense that in this case the suboptimal balancing factors in (3.9) are equal to $\hat{w}_{(k)s}$=1 $\forall k$.

Residuals not orthogonal to common latent variables

We consider the INDSCAL model $\mathbf{S}_k = \mathbf{M}_k + \mathbf{E}_k = \mathbf{X}\mathbf{W}_k\mathbf{X}' + \mathbf{E}_k$, $\forall k$, as formulated in (3.13). In analogy with the PCA model some users of the INDSCAL program might erroneously think that at convergence we have strong orthogonality between residuals $\mathbf{E}_k$ and the dimensions given by $\mathbf{X}$,

$$\mathbf{X}'\mathbf{E}_k = \mathbf{0}\mathbf{0}'. \qquad \forall k \quad (3.18)$$

where $\mathbf{0}$ denotes a column vector of appropriate size with elements 0.

However this is usually not true and generally not possible, neither for the INDSCAL model, nor for the orthonormal INDSCAL model, where $\mathbf{X}$ is required to satisfy $\mathbf{X}'\mathbf{X}=\mathbf{I}$. We will refer to this statement as the *residual rule* for the INDSCAL models. The residual rule will be proved later in this section. At convergence of the INDSCAL program we do not have (3.18), but only the weak orthogonality

$$\sum_{k=1}^{K} \text{tr } \mathbf{M}_k\mathbf{E}_k = 0. \tag{3.19}$$

The weak orthogonality of residuals can be an undesirable property, because it implies that important information of $\mathbf{S}_k$ related to $\mathbf{X}$ can be left undetected in the residuals. The recovery of true INDSCAL dimensions will be less effective, if the estimates $\mathbf{M}_k$ give a distorted image of the original matrix $\mathbf{S}_k$. These distortions can be understood by examining the orthonormal INDSCAL model. If we elaborate the orthogonality restriction $\mathbf{E}_k\mathbf{X}=\mathbf{0}\mathbf{0}'$, we obtain

$$\mathbf{E}_k\mathbf{X} = (\mathbf{S}_k - \mathbf{M}_k)\mathbf{X} = \mathbf{S}_k\mathbf{X} - \mathbf{X}\mathbf{W}_k = \mathbf{0}\mathbf{0}'. \qquad \forall k \quad (3.20)$$

The last equality in (3.20) implies that the restriction $\mathbf{E}_k\mathbf{X}=\mathbf{0}\mathbf{0}'$ is only valid if the columns of $\mathbf{S}_k\mathbf{X}$ are proportional to the respective columns of $\mathbf{X}$. In figure 3.5 we show geometrically for three columns of $\mathbf{S}_k$ how the ideal projections on the space $\mathbf{X}$ would be distorted by the multiset restrictions of the INDSCAL model. The column vectors of $\mathbf{E}_k$ are clearly not perpendicular to INDSCAL dimensions $\mathbf{X}$.

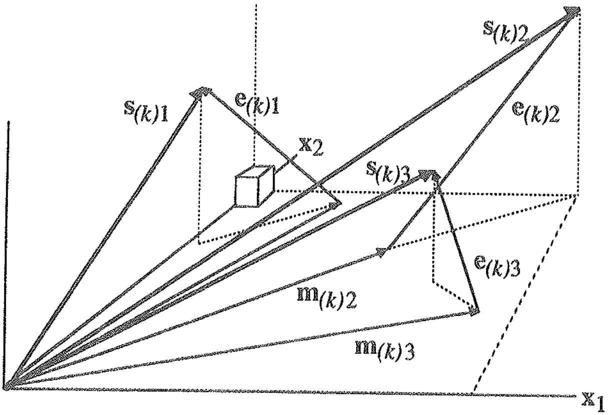


Figure 3.5 *Distorted projections of* $\mathbf{S}_k$ *on space* $\mathbf{X}$.

If we are not satisfied with the weak orthogonality of residuals in the INDSCAL model, we cannot improve orthogonality by applying other computational methods (like De Leeuw & Pruzansky, 1978) to fit the INDSCAL model. We have to adapt the model for instance by penalizing non-orthogonality between the INDSCAL dimensions and the residuals for each set k. We call the resulting model the INDRES model. The INDRES model in scalar product form is given by:

$$\left\{\begin{array}{ccccc} \mathbf{S}_k & = & \mathbf{X}\mathbf{W}_k\mathbf{X}' & + & \mathbf{E}_k \\ & & & & \\ \mathbf{S}_k\mathbf{P}_\mathbf{X} & = & \mathbf{X}\mathbf{W}_k\mathbf{X}'\mathbf{P}_\mathbf{X} & + & \mathbf{E}_k\mathbf{P}_\mathbf{X} \end{array}\right\}, \qquad \forall k \qquad (3.21)$$

where $\mathbf{P}_\mathbf{X}$ denotes an orthonormal basis of $\mathbf{X}$.
and all other parameters have the same notation as for the INDSCAL model in (3.13).

The first line of the INDRES model specifies the INDSCAL model and the second line penalizes non-orthogonality between the residuals $\mathbf{E}_k$ and the INDSCAL dimensions $\mathbf{X}$. The dimensions in the INDRES model do not have to be orthogonal. The orthonormal basis $\mathbf{P}_\mathbf{X}$ is introduced in order to allow for this possibility. Many functions can be proposed to fit the INDRES model. One possibility is to minimize a weighted INDSCAL loss function

$$\text{INDWEI: Loss}(\mathbf{X},\mathbf{W}_k) = \text{tr }(\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}'-\mathbf{I})\mathbf{E}_k\mathbf{E}_k(\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}'-\mathbf{I}) + v \text{ tr } \mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}'\mathbf{E}_k\mathbf{E}_k\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}'$$

$$= \text{tr} \sum_{k=1}^{K} (\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}'\mathbf{S}_k-\mathbf{S}_k)'(\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}'\mathbf{S}_k-\mathbf{S}_k) +$$

$$v \text{ tr} \sum_{k=1}^{K} (\mathbf{X}\mathbf{W}_k\mathbf{X}'-\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}'\mathbf{S}_k)'(\mathbf{X}\mathbf{W}_k\mathbf{X}'-\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}'\mathbf{S}_k), \qquad (3.22)$$

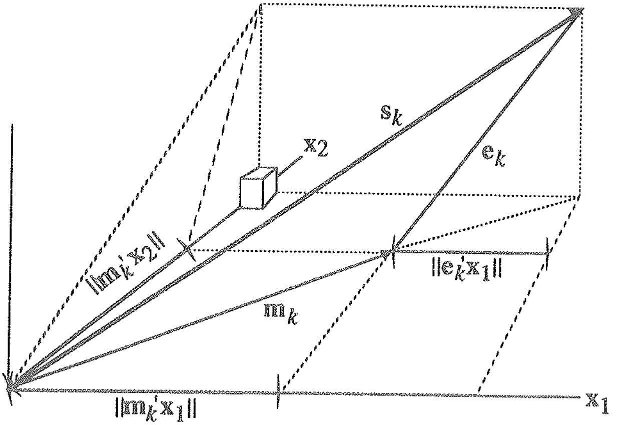
where v denotes a balancing constant.

For v=1, (3.22) gives a decomposition of the error $\mathbf{E}_k$ in the INDSCAL loss function (3.14). It is interesting to notice that the INDSCAL weights $\mathbf{W}_k$ can only minimize the error $\mathbf{E}_k\mathbf{P}_\mathbf{X}$. For v=1, (3.22) also minimizes the error $\mathbf{E}_k$ of the INDRES model (3.21). The second part of (3.22) is equal to $v\text{tr}\Sigma_k\mathbf{P}_\mathbf{X}'\mathbf{E}_k\mathbf{E}_k\mathbf{P}_\mathbf{X}$ and minimizes the error $\mathbf{E}_k\mathbf{P}_\mathbf{X}$ of the INDRES model. Therefore the parameters of the INDRES model can be estimated by minimizing the INDSCAL loss function. More emphasis on minimizing the error $\mathbf{E}_k\mathbf{P}_\mathbf{X}$ can be given by minimizing (3.22) with v>1.

We will now confine ourselves within the scope of this section to define a fit function for the INDRES model that gives a relation with SCA. Therefore we first impose the restriction $\mathbf{P}_\mathbf{X}=\mathbf{X}$, which implies $\mathbf{X}'\mathbf{X}=\mathbf{I}$. With this restriction least squares fitting of the INDRES model could involve the minimization of the product of two subfunctions $\mathrm{tr}\Sigma_k\mathbf{E}_k\mathbf{E}_k$ and $\mathrm{tr}\Sigma_k\mathbf{X}'\mathbf{E}_k\mathbf{E}_k\mathbf{X}$. Because minimization of the second subfunction $\mathrm{tr}\Sigma_k\mathbf{X}'\mathbf{E}_k\mathbf{E}_k\mathbf{X}$ induces the orthogonalization of $\mathbf{X}$ and $\mathbf{E}_k$, we could instead of $\mathrm{tr}\Sigma_k\mathbf{E}_k\mathbf{E}_k$ just as well minimize $(\mathrm{tr}\Sigma_k\mathbf{M}_k\mathbf{M}_k)^{-1} = (\mathrm{tr}\Sigma_k\mathbf{X}'\mathbf{M}_k\mathbf{M}_k\mathbf{X})^{-1}$. The last equation is valid, because $\mathbf{X}'\mathbf{X}=\mathbf{I}$. Due to the same orthogonality restriction we can split $\mathbf{X}'\mathbf{E}_k\mathbf{E}_k\mathbf{X}$ and $\mathbf{X}'\mathbf{M}_k\mathbf{M}_k\mathbf{X}$ respectively in $\Sigma_s\mathbf{x}_s'\mathbf{E}_k\mathbf{E}_k\mathbf{x}_s$ and $\Sigma_s\mathbf{x}_s'\mathbf{M}_k\mathbf{M}_k\mathbf{x}_s$ for each set k and minimize the product $(\mathbf{x}_s'\mathbf{E}_k\mathbf{E}_k\mathbf{x}_s)(\mathbf{x}_s'\mathbf{M}_k\mathbf{M}_k\mathbf{x}_s)^{-1}$ for each dimension separately. In this way orthogonality of $\mathbf{X}$ and $\mathbf{E}_k$ is approximated equally for all dimensions. We propose to fit the orthonormal INDRES model in least squares sense by minimizing the sum of the ratio's:

$$\text{INDRES: Loss}(\mathbf{X},\mathbf{W}_k) = \sum_{s=1}^{p}\sum_{k=1}^{K}\frac{\mathbf{x}_s'\mathbf{E}_k\mathbf{E}_k\mathbf{x}_s}{\mathbf{x}_s'\mathbf{M}_k\mathbf{M}_k\mathbf{x}_s}, \qquad (3.23)$$

subject to $\mathbf{X}'\mathbf{X} = \mathbf{I}$.



$\mathbf{s}_k$
$\mathbf{x}_2$
$\mathbf{e}_k$
$\|\mathbf{m}_k'\mathbf{x}_2\|$
$\|\mathbf{e}_k'\mathbf{x}_1\|$
$\mathbf{m}_k$
$\|\mathbf{m}_k'\mathbf{x}_1\|$
$\mathbf{x}_1$

Figure 3.6 *Approximation of orthogonality between $\mathbf{E}_k$ and $\mathbf{X}$.*

In figure 3.6 we have redrawn column vector $\mathbf{s}_{(k)2}$, from figure 3.5 as a representative column vector of $\mathbf{S}_k$. With the representative column vectors denoted by $\mathbf{s}_k$, $\mathbf{m}_k$ and $\mathbf{e}_k$ we want to show how the columns of $\mathbf{E}_k$ are made as orthogonal as possible to the columns of $\mathbf{X}$ and $\mathbf{M}_k$ by minimizing for each dimension s and set k the *sum of all* squared projection lengths $(\mathbf{e}_k'\mathbf{x}_s)^2$ of the column vectors of $\mathbf{E}_k$ divided by the *sum of all* squared projection lengths $(\mathbf{m}_k'\mathbf{x}_s)^2$ of the column vectors of $\mathbf{M}_k$.

The INDRES loss in (3.23) can be elaborated as follows

$$\text{INDRES: Loss}(\mathbf{X},\mathbf{W}_k) = \sum_{s=1}^{p}\sum_{k=1}^{K} \mathbf{x}_s'(\mathbf{x}_s w_{(k)s}\mathbf{x}_s' - \mathbf{S}_k)(\mathbf{x}_s w_{(k)s}\mathbf{x}_s' - \mathbf{S}_k)'\mathbf{x}_s w_{(k)s}^{-2}$$

$$= \text{tr} \sum_{k=1}^{K} (\mathbf{X}-\mathbf{S}_k\mathbf{X}\mathbf{W}_k^{-1})'(\mathbf{X}-\mathbf{S}_k\mathbf{X}\mathbf{W}_k^{-1}), \qquad (3.24)$$

with $\mathbf{X}'\mathbf{X}=\mathbf{I}$. We take the first line of formula (3.7) for defining a corresponding SCA loss function

$$\text{SCA:} \quad \text{Loss}(\mathbf{X},\mathbf{W}_k) = pK - \text{SCA}_{\text{fit}}(\mathbf{x}_s, w_{(k)s})$$

$$= \text{tr} \sum_{k=1}^{K} (\mathbf{X}-\mathbf{S}_k\mathbf{X}\mathbf{W}_k^{-1})'(\mathbf{X}-\mathbf{S}_k\mathbf{X}\mathbf{W}_k^{-1}), \qquad (3.25)$$

with $\mathbf{X}'\mathbf{X}=\mathbf{I}$. The relation with the loss function for INDRES in (3.24) is obvious. If the rows and columns of $\mathbf{S}_k$ have zero mean minimization of the INDRES loss function comes down to maximizing the sum of squared set correlations $\rho^2_{(\mathbf{x}_s;\mathbf{S}_k\mathbf{x}_s)}$ between $\mathbf{x}_s$ and $\mathbf{S}_k\mathbf{x}_s$ over all s and k, as we can verify in (3.5).

We promised to prove the residual rule for the INDSCAL models. This rule states that it is in general not possible to find INDSCAL dimensions $\mathbf{X}$ that are orthogonal to the residuals $\mathbf{E}_k$ $\forall k$, neither for the INDSCAL model, nor for the orthonormal INDSCAL model, where $\mathbf{X}'\mathbf{X}=\mathbf{I}$. If we prove this rule for the orthonormal INDSCAL model, it is also valid for the general INDSCAL model, because the two models have the same solution in the one dimensional case.

Proof: For the orthonormal INDSCAL model we know that the residual rule is only violated if the INDRES loss in (3.23) and therefore the SCA loss in (3.25) is equal to zero. This implies that the SCA fit must be equal to pK and that all squared set correlations in (3.5) must be equal to 1. It also implies that all squared canonical correlations of the MCCA method described in (2.20) must be equal to 1 for p

dimensions, because the MCCA fit function gives the upper bound for the SCA fit function, as we have shown in (3.12). So perfect fit for the MCCA method is a necessary, but usually not sufficient condition to violate the residual rule for the INDSCAL models. It is clear that perfect canonical fit in p dimensions is a very special case and will *in general* not occur, which proves the validity of our residual rule. □

3.4.5 Summary of INDSCAL and SCA properties

Summarising the comparison between INDSCAL and SCA we found that the INDSCAL solution is dominated by sets with low information span, that it has a tendency to exaggerate the differences between the sets, that it is dependent on the normalizations of $\mathbf{S}_k$ and that it can leave distortions of the original data undetected. The SCA solution is more balanced in the weighting of sets, invariant under different normalizations of $\mathbf{S}_k$ and gives a more complete relation with the original data by making the residuals as much as possible orthogonal to the $\mathbf{x}_s$. It can be expected that this property improves the recovery of true INDSCAL dimensions. In the next section 3.5 we compare the INDSCAL and the SCA solutions in a simulation study. In chapter 7 the theoretical properties of SCA and INDSCAL are confirmed in an analysis of Miller-Nicely data.

3.5 Simulation study of INDSCAL compared with SCA

The main purpose of this section is to investigate if the SCA solution improves the recovery of true INDSCAL dimensions. This improvement could be attained by making the residuals as much as possible orthogonal to the recovered dimensions.

In order to compare the properties of INDSCAL and SCA we set up a little simulation study with 6 individuals or sets and 20 stimuli. As we saw in the previous section 3.4.4, fitting of the orthonormal INDRES model leads to the SCA fit function. In this study we suppose that each scalar product matrix $\mathbf{S}_k$ is decomposed in a common part of the orthonormal INDRES model

$$\mathbf{X}\mathbf{W}_k\mathbf{X}' = \mathbf{M}_k, \tag{3.26}$$

where $\mathbf{X}$ denotes unit orthonormalized dimensions (20×2),

$\mathbf{W}_k$ denotes a 2×2 diagonal matrix with common dimension weights $w_{(k)s}$,

and a unique part of the orthonormal INDRES model

$$\mathbf{Y}_k\mathbf{U}_k\mathbf{Y}_k', \tag{3.27}$$

where $\mathbf{Y}_k$ denotes unit orthonormalized dimensions (20×2),

$\mathbf{U}_k$ denotes a 2×2 diagonal matrix with unique dimension weights $u_{(k)s}$.

In the terminology of chapter 4 the true stimulus configuration $\mathbf{S}_k = \mathbf{X}\mathbf{W}_k\mathbf{X}' + \mathbf{Y}_k\mathbf{U}_k\mathbf{Y}_k'$ is an external decomposition, because $\mathbf{X}$ and $\mathbf{Y}_k$ can usually not be written as a linear combination of $\mathbf{S}_k$. The true common stimulus configuration $\mathbf{X}\mathbf{W}_k\mathbf{X}'$ (3.26) has 2 dimensions $\mathbf{X}$, for all 6 sets the same, and positive weights $\mathbf{W}_k$ on a circle with its centre in point (0,0). The true unique stimulus configuration $\mathbf{Y}_k\mathbf{U}_k\mathbf{Y}_k'$ (3.27) has 2 dimensions and is orthogonal to all other true dimensions, common or unique. The weights $\mathbf{U}_k$ are chosen identical to the corresponding $\mathbf{W}_k$. The common-to-total ratio of each true configuration *k* is defined by

$$CT_k = \frac{\text{tr } \mathbf{W}_k}{\text{tr } (\mathbf{W}_k + \mathbf{U}_k)}. \tag{3.28}$$

With $CT=a$, we will refer to $CT_k=a$, $\forall k$.

To each true stimulus configuration $\mathbf{S}_k$ we add constructed error $\mathbf{E}_k$

$$\mathbf{S}_k + \mathbf{E}_k = \mathbf{S}_k + \mathbf{E}_k(\mathbf{W}_k + \mathbf{U}_k)\mathbf{E}_k'. \tag{3.29}$$

We want to approximate the constructed configurations $\mathbf{S}_k + \mathbf{E}_k$ with recovered dimensions $\hat{\mathbf{X}}\hat{\mathbf{W}}_k\hat{\mathbf{X}}'$ and residuals $\hat{\mathbf{E}}_k$,

$$\mathbf{S}_k + \mathbf{E}_k = \hat{\mathbf{X}}\hat{\mathbf{W}}_k\hat{\mathbf{X}} + \hat{\mathbf{E}}_k. \tag{3.30}$$

The recovered dimensions are denoted by $\hat{\mathbf{X}}$ and can be non orthogonal for the INDSCAL model. It appeared to be most efficient in this simulation study to use a 2 dimensional INDORT solution (3.15) as starting configuration for computing the INDSCAL solution. The recovered weights $\hat{\mathbf{W}}_k$ for INDSCAL *and* SCA dimensions are computed according to the INDSCAL procedure of Carroll & Chang (1970) for fixed $\mathbf{X}$, because the optimal INDSCAL weights $\hat{\mathbf{W}}_k$ minimize $\mathrm{tr}\Sigma_k \mathbf{P}_{\mathbf{X}}'\mathbf{E}_k\mathbf{E}_k\mathbf{P}_{\mathbf{X}}$ of the INDRES model. This can be verified in (3.22) for v=1. For the SCA dimensions this implies that they can be derived from $\hat{\mathbf{X}}$ using formula (3.16) of suboptimal dimension weights. We varied the common-to-total ratio (3.28), CT = 1 0.7 0.4 0.2. The error level of $\mathbf{E}_k$ (3.29) was equal to the standard deviation of a unit normalized random normal variable. We chose error level = 0 0.1 0.4 0.7. For each combination of CT and error level we computed four measures, V, δ, M and ϕ^2, for 150 constructed configurations:

$$V = \frac{\mathrm{tr}\ \mathbf{X}'\hat{\mathbf{X}}\hat{\mathbf{X}}'\mathbf{X}}{\mathrm{tr}\ \hat{\mathbf{X}}'\hat{\mathbf{X}}} \tag{3.31}$$

denotes the proportion of variance of the recovered dimensions $\hat{\mathbf{X}}$ accounted for by the true dimensions $\mathbf{X}$. It measures the recovery of true stimulus dimensions with rotational freedom. The following measure is the only distance measure. For perfect recovery δ is zero.

$$\delta = \left(\frac{\mathrm{tr}\ (\mathbf{X}-\hat{\mathbf{X}})'(\mathbf{X}-\hat{\mathbf{X}})}{p}\right)^{1/2} \tag{3.32}$$

denotes the square root of the mean squared difference between all true unit orthonormalized and recovered unit normalized stimulus scores for p dimensions (MacCallum, 1977). It measures the recovery of true stimulus dimensions with unique directions. For the optimal arrangement of true and recovered dimensions with respect to permutations and/or reflections of the dimensions, see MacCallum, 1977. In the next two measures we use a centring operator $\mathbf{J}$ and we concatenate all possible true interpoint stimulus distances between row i and j of $\mathbf{X}\mathbf{W}_k^{1/2}$, $\forall i>j$, and for all k successively in vector $\mathbf{d}$, and all corresponding recovered stimulus distances based on $\hat{\mathbf{X}}\hat{\mathbf{W}}_k^{1/2}$ in vector $\hat{\mathbf{d}}$.

$$M = \frac{(\mathbf{d}'\mathbf{J}\hat{\mathbf{d}})^2}{\mathbf{d}'\mathbf{J}\mathbf{d}.\hat{\mathbf{d}}'\mathbf{J}\hat{\mathbf{d}}} \tag{3.33}$$

denotes the squared correlation between true and recovered distances across all stimulus pairs and sets, normalized matrix conditional. It measures the recovery of true interpoint distances and is also called the index of metric determinacy (Young, 1970).

$$\phi^2 = \frac{(\mathbf{d}'\hat{\mathbf{d}})^2}{\mathbf{d}'\mathbf{d}.\hat{\mathbf{d}}'\hat{\mathbf{d}}} \tag{3.34}$$

denotes the squared coefficient of congruence between true and recovered distances across all stimulus pairs and sets, normalized matrix conditional. It measures the recovery of true interpoint distances with the coefficient of congruence (Tucker, 1951).

3.5.1 Results of simulation study

In the following tables we present the mean values of the four above mentioned recovery measures over 150 constructed configurations for each combination of *CT* (3.28) and error level. The INDSCAL and SCA solutions are computed using the same 150 constructed configurations.

Table 3.1 *V: recovery of true stimulus dimensions with rotational freedom.*

Error		INDSCAL				SCA			
level	*CT*:	1	0.7	0.4	0.2	1	0.7	0.4	0.2
0		1	1	1	0	1	1	1	1
0.1		0.98	0.97	0.85	0.01	0.98	0.97	0.97	0.96
0.4		0.90	0.90	0.44	0.05	0.93	0.92	0.91	0.87
0.7		0.84	0.82	0.33	0.07	0.89	0.87	0.84	0.79

We did some extra computation in order to find the *CT* value below which the INDSCAL solution degenerates with zero error. This threshold was found at *CT*=0.36. In table 3.2 the standard deviations of *V* in table 3.1 are presented. For δ and *M* the standard deviations will not be given.

Table 3.2 *Standard deviations of V.*

Error		INDSCAL				SCA			
level	*CT*:	1	0.7	0.4	0.2	1	0.7	0.4	0.2
0		0	0	0	0	0	0	0	0
0.1		0.01	0.01	0.19	0.01	0.01	0.01	0.01	0.01
0.4		0.03	0.03	0.24	0.04	0.02	0.02	0.02	0.03
0.7		0.05	0.06	0.19	0.04	0.03	0.04	0.04	0.08

As we could expect are the standard deviations of *V* for the INDSCAL solutions higher near the degeneration threshold *CT*=0.36.

Table 3.3 *δ: recovery of true stimulus dimensions with unique directions (distance measure).*

Error		INDSCAL				SCA			
level	*CT*:	1	0.7	0.4	0.2	1	0.7	0.4	0.2
0		0	0	0	1.41	0	0	0	0
0.1		0.16	0.17	0.39	1.35	0.16	0.18	0.19	0.21
0.4		0.33	0.34	0.90	1.29	0.28	0.31	0.32	0.38
0.7		0.43	0.46	1.02	1.27	0.35	0.38	0.43	0.52

Table 3.4 *M: recovery of true interpoint distances, index of metric determinacy.*

Error		INDSCAL				SCA			
level	*CT*:	1	0.7	0.4	0.2	1	0.7	0.4	0.2
0		1	1	1	0	1	1	1	1
0.1		0.88	0.88	0.66	0	0.87	0.86	0.86	0.84
0.4		0.62	0.59	0.14	0.01	0.67	0.64	0.64	0.54
0.7		0.40	0.39	0.05	0.01	0.50	0.51	0.47	0.34

Table 3.5 *ϕ^2: recovery of true interpoint distances, squared coefficient of congruence.*

Error		INDSCAL				SCA			
level	*CT*:	1	0.7	0.4	0.2	1	0.7	0.4	0.2
0		1	1	1	0.17	1	1	1	1
0.1		0.97	0.97	0.86	0.17	0.97	0.96	0.96	0.95
0.4		0.88	0.87	0.50	0.21	0.90	0.89	0.89	0.84
0.7		0.78	0.78	0.40	0.22	0.83	0.84	0.82	0.74

3.5.2 Conclusions

The properties of INDSCAL and SCA are evaluated with an exploratory simulation study. The results for SCA are promising. The SCA fit function almost always gives a better reconstruction both of the true stimulus dimensions, including its unique directions, and of the true interpoint distances. It is remarkable that the improvement can also be significant if there is no unique part in the true stimulus configuration. These values can be found in the left-hand columns of tables 3.1, 3.3, 3.4 and 3.5 for *CT*=1. The 'no unique part' improvement is due to the fact that the SCA fit function minimizes mainly the error projected on to the recovered common dimensions. This projection reduces the effect of the error on the solution. Another interesting result is found in the first row of the tables 3.1, 3.3, 3.4 and 3.5 with error level=0. Because it is dominated by the true unique stimulus configurations of the sets, the INDSCAL solution is seen to degenerate. In the SCA solution these true unique stimulus configurations are incorporated in the error, because they have zero projections on to the recovered dimensions. As the number of dimensions for the SCA solution is raised above the dimensionality of the true common stimulus configuration the true unique stimulus configurations emerge in badly fitting dimensions.

Chapter 4

SET VARIANCE WITH SET CORRELATION CONSTRAINTS OR REFLECTED VARIANCE

In chapter 3 the adjusted method of Set Component Analysis was formulated from several points of view. We repeat this approach for *Reflected Variance methods.* (1) The Reflected Variance methods integrate a set variance and a set correlation part by maximizing the variance accounted for by set variates and adjusting the set variates with set correlation constraints. (2) The Reflected Variance methods project variables from one set on to another set, project these variables back and then compute principal components of the *reflected variables.* (3) By defining reflecting filters, Reflected Variance methods are related with the filter theory formulated in chapter 2. The principle of reflected variables is elaborated by defining *Reflected Component Analysis* (RCA) and *Reflected Discriminant Analysis* (RDA). It will be shown theoretically how and under which conditions RDA can improve group prediction compared to Discriminant Analysis (DA) and Principal Component - Discriminant Analysis (PC-DA). In a simulation study theoretical results are confirmed. Some multiset and nonlinear extensions are proposed.

Introduction

In chapter 3 we combined one *fit function* with the constraint of an *adjusting function.* The fit function was the sum of squared canonical correlations and the adjusting function was improving variance accounted for. In this chapter the roles of the functions are interchanged and one function slightly changed: We maximize variance accounted for and improve the squared canonical correlations (not the *sum* of squared canonical correlations). This slight change of correlation function already indicates that we are always dealing in this chapter with two sets of variables. There are only two additive multiset extensions.

From the geometrical point of view we are maximizing *reflected* variance (Nierop, 1991) accounted for. Reflected means that we look at the variables through the mirror of other relevant external information and in this way filter out irrelevant information. Therefore the constraint of the adjusting 'squared canonical correlations' function is called *reflecting constraint.*

The integration of set correlation and set variance follows the same lines as in the previous chapter. The basic method is *Reflected Component Analysis* (RCA). The relation with the filter theory in chapter 2 is given in section 4.2. *Reflected Discriminant Analysis* (RDA) is formulated in section 4.3 as a special case of RCA and it will serve as an illustrative method for this chapter. In RDA the external mirror mentioned above consists of information about the group design. RDA will be compared with linear Discriminant Analysis (DA) and Principal Component - Discriminant Analysis (PC-DA) and it will be shown theoretically how and under which conditions RDA can improve group prediction. The improvement can theoretically also be expected in relation to other shrunken estimators in DA like Campbell (1980), because here the discriminant weights are estimated by ridge regression procedures. Both PC-DA and ridge regression are hybrid methods based on compound filters described in chapter 2. The *reduced constant* PC-DA filter is a discrete compound filter of a sequential hybrid method and the *ridge filter* is a continuous compound filter of a weighted hybrid method. We will confine ourselves to PC-DA being representative for hybrid methods with a compound filter. In section 4.5 we give some variations on reflecting the variance. In 4.5.1 we discuss Reflected Redundancy Analysis and two multiset extensions are briefly discussed in 4.5.2. We give Multiset Reflected Image Analysis (MRIA) and Multiset Reflected Component Analysis (MRCA). Section 4.5.3 gives nonlinear extensions of the reflected variance methods. It is shown why nonlinear reflected variance methods make new fields of application readily accessible.

4.1 Reflected Component Analysis (RCA)

In this chapter we have two sets of variables, the external variables $\mathbf{H}_\mathbf{U}$, with orthonormal basis $\mathbf{U}$ and the variables $\mathbf{H}$, with singular value decomposition $\mathbf{H}=\mathbf{P}\Phi\mathbf{Q}'$, selecting only non-zero singular values. We term $\mathbf{U}$ the mirror matrix. $\mathbf{U}$ can be equal to $\mathbf{H}_\mathbf{U}$ in the form of some orthonormal design matrix or extracted from external variables $\mathbf{H}_\mathbf{U}$. We have latent variables $\mathbf{X}$, which are a linear combination of the variables $\mathbf{H}$.

For the construction of the basic adjusted method of this chapter, Reflected Component Analysis (RCA), we integrate the *maximization* of variance accounted for with the *improvement* of squared canonical correlations. Latent variables $\mathbb{X}$, that best account for the variance of the variables $\mathbb{H}$, can be obtained by maximizing

$$\mathrm{Fit}(\mathbb{X}) = \mathrm{tr}\ \mathbb{X}'\mathbb{H}\mathbb{H}'\mathbb{X} = \mathrm{tr}\ \mathbb{X}'\mathbb{S}\mathbb{X}, \tag{4.1}$$

with $\mathbb{X}'\mathbb{X}=\mathbb{I}$ and $\mathbb{H}\mathbb{H}'=\mathbb{S}$. For the improvement of squared canonical correlations between $\mathbb{X}$ and $\mathbb{H}_U$ we use one of the CCA fit functions. Formulated in the format of MCCA (2.25) we have

^{1}MCCA: $\mathrm{Fit}(\mathbb{V}_1) = \mathrm{tr}\ \mathbb{V}_1'\mathbb{V}_1 + \mathbb{V}_1'\mathbb{P}_1'\mathbb{P}_2\mathbb{P}_2'\mathbb{P}_1\mathbb{V}_1$

with $\mathbb{X}_{(1)}'\mathbb{X}_{(1)}=\mathbb{V}_1'\mathbb{V}_1=\mathbb{I}$, and $\mathbb{V}_1=\mathbb{P}_1'\mathbb{X}_{(1)}$. Omitting the constant term $\mathbb{V}_1'\mathbb{V}_1=\mathbb{I}$, the ^{1}MCCA fit function translated in the two sets notation of this chapter is

$$\mathbb{V}'\mathbb{P}'\mathbb{U}\mathbb{U}'\mathbb{P}\mathbb{V}=\mathbb{X}'\mathbb{P}\mathbb{P}'\mathbb{U}\mathbb{U}'\mathbb{P}\mathbb{P}'\mathbb{X},$$

with $\mathbb{X}'\mathbb{X}=\mathbb{V}'\mathbb{V}=\mathbb{I}$, and $\mathbb{V}=\mathbb{P}'\mathbb{X}$. If one of the sets has only one variable, the canonical correlation is equal to the multiple correlation of this variable with the other set. For the improvement of each of the squared multiple correlations of the latent variables $\mathbb{X}$ with the external variables $\mathbb{H}_U$ we take one step of the Power Method (Wilkinson, 1965) for matrix $\mathbb{P}\mathbb{P}'\mathbb{U}\mathbb{U}'\mathbb{P}\mathbb{P}'$. Therefore the resulting *reflected latent variables*

$$\mathbb{P}\mathbb{P}'\mathbb{U}\mathbb{U}'\mathbb{P}\mathbb{P}'\mathbb{X} = \mathbb{P}\mathbb{U}\mathbb{P}\mathbb{X}, \tag{4.2}$$

with $\mathbb{P} = \mathbb{P}\mathbb{P}'$ and $\mathbb{U} = \mathbb{U}\mathbb{U}'$, all have higher squared multiple correlations with the external variables $\mathbb{H}_U$ than their respective original variables $\mathbb{X}$. Note that reflection is not conceived in the sense that we change the sign of vectors, but as a double projection. The reflected latent variables (4.2) are inserted for the original latent variables $\mathbb{X}$ in (4.1) and we obtain the Reflected Component Analysis (RCA) fit function

RCA: $$\mathrm{Fit}(\mathbb{X}) = \mathrm{tr}\ \mathbb{X}'\mathbb{P}\mathbb{U}\mathbb{S}\mathbb{U}\mathbb{P}\mathbb{X}, \tag{4.3}$$

with $\mathbb{X}'\mathbb{X}=\mathbb{I}$. Because $\mathbb{P}$ defines the orthonormal space of $\mathbb{H}$ we have $\mathbb{P}\mathbb{S}\mathbb{P}=\mathbb{S}$.

Instead of improving the squared multiple correlations of the latent variables $\mathbf{X}$ with the external variables $\mathbf{H}_\mathbf{U}$ we can also improve the squared multiple correlations of the variables $\mathbf{H}$ with these external variables. Therefore the resulting *reflected variables*

$$\mathbf{PP'UU'PP'H} = \mathbf{PP'UU'H} = \mathbf{PUH}, \tag{4.4}$$

all have higher squared multiple correlations with the external variables $\mathbf{H}_\mathbf{U}$ than their respective original variables $\mathbf{H}$. The variables in the space of $\mathbf{H}$ are projected on to the mirror space $\mathbf{U}$, which gives the *mirror variables* $\mathbf{UH}$. The mirror variables are then projected back on to the space of $\mathbf{H}$, which gives the reflected variables. The rank of the reflected variables is never higher than the rank of $\mathbf{U}$. The size of the images of the variables after reflection by the mirror matrix $\mathbf{U}$ is influenced by the angle of reflection. This is illustrated in figure 4.1 for two different reflection angles.

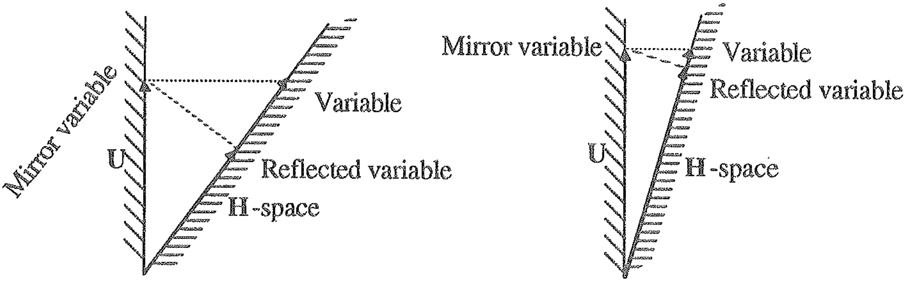


Figure 4.1 *Reflecting variables under different angles.*

It is important to bear in mind that the 'reflected variable' and the 'variable' are usually not exactly on the same line, but that they are both in the space $\mathbf{P}$ of $\mathbf{H}$. Insertion of the reflected variables (4.4) for the original variables $\mathbf{H}$ in (4.1) gives again the RCA fit function (4.3). From the geometric projections in figure 4.1 we can infer that the RCA solution can also be found by maximizing the variance of the mirror variables $\mathbf{UH}$ accounted for by latent variables $\mathbf{X}$ in the space of $\mathbf{H}$.

4.2 A filter view on RCA

The relation of RCA with the filter theory formulated in chapter 2 is given by defining the reflecting filter

RVAR: $$\Omega_k(\Phi_k^2) = \mathbb{P}_k'\mathbb{U}_k\mathbb{U}_k'\mathbb{P}_k\Phi_k^2\mathbb{P}_k'\mathbb{U}_k\mathbb{U}_k'\mathbb{P}_k. \tag{4.5}$$

For the RCA method (4.3) we insert (4.5) with k=1 and $\mathbb{U}_1$=$\mathbb{U}$ in (2.1). In section 4.5.2 we shall formulate two other multiset generalizations of RCA.

4.3 Discriminant methods

Reflected Discriminant Analysis (RDA) is formulated as a special case of RCA (4.3). The RDA method will be elaborated extensively. In RDA the external mirror matrix $\mathbb{U}$ of RCA is specified as an orthonormal group design matrix. RDA will be compared with other discriminant methods like linear Discriminant Analysis (DA) and Principal Component - Discriminant Analysis (PC-DA). The effectiveness of group prediction is assessed with the stability and exactness of group prediction and it is shown theoretically how and under which conditions RDA can improve group prediction.

The comparison between discriminant methods is greatly facilitated by an object-wise formulation of the methods with explicit latent variables. To enable this object-wise formulation we first give in section 4.3.1 a definition of Between-Within decomposition of variables. In sections 4.3.2 to 4.3.5 we give a description of the following discriminant methods:

Model	Abbreviation	Section
Discriminant Analysis	DA	4.3.2
Canonical Variate Analysis	CVA	4.3.3
Principal Component - Discriminant Analysis	PC-DA	4.3.4
Reflected Discriminant Analysis	RDA	4.3.5

A summary table with theoretical discussion of properties is provided in section 4.3.6. Two theoretically interesting special cases of RDA are presented in section 4.3.7. In section 4.4 it is shown in a simulation study that the theoretical properties of the discriminant methods can be demonstrated with simulated data.

4.3.1 External decomposition: the Between-Within decomposition

The comparison between discriminant methods is simplified by an object-wise formulation of the methods with explicit latent variables. As a preliminary step we first give a definition of *internal decomposition* of the variables $\mathbb{H}$:

$$\mathbb{H} = \mathbb{H}_1 + \mathbb{H}_2, \qquad (4.6)$$

with $\mathbb{P}\mathbb{P}'\mathbb{H}_1 = \mathbb{H}_1$,

$\mathbb{P}\mathbb{P}'\mathbb{H}_2 = \mathbb{H}_2$

and $\mathbb{H}_2'\mathbb{H}_1 = 0.$

Following previous notation the matrix $\mathbb{P}$ is derived from the SVD $\mathbb{H}=\mathbb{P}\Phi\mathbb{Q}'$, but any other orthonormal basis would also be suitable. The Eckart-Young decomposition is an example of internal decomposition. It always gives orthogonal submatrices within the orthonormal basis $\mathbb{P}$ of $\mathbb{H}$. The decomposition is internal, because the orthogonal submatrices $\mathbb{H}_1$ and $\mathbb{H}_2$ can always be expressed as linear combinations of the variables $\mathbb{H}$. The sum of the rank of $\mathbb{H}_1$ and the rank $\mathbb{H}_2$ is always equal to the rank of $\mathbb{H}$. In figure 4.2 we show an example of internal decomposition of $\mathbb{H}$ with two variables $\mathbb{m}$ and $\mathbb{n}$. After substitution of these variables in (4.6) we obtain $\mathbb{H} = (\mathbb{m},\mathbb{n}) = (\mathbb{m}_1,\mathbb{n}_1) + (\mathbb{m}_2,\mathbb{n}_2)$.

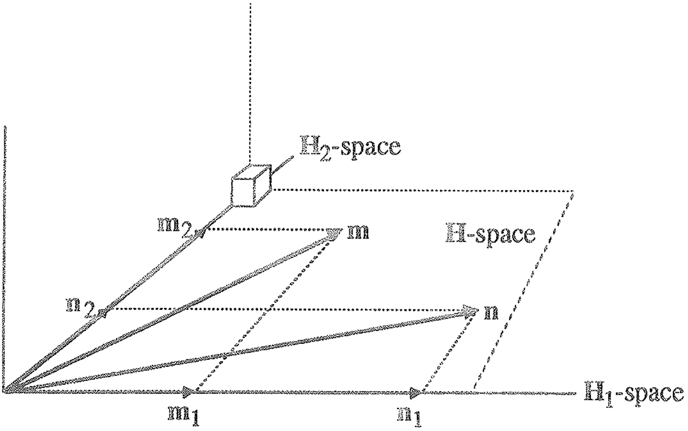


Figure 4.2 *Internal decomposition of* $\mathbb{H} = (\mathbb{m}\ ,\mathbb{n})$.

We observe in figure 4.2 that the decomposing parts $\mathbb{H}_1 = (\mathbb{m}_1,\mathbb{n}_1)$ and $\mathbb{H}_2 = (\mathbb{m}_2,\mathbb{n}_2)$ remain in the space of $\mathbb{H}$, which is a plane in this example with two variables. Within this plane $\mathbb{H}_1$ and $\mathbb{H}_2$ occupy mutually exclusive subspaces, which are in figure 4.2 two orthogonal lines.

The *external decomposition* of the variables $\mathbb{H}$ is less restricted than the internal decomposition and is defined by

$$\mathbb{H} = \mathbb{H}_1 + \mathbb{H}_2, \tag{4.7}$$

with $\mathbb{H}_2'\mathbb{H}_1 = 0.$

The decomposition is external, because the orthogonal submatrices $\mathbb{H}_1$ and $\mathbb{H}_2$ can *not* always be expressed as linear combinations of the variables $\mathbb{H}$. In figure 4.3 we show an example of external decomposition of $\mathbb{H}$ with two variables $\mathbb{m}$ and $\mathbb{n}$. After substitution of these variables in (4.7) we obtain the same decomposition formula as for figure 4.2, $\mathbb{H} = (\mathbb{m},\mathbb{n}) = (\mathbb{m}_1,\mathbb{n}_1) + (\mathbb{m}_2,\mathbb{n}_2)$.

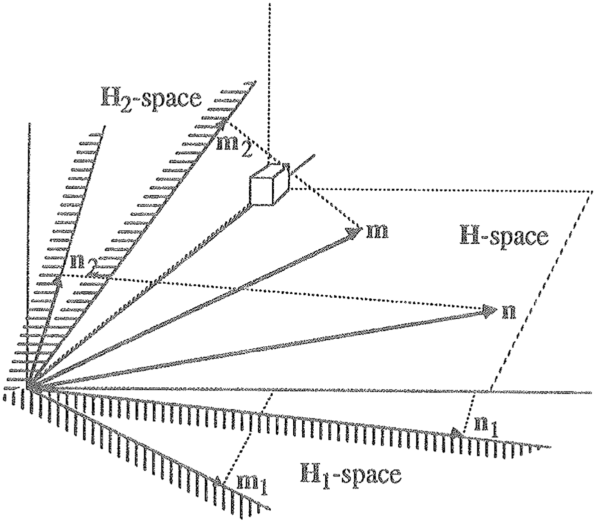


Figure 4.3 *External decomposition of* $\mathbb{H} = (\mathbb{m}, \mathbb{n})$.

The difference with figure 4.2 is that in figure 4.3 $\mathbb{H}_1$ and $\mathbb{H}_2$ are not restricted to be in the space of $\mathbb{H}$. They can be located in mutually orthogonal spaces outside the $\mathbb{H}$-space. The sum of the rank of $\mathbb{H}_1$ and the rank $\mathbb{H}_2$ can be greater than the rank of $\mathbb{H}$.

In figure 4.2 the spaces of $\mathbb{H}_1$ and $\mathbb{H}_2$ were two orthogonal lines, in figure 4.3 they are two orthogonal planes.

The Between-Within decomposition is an example of external decomposition of $\mathbb{H}$ by using

$\mathbb{G}$ the $(n{\times}g)$ orthogonal group indicator matrix for n objects and g groups, with $\mathbb{G}'\mathbb{G}=\mathbb{D}$, a diagonal matrix with group frequencies,

and the projector

$$\mathbb{G} = G(G'G)^{-1}G' = GD^{-1}G'. \qquad (4.8)$$

The *Between-Within decomposition* of the variables $\mathbb{H}$ is given and elaborated with a SVD of the two orthogonal submatrices

$$\begin{aligned} \mathbb{H} &= \mathbb{G}\mathbb{H} + (\mathbb{I}-\mathbb{G})\mathbb{H} \\ &= \mathbb{H}_B + \mathbb{H}_W \\ &= \mathbb{P}_B\Phi_B\mathbb{Q}_B' + \mathbb{P}_W\Phi_W\mathbb{Q}_W' \\ &= \mathbb{P}_B\mathbb{P}_B'\mathbb{H} + \mathbb{P}_W\mathbb{P}_W'\mathbb{H} \qquad (4.9) \end{aligned}$$

with $\mathbb{G}\mathbb{H} = \mathbb{H}_B = \mathbb{P}_B\Phi_B\mathbb{Q}_B' = \mathbb{P}_B\mathbb{P}_B'\mathbb{H}$,

and $\mathbb{H}'\mathbb{G}'(\mathbb{I}-\mathbb{G})\mathbb{H} = \mathbb{H}_B'\mathbb{H}_W = \mathbb{P}_B'\mathbb{P}_W = 0$, but not necessarily with

$$\mathbb{P}\mathbb{P}'\mathbb{H}_B = \mathbb{H}_B,$$

and $\mathbb{P}\mathbb{P}'\mathbb{H}_W = \mathbb{H}_W$.

Equation $\mathbb{H}_B=\mathbb{P}_B\mathbb{P}_B'\mathbb{H}$ holds, because $\mathbb{P}_B'\mathbb{H}_W=0$. The interpretation of the matrices $\mathbb{H}_B$ and $\mathbb{H}_W$ is very straightforward. In matrix $\mathbb{H}_B$ the elements of the variables $\mathbb{H}$ are replaced by their group means, whereas matrix $\mathbb{H}_W$ gives the deviations of the groups means. In the sequel we will denote the newly defined matrices as

Matrix	*Referred to as*
$\mathbb{H}_B$	Between-variables
$\mathbb{H}_W$	Within-variables
$\mathbb{P}_B$	between-variables space
$\mathbb{P}_W$	within-variables space

It is important to bear in mind that the four above mentioned matrices can usually not be expressed as linear combinations of the variables $\mathbb{H}$, as is illustrated in figure 4.3. After subscript substitution $1=\mathbb{B}$ and $2=\mathbb{W}$ we have a geometric example of the Between-Within decomposition with $\mathbb{H}_\mathbb{B} = (\mathbb{m}_\mathbb{B},\mathbb{n}_\mathbb{B})$, $\mathbb{H}_\mathbb{W} = (\mathbb{m}_\mathbb{W},\mathbb{n}_\mathbb{W})$, $\mathbb{P}_\mathbb{B}=\mathbb{H}_\mathbb{B}$-space and $\mathbb{P}_\mathbb{W}=\mathbb{H}_\mathbb{W}$-space.

The Between-Within decomposition of the variables $\mathbb{H}$ (4.9) including its orthogonality restriction implies for the variance-covariance matrix

$$\begin{array}{lclcll}
\mathbb{H}'\mathbb{H} & = & \mathbb{H}_\mathbb{B}'\mathbb{H}_\mathbb{B} & + & \mathbb{H}_\mathbb{W}'\mathbb{H}_\mathbb{W} & = \\
\mathbb{T} & = & \mathbb{B} & + & \mathbb{W} & = \\
\mathbb{Q}\Phi^2\mathbb{Q}' & = & \mathbb{Q}_\mathbb{B}\Phi_\mathbb{B}^2\mathbb{Q}_\mathbb{B}' & + & \mathbb{Q}_\mathbb{W}\Phi_\mathbb{W}^2\mathbb{Q}_\mathbb{W}' & = \\
 & = & \mathbb{H}'\mathbb{P}_\mathbb{B}\mathbb{P}_\mathbb{B}'\mathbb{H} & + & \mathbb{H}'\mathbb{P}_\mathbb{W}\mathbb{P}_\mathbb{W}'\mathbb{H}, &
\end{array} \tag{4.10}$$

where $\mathbb{T} = \mathbb{H}'\mathbb{H}$ denotes the total variance-covariance matrix,
$\mathbb{B} = \mathbb{H}'\mathbb{P}_\mathbb{B}\mathbb{P}_\mathbb{B}'\mathbb{H}$ denotes the between group var.-cov. matrix,
and $\mathbb{W} = \mathbb{H}'\mathbb{P}_\mathbb{W}\mathbb{P}_\mathbb{W}'\mathbb{H}$ denotes the within group var.-cov. matrix.

With the theory developed so far we know that variants of discriminant analysis based only on the analysis of $\mathbb{B}$ (or rescaled $\mathbb{B}$) usually result in optimizing linear combinations outside the space of $\mathbb{H}$. Although in a second step the optimal variable weights are very often applied to the variables of $\mathbb{H}$, these linear combinations of $\mathbb{H}$ are not optimized in the first place to predict group membership. Therefore these factors will generally be less discriminating. See for instance the method of Discriminant Principal Components Analysis (DPCA) proposed by Yendle & Macfie (1989).

4.3.2 Linear Discriminant Analysis

Linear Discriminant Analysis (Fisher, 1936) or Canonical Variate Analysis (Maxwell, 1977, p.97) maximizes the variance accounted for of between group variance divided by within group variance. Although linear Discriminant Analysis (DA) and Canonical Variate Analysis (CVA) are often formulated as identical methods, we define DA and CVA in this monograph slightly different with respect to scaling parameters. For DA we maximize the ratio of *between* to *within* sum of squares for g groups on composite variates and for CVA we maximize the ratio of *between* to *total* sum of squares.

The fit function for DA is reformulated object-wise with latent variables $\mathbf{Ha}_s$ by substituting for $\mathbf{B}$ and $\mathbf{W}$ the matrices found in (4.10)

$$\mathrm{DA}_{\mathrm{BW}}: \quad \mathrm{Fit}_{\mathrm{BW}}(\mathrm{A}) = \sum_{s=1}^{p} \frac{\mathbf{a}_s'\mathbf{B}\mathbf{a}_s}{\mathbf{a}_s'\mathbf{W}\mathbf{a}_s} = \sum_{s=1}^{p} \frac{\mathbf{a}_s'\mathbf{H}'\mathbf{P}_{\mathbf{B}}\mathbf{P}_{\mathbf{B}}'\mathbf{H}\mathbf{a}_s}{\mathbf{a}_s'\mathbf{H}'\mathbf{P}_{\mathbf{W}}\mathbf{P}_{\mathbf{W}}'\mathbf{H}\mathbf{a}_s}, \tag{4.11}$$

where $_m\mathrm{A}_p = \mathbf{a}_1,\ldots,\mathbf{a}_s,\ldots,\mathbf{a}_p,$ denote the discriminant weights for p dimensions in descending order as for the value of $\mathrm{DA}_{\mathrm{BW}}(\mathbf{a}_s)$.

After maximization of $\mathrm{Fit}_{\mathrm{BW}}(\mathrm{A})$ with normalization $\mathbf{A}'\mathbf{W}\mathbf{A}=\mathbf{I}$ gives $\mathbf{HA}$ the discriminant space with *between* to *within* normalization. The optimal discriminant weights $\mathbf{a}_s$ are the parameters of the discriminant functions $\mathbf{Ha}_s$. As mentioned by Maxwell (1977, p.98), the rank of the full discriminant space can be reduced.

4.3.3 Canonical Variate Analysis

Gittins (1985) showed that the DA solution can also be found by maximizing the B/T ratio instead of the B/W ratio. Only the normalization of the discriminant space $\mathbf{HA}$ is somewhat different. Maximizing the B/T ratio he called Canonical Variate Analysis (CVA). For each dimension the squared canonical correlation between the variable space $\mathbf{P}$ and the between-variables space $\mathbf{P}_{\mathbf{B}}$ is maximized. The fit function for CVA is reformulated object-wise with latent variables $\mathbf{Ha}_s$ by substituting for $\mathbf{B}$ and $\mathbf{T}$ the matrices found in (4.10)

$$\mathrm{CVA}_{\mathrm{BT}}: \quad \mathrm{Fit}_{\mathrm{BT}}(\mathrm{A}) = \sum_{s=1}^{p} \frac{\mathbf{a}_s'\mathbf{B}\mathbf{a}_s}{\mathbf{a}_s'\mathbf{T}\mathbf{a}_s} = \sum_{s=1}^{p} \frac{\mathbf{a}_s'\mathbf{H}'\mathbf{P}_{\mathbf{B}}\mathbf{P}_{\mathbf{B}}'\mathbf{H}\mathbf{a}_s}{\mathbf{a}_s'\mathbf{H}'\mathbf{H}\mathbf{a}_s}, \tag{4.12}$$

where $_m\mathrm{A}_p = \mathbf{a}_1,\ldots,\mathbf{a}_s,\ldots,\mathbf{a}_p,$ denote the discriminant weights for p dimensions in descending order as for the value of $\mathrm{CVA}_{\mathrm{BT}}(\mathbf{a}_s)$.

In addition we reformulate (4.12) by using the orthonormal basis $\mathbf{P}$ of $\mathbf{H}$. By inserting $\mathbf{Pv}_s$ for $\mathbf{Ha}_s$ we obtain

CVA$_{BT}$: $\text{Fit}_{BT}(\mathbb{V}) = \text{tr}\ \mathbb{V}'\mathbb{P}'\mathbb{P}_{\mathbb{B}}\mathbb{P}_{\mathbb{B}}'\mathbb{P}\mathbb{V},$ (4.13)

with $\mathbb{V}'\mathbb{P}'\mathbb{P}\mathbb{V} = \mathbb{V}'\mathbb{V} = \mathbb{I},$ for the unit orthonormalized discriminant space $\mathbb{P}\mathbb{V} = \mathbb{H}\mathbb{A}.$

> The unit orthonormalized discriminant space is given by $\mathbb{P}\mathbb{V}$ and the correlations with this D-space (sometimes called D-factor loadings) are given by $\mathbb{H}'\mathbb{P}\mathbb{V}=\mathbb{Q}\Phi\mathbb{V}$. These loadings are mostly non-orthogonal. A complete solution of Discriminant Analysis, where $\mathbb{V}$ is a square matrix with $\mathbb{V}'\mathbb{V} = \mathbb{V}\mathbb{V}' = \mathbb{I}$, gives a decomposition of the datamatrix $\mathbb{H}$ in the unit orthonormalized discriminant space and the D-factor loadings:
>
> $$\mathbb{H} = \mathbb{P}\mathbb{V}\mathbb{V}'\Phi\mathbb{Q}' = (\mathbb{P}\mathbb{V})(\mathbb{Q}\Phi\mathbb{V})'. \qquad (4.14)$$

At this point we can mould the least squares fit functions of DA and CVA in the form of a model

$$\mathbb{P}_{\mathbb{B}} = \mathbb{P}\mathbb{V}\mathbb{A}_{\mathbb{P}}' + \mathbb{E} \qquad (4.15)$$

where $\mathbb{A}_{\mathbb{P}}$ denotes the loadings of the columns of $\mathbb{P}_{\mathbb{B}}$ on the unit orthonormalized discriminant space $\mathbb{P}\mathbb{V}$. So both DA and CVA give an optimum prediction of the between-variables space $\mathbb{P}_{\mathbb{B}}$ in least squares sense.

4.3.4 Principal Component - Discriminant Analysis

In Principal Component - Discriminant Analysis (PC-DA) a DA is performed not on the original variables $\mathbb{H}$, but on a reduced rank matrix of $\mathbb{H}$. This reduced rank matrix is constructed by taking the first p principal components of a PCA solution (4.1). PCA($\mathbb{Z}$) is maximal for optimal $\mathbb{Z} = \mathbb{P}_p$, where $\mathbb{P}_p$ are the first p left singular vectors from the SVD $\mathbb{H} = \mathbb{P}\Phi\mathbb{Q}'$, with Φ in descending order and $\mathbb{P}_p'\mathbb{P}_p=\mathbb{Z}'\mathbb{Z}=\mathbb{I}$. As for the choice of p Yendle & Macfie (1989, page 595) give many tests for determining the dimensionality of the space produced by PCA. The choice of rank in the dimension reduction step introduces in fact an extra analysis step with its own problems, which we will not discuss here. The reduced rank matrix of $\mathbb{H}$ is given by

$$\mathbb{Z}\mathbb{Z}'\mathbb{H} \qquad (4.16)$$

In order to describe the PC-DA model in the same way as we did for the DA model in (4.15) we need a Between-Within decomposition of $\mathbb{Z}\mathbb{Z}'\mathbb{H}$. The decomposition is made by first decomposing $\mathbb{Z} = \mathbb{P}_p$ as we did for $\mathbb{H}$ in (4.9),

$$\begin{aligned} \mathbf{Z} &= \mathbf{Z}_B + \mathbf{Z}_W \\ &= \mathbf{K}_B\Lambda_B\mathbf{L}_B' + \mathbf{K}_W\Lambda_W\mathbf{L}_W' \\ &= \mathbf{K}_B\mathbf{K}_B'\mathbf{Z} + \mathbf{K}_W\mathbf{K}_W'\mathbf{Z}, \end{aligned} \tag{4.17}$$

With (4.17) the required Between-Within decomposition of the reduced rank matrix $\mathbf{ZZ'H}$ can easily be made

$$\mathbf{ZZ'H} = \mathbf{K}_B\mathbf{K}_B'\mathbf{ZZ'H} + \mathbf{K}_W\mathbf{K}_W'\mathbf{ZZ'H}. \tag{4.18}$$

The fit function of PC-DA is obtained by replacing $\mathbf{P}_B$ in (4.13) by $\mathbf{K}_B$:

PC-DA: $\text{Fit}(\mathbf{V}) = \text{tr } \mathbf{V'P'K}_B\mathbf{K}_B'\mathbf{PV}$ (4.19)

with $\mathbf{V'V} = \mathbf{I}$.

The corresponding PC-DA model is

$$\mathbf{K}_B = \mathbf{PVA}_K' + \mathbf{E} \tag{4.20}$$

where $\mathbf{A}_K$ denotes the loadings of the columns of $\mathbf{K}_B$ on the unit orthonormalized discriminant space $\mathbf{PV}$.

4.3.5 *Reflected Discriminant Analysis*

Reflected Discriminant Analysis (RDA) maximizes the following fit function:

RDA: $\text{Fit}(\mathbf{V}) = \text{tr } \mathbf{V'P'H}_B\mathbf{H}_B'\mathbf{PV}$ (4.21)

with $\mathbf{V'V} = \mathbf{I}$.

The corresponding RDA model is

$$\mathbf{H}_B = \mathbf{PVA}_H' + \mathbf{E} \tag{4.22}$$

where $\mathbf{A}_H$ denotes the loadings of the between-variables $\mathbf{H}_B$ on the unit orthonormalized discriminant space $\mathbf{PV}$. RDA optimizes the prediction of the between-variables $\mathbf{H}_B$ in least squares sense, whereas DA optimizes the prediction of the between-variables *space* $\mathbf{P}_B$.

To show that (4.21) maximizes reflected variance we substitute $\mathbf{G}$ (4.8) for $\mathbf{U}$ in (4.4). The reflected variables in discriminant context are given by

$$\mathbb{P}\mathbb{G}\mathbb{H} = \mathbb{P}\mathbb{H}_{\mathbb{B}}. \tag{4.23}$$

Searching for a discriminant space $\mathbb{P}\mathbb{V}$, that best accounts for the variance of the reflected variables, we obtain the RDA fit function (4.21)

RDA: $\mathrm{Fit}(\mathbb{V}) = \mathrm{tr}\ \mathbb{V}'\mathbb{P}'(\mathbb{P}\mathbb{P}'\mathbb{H}_{\mathbb{B}})(\mathbb{H}_{\mathbb{B}}'\mathbb{P}\mathbb{P}')\mathbb{P}\mathbb{V} = \mathrm{tr}\ \mathbb{V}'\mathbb{P}'\mathbb{H}_{\mathbb{B}}\mathbb{H}_{\mathbb{B}}'\mathbb{P}\mathbb{V}.$

The maximum rank unit orthonormalized discriminant space $\mathbb{P}\mathbb{V}_r$ of the RDA solution, is always a rotation of the maximum rank unit orthonormalized discriminant space $\mathbb{P}\mathbb{V}_d$ of a comparable DA solution, if $\mathbb{H}_W$ is non-singular. In matrix notation we state that $\mathbb{P}\mathbb{V}_r = \mathbb{P}\mathbb{V}_d\mathbb{C}$, with $\mathbb{C}'\mathbb{C}=\mathbb{C}\mathbb{C}'=\mathbb{I}$.

Proof. We exclude cases with singular $\mathbb{H}_W$, because in that case the DA solution degenerates. The optimal unit orthonormalized $\mathbb{V}_r$ of the RDA solution (4.21) are equal to the left singular vectors of $\mathbb{P}'\mathbb{H}_{\mathbb{B}}=\mathbb{P}'\mathbb{P}_{\mathbb{B}}\Phi_{\mathbb{B}}\mathbb{Q}_{\mathbb{B}}'$ and equal to the left singular vectors of $\mathbb{P}'\mathbb{P}_{\mathbb{B}}\Phi_{\mathbb{B}}$, because $\mathbb{Q}_{\mathbb{B}}'\mathbb{Q}_{\mathbb{B}}=\mathbb{I}$. We compute the optimal unit orthonormalized $\mathbb{V}_d$ of the DA solution by maximizing (4.13) and they are equal to the left singular vectors of $\mathbb{P}'\mathbb{P}_{\mathbb{B}}$. The left singular vectors $\mathbb{V}_r$ and $\mathbb{V}_d$ are an orthonormal basis for *both* $\mathbb{P}'\mathbb{P}_{\mathbb{B}}\Phi_{\mathbb{B}}$ *and* $\mathbb{P}'\mathbb{P}_{\mathbb{B}}$, because $\Phi_{\mathbb{B}}$ is a diagonal matrix with diagonal values >0. This implies that $\mathbb{P}\mathbb{V}_r = \mathbb{P}\mathbb{V}_d\mathbb{C}$, with $\mathbb{C}'\mathbb{C}=\mathbb{C}\mathbb{C}'=\mathbb{I}$. □

If the optimal DA discriminant space $\mathbb{P}\mathbb{V}_d$ exists, the optimal $\mathbb{C}$ for computing $\mathbb{V}_r$ with $\mathbb{V}_d\mathbb{C}$ are equal to the eigenvectors of $\mathbb{V}_d'\mathbb{P}'\mathbb{H}_{\mathbb{B}}\mathbb{H}_{\mathbb{B}}'\mathbb{P}\mathbb{V}_d$. In this way the optimal DA discriminant space is rotated in RDA in such a way that the group or 'between' variance accounted for by the successive optimal RDA variates is decreasing. We can expect that the stability of group prediction will also decrease.

4.3.6 Summary of models and expected properties

In table 4.1 we give a summary of the discriminant models and fit functions of the preceding sections. The PCA model is also included, because it is needed as a first step in the sequential hybrid PC-DA model.

The summary table 4.1 makes it easy to compare the different discriminant methods. We discuss two theoretical properties which are indirectly related to the effectiveness of group prediction. The first property is *exactness* of group prediction by filtering out irrelevant 'within' information, and the second one is stability of group prediction by avoiding solutions in spurious regions, i.e. solutions with very small variance accounted for.

Table 4.1 *Summary of discriminant models.*

Name	Model	Fit function
DA(CVA)	$\mathbf{P_B} = \mathbf{PVA_P}' + \mathbf{E}$	tr $\mathbf{V'P'P_BP_B'P\,V}$
(PCA)	$\mathbf{H_B} + \mathbf{H_W} = \mathbf{ZA}' + \mathbf{E}$	tr $\mathbf{Z}'(\mathbf{H_BH_B}'+\mathbf{H_WH_W}')\mathbf{Z}$
PC-DA	$\mathbf{K_B} = \mathbf{PVA_K}' + \mathbf{E}$	tr $\mathbf{V'P'K_BK_B'P\,V}$
RDA	$\mathbf{H_B} = \mathbf{PVA_H}' + \mathbf{E}$	tr $\mathbf{V'P'H_BH_B'PV}$ = tr $\mathbf{V'P'P_B\Phi_B^2P_B'P\,V}$

with $\mathbf{H} = \mathbf{P\Phi Q}' = \mathbf{H_B} + \mathbf{H_W}$,

$\mathbf{H_B}'\mathbf{H_B} = \mathbf{B}$,

$\mathbf{H_B} = \mathbf{P_B\Phi_BQ_B}'$,

$\mathbf{Z} = \mathbf{Z_B} + \mathbf{Z_W}$, with $\mathbf{Z'Z} = \mathbf{I}$,

and $\mathbf{Z_B} = \mathbf{K_B\Lambda_BL_B}'$.

If we look at the exactness of group prediction by filtering out irrelevant 'within' information DA, PC-DA and RDA all seem to predict only group or 'between' information. But if we look at the first PCA step of the PC-DA solution we see that in making the reduced rank matrix the PCA solution $\mathbf{Z}$ can capitalize on within information if the within variance comprises a substantial part of the total variance. To make this clear we replaced the matrix $\mathbf{H}$ in table 4.1 by the orthogonal decomposition $\mathbf{H_B} + \mathbf{H_W}$. So on the whole, exactness of group prediction is not optimal for PC-DA.

The stability of group prediction by avoiding spurious regions is only effective if the method in some way capitalizes on the variance of $\mathbf{H}$. In other words the method has to be dependent on the scaling of the variables. This is the case for PC-DA. RDA even capitalizes on the variance of the relevant between part of $\mathbf{H}$, but DA turns out badly in this respect, because it is independent of scale and not interested in variance whatsoever.

Table 4.2 *Effectiveness of group prediction.*

	DA	PC-DA	RDA
Exactness of group prediction:	+	–	+
Stability of group prediction:	–	+	+

In table 4.2 we give a summary of the effectiveness of group prediction that can be theoretically expected with respect to exactness and stability of group prediction. From this summary we can expect RDA to have more effective group prediction compared to DA and PC-DA.

4.3.7 Six special cases of RDA

Six theoretically interesting special cases of RDA are presented, because they show the intricate integration of set correlation and set variance in RDA. The six cases are: The linear independence case for the variables, where the variables of $\mathbf{H}$ are not correlated. The complete rank case for the variables, where ${}_n\mathbf{H}_m$ has n non-zero eigenvalues. The 2 group and the n group case. The p=1 and p=g–1 case.

The linear independence case for the variables

RDA gives the same solution as linear DA if all variables are linear independent and have the same normalization.

Proof. In the linear independent case we have $\mathbf{H}=c\mathbf{P}$ and without loss of generality we take c=1. From (4.9) we have the equality $\mathbf{H_B}=\mathbf{P_B}\mathbf{P_B}'\mathbf{H}$. Substitution in (4.21) gives

RDA: $\quad \text{Fit}(\mathbf{V}) = \text{tr }\mathbf{V}'\mathbf{P}'\mathbf{P_B}\mathbf{P_B}'\mathbf{H}\mathbf{H}'\mathbf{P_B}\mathbf{P_B}'\mathbf{P}\mathbf{V}$ (4.24)

with $\quad \mathbf{V}'\mathbf{V} = \mathbf{I}.$

After substituting $\mathbf{H}=\mathbf{P}$ maximizing (4.24) boils down to finding the first p eigenvectors of $\mathbf{P}'\mathbf{P_B}\mathbf{P_B}'\mathbf{P}\mathbf{P}'\mathbf{P_B}\mathbf{P_B}'\mathbf{P}$, which is equal to the first p eigenvectors of $\mathbf{P}'\mathbf{P_B}\mathbf{P_B}'\mathbf{P}$. These eigenvectors also give the solution for the maximization of DA in (4.13).

□

In general we can say that the RDA and DA solution are equal if all variables are uncorrelated and have the same normalization and that these solutions *can* diverge more if there is more linear dependence between the predictor variables, but this is not necessary.

The complete rank case for the variables

RDA gives the same solution as a PCA of the between variables $\mathbb{H}_B$, if ${}_n\mathbb{H}_m$ has n non-zero eigenvalues (only possible if $m \geq n$). This implies that all within group variances in the discriminant space $\mathbb{PV}$ of formula (4.21) are zero and that the linear DA solution is degenerated.

Proof. In the complete rank case we have $\mathbb{PP}' = \mathbb{P} = \mathbb{I}$. As we saw in section 4.3.5 the reflected variables in discriminant context are given by $\mathbb{PGH} = \mathbb{PH}_B$ (4.23). In this section it was also shown that RDA maximizes reflected variance accounted for by the discriminant space $\mathbb{PV}$. In the complete rank case after substitution of $\mathbb{P} = \mathbb{I}$ in (4.23) this comes down to maximizing the variance of the between variables $\mathbb{H}_B$ accounted for by an unrestricted discriminant space $\mathbb{PV}$. □

The $(n \times m)$ datamatrix $\mathbb{H}$ has in many applications of DA many more columns than rows and in that case the SVD $\mathbb{H} = \mathbb{P}\Phi\mathbb{Q}'$ has usually an orthonormal singular vector matrix $\mathbb{P}$ $(n \times p)$ with $p=n$ or, if the columns of $\mathbb{H}$ have zero mean, $p=n\text{-}1$ non-zero singular values. In the n-1 case the columns of $\mathbb{P}$ in formula (4.23) can without loss of generality also be completed with an extra column to a full square orthonormal matrix, because in the $n-1$ case the columns of $\mathbb{PGH}$ are the result of two consecutive projections of $\mathbb{H}$ on spaces that fully contain the vector with elements 1 and therefore the columns of $\mathbb{PGH}$ also have zero mean.

The 2 group and the n group case

For g=2, RDA gives the same solution as linear DA.

Proof. For g=2 $\mathbb{H}_B$ has always rank one and can be written as $\mathbb{h}_B$. This implies that $\mathbb{h}_B = \mathbb{p}_B$ and in table 4.1 we can verify that the models and fit functions of RDA and DA become identical. □

For $g{=}n$, RDA gives the same solution as PCA.

Proof. For $g{=}n$ we have $\mathbb{G} = \mathbb{I}$ (4.8). In this case the reflected variables are given by $\mathbb{PGH} = \mathbb{H}$ (4.23). Because RDA maximizes reflected variance accounted for by the discriminant space $\mathbb{PV}$, it will in this case maximize tr $\mathbb{V}'\mathbb{P}'\mathbb{HH}'\mathbb{PV}$, which is the fit function for PCA. □

The $p{=}1$ and $p{=}g{-}1$ case

For $p{=}g{-}1$ the optimal RDA discriminant space, is always a rotation of a comparable optimal DA discriminant space, if $\mathbb{H}_W$ is non-singular. The optimal DA discriminant space is rotated in such a way that the group or 'between' variance accounted for by the successive optimal RDA variates is decreasing. We can expect that the stability of group prediction will also decrease for the respective RDA variates.

Proof. Because the maximum rank of $\mathbb{H}_B$ is $g{-}1$, the maximum rank for the optimal RDA and DA discriminant space we will always be $g{-}1$. Therefore we know that for $p{=}g{-}1$ the discriminant space of RDA and DA has maximum rank. For this condition the previous statement is proved at the end of section 4.3.5. □

For $p{=}1$ we can expect that the stability of RDA group prediction will increase compared to DA group prediction if the number of groups g increases. This statement can be derived from the previous '$p{=}g{-}1$' statement.

4.4 Simulation study of discriminant methods

The theoretical properties of the discriminant methods derived in section 4.3.6 and 4.3.7 can be investigated with simulated data. If the properties have predictive value RDA must generally give better group predictions than both DA and PC-DA. In the complete rank case for the variables a comparison between RDA and DA is not possible, because the DA solution degenerates. In chapter 7 we give a real-life example of a complete rank case for the variables. The previous section indicates an increasing stability of RDA group prediction compared to DA, if the number of groups increases and if the RDA solution is of reduced rank. Therefore we decided to explore the predictive properties of DA, PC-DA and RDA for 5 groups ($g{=}5$) and 2 dimensions (p=2). Of course this limited simulation study needs to be extended in the future.

In section 4.4.1 we discuss the construction of the artificial data, containing a true between-group configuration, a true within-group configuration and an error part. In 4.4.2 we give some measures of recovery of the true between-group configuration and in 4.4.3 the results are presented. The expected differences in group prediction are verified by calculating the leaving-one-out error rate for DA, PC-DA and RDA.

4.4.1 Construction of artificial data

The artificial data $\mathbf{H}_{\mathrm{art}}$ are decomposed in three parts, a true between-group configuration, a true within-group configuration and an error part.

For the construction of the true object configuration we rewrite the Between-Within decomposition in (4.9) as follows

$$\begin{aligned}\mathbf{H} &= \mathbf{P}_{\mathrm{B}}\Phi_{\mathrm{B}}\mathbf{Q}_{\mathrm{B}}' + \mathbf{P}_{\mathrm{W}}\Phi_{\mathrm{W}}\mathbf{Q}_{\mathrm{W}}' \\ &= (\mathbf{P}_{\mathrm{B}}\Phi_{\mathrm{B}},\mathbf{P}_{\mathrm{W}}\Phi_{\mathrm{W}})(\mathbf{Q}_{\mathrm{B}}',\mathbf{Q}_{\mathrm{W}}'). \qquad (4.25)\end{aligned}$$

By taking the matrix $(\mathbf{P}_{\mathrm{B}}\Phi_{\mathrm{B}},\mathbf{P}_{\mathrm{W}}\Phi_{\mathrm{W}})$, with $\mathbf{Q}_{\mathrm{B}}'\mathbf{Q}_{\mathrm{B}}=\mathbf{I}$ and $\mathbf{Q}_{\mathrm{W}}'\mathbf{Q}_{\mathrm{W}}=\mathbf{I}$, as our true object configuration instead of $\mathbf{H}$ we can separate the between and the within part nicely in different latent variables. In this way we can control the distribution of the error over the between-variables space $\mathbf{P}_{\mathrm{B}}$ and the within-variables space $\mathbf{P}_{\mathrm{W}}$. We can use $(\mathbf{P}_{\mathrm{B}}\Phi_{\mathrm{B}},\mathbf{P}_{\mathrm{W}}\Phi_{\mathrm{W}})$ for our simulation study, because we still have the equality $(\mathbf{P}_{\mathrm{B}}\Phi_{\mathrm{B}},\mathbf{P}_{\mathrm{W}}\Phi_{\mathrm{W}})(\mathbf{P}_{\mathrm{B}}\Phi_{\mathrm{B}},\mathbf{P}_{\mathrm{W}}\Phi_{\mathrm{W}})'=\mathbf{H}\mathbf{H}'$, with $\mathrm{tr}\mathbf{H}\mathbf{H}' = \mathrm{tr}\Phi_{\mathrm{B}}^2 + \mathrm{tr}\Phi_{\mathrm{W}}^2 = \mathrm{tr}\mathbf{T} = \mathrm{tr}\mathbf{B} + \mathrm{tr}\mathbf{W}$. From $(\mathbf{P}_{\mathrm{B}}\Phi_{\mathrm{B}},\mathbf{P}_{\mathrm{W}}\Phi_{\mathrm{W}})$ we can derive a lower bound for the number of variables m of $\mathbf{H}$, because $m \geq (\mathbf{H})_{\mathrm{rank}} \geq ((\mathbf{P}_{\mathrm{B}})_{\mathrm{rank}},(\mathbf{P}_{\mathrm{W}})_{\mathrm{rank}})_{\max}$, with $(\mathbf{H})_{\mathrm{rank}}$ giving the rank of $\mathbf{H}$ and $(a,b)_{\max}$ giving the maximum value of a and b.

By adding error $\mathbf{E}_b$ and $\mathbf{E}_w$ to respectively $\mathbf{P}_{\mathrm{B}}$ and $\mathbf{P}_{\mathrm{W}}$ we have set up the following decomposition of the artificial data:

$$\mathbf{H}_{\mathrm{art}} = (\mathbf{P}_{\mathrm{B}}\Phi_{\mathrm{B}},\mathbf{P}_{\mathrm{W}}\Phi_{\mathrm{W}}) + \mathbf{E} = (\mathbf{P}_{\mathrm{B}}\Phi_{\mathrm{B}},\mathbf{P}_{\mathrm{W}}\Phi_{\mathrm{W}}) + (\mathbf{E}_b\Phi_{\mathrm{B}},\mathbf{E}_w\Phi_{\mathrm{W}}) \qquad (4.26)$$

Note that generally the error can not be simulated by omitting $\mathbf{E}$ and changing the values of Φ_{B} and Φ_{W}. The error can only be simulated this way if the rank of $\mathbf{H}_{\mathrm{art}}$ is

$n-1$, because the maximum rank of $\mathbb{P}_\mathbb{B}$ is $g-1$ and the maximum rank of $\mathbb{P}_\mathbb{W}$ is $n-g+1$.

For $\mathbb{H}_{\text{art}}$ we chose 40 objects and 5 groups, with 8 objects in each group. The true between-group configuration $\mathbb{P}_\mathbb{B}\Phi_\mathbb{B}$ had 4 latent variables with weights $\Phi_\mathbb{B}^2$ proportional to 5, 4, 2, and 1. This gave a moderate gap between the second and third eigenvalue. The true within-group configuration $\mathbb{P}_\mathbb{W}\Phi_\mathbb{W}$ had also 4 latent variables with weights $\Phi_\mathbb{W}^2$ proportional to 5, 4, 2, and 1. Therefore the minimum number of variables is 4.

Two factors were systematically changed during the construction of the artificial data $\mathbb{H}_{\text{art}}$ (4.26). This were the Between-to-total ratio BT and the error level. The Between-to-total ratio

$$BT = \frac{\text{tr } \mathbb{B}}{\text{tr } \mathbb{T}} = \frac{\text{tr } \Phi_\mathbb{B}^2}{\text{tr } (\Phi_\mathbb{B}^2 + \Phi_\mathbb{W}^2)}, \tag{4.27}$$

had the values $BT = 1 \;\; 0.4 \;\; 0.2 \;\; 0.1$. The diagonal matrices $\Phi_\mathbb{B}$ and $\Phi_\mathbb{W}$ in (4.26) were computed as follows: $\Phi_\mathbb{B}=BT^{1/2}\Phi_{\text{sim}}$ and $\Phi_\mathbb{W}=(1-BT)^{1/2}\Phi_{\text{sim}}$, where Φ_{sim}^2 is a diagonal simulation matrix with eigenvalues 5, 4, 2, and 1. The error level for $\mathbb{E}_b$ and $\mathbb{E}_w$ is equal to the standard deviation of a unit normalized random normal variable. We chose error level = 0 0.1 0.4 0.7. For each combination of BT and error level we computed three reconstruction measures for 150 artificial configurations. Each of the 150 configurations was constructed with a different error. The three reconstruction measures are given in the following section.

4.4.2 *Measures of recovery*

In the simulation study we explored only reduced rank predictions and computed 2 dimensional solutions. The recovery measures were all adapted to this restriction by dividing each fit measure by the upper bound for two dimensions. In this way we defined a set correlation DA measure *DA2*, a set variance measure *V2*, and a reflected variance measure *RV2*. Successively

$$DA2 = \frac{\text{tr } \mathbb{V}'\mathbb{P}'\mathbb{P}_\mathbb{B}\mathbb{P}_\mathbb{B}'\mathbb{P}\,\mathbb{V}}{2}, \tag{4.28}$$

denotes the variance of the orthonormal basis $\mathbf{P}_\mathrm{B}$ of the true between-group configuration accounted for by the discriminant space $\mathbf{PV}$, divided by the upper bound for a 2 dimensional solution. *DA2* indicates the exactness of group prediction.

$$V2 = \frac{\mathrm{tr}\ \mathbf{V}'\mathbf{P}'\mathbf{H}\mathbf{H}'\mathbf{P}\mathbf{V}}{\phi_1^2 + \phi_2^2} = \frac{\mathrm{tr}\ \mathbf{V}'\mathbf{P}'(\mathbf{H}_\mathrm{B}\mathbf{H}_\mathrm{B}'+\mathbf{H}_\mathrm{W}\mathbf{H}_\mathrm{W}')\mathbf{P}\mathbf{V}}{\phi_1^2 + \phi_2^2}, \tag{4.29}$$

denotes the variance of the true object configuration $(\mathbf{P}_\mathrm{B}\Phi_\mathrm{B},\mathbf{P}_\mathrm{W}\Phi_\mathrm{W})$ accounted for by the discriminant space $\mathbf{PV}$, divided by the upper bound for a 2 dimensional solution. ϕ_1^2 and ϕ_2^2 denote the largest two eigenvalues of $\mathbf{H}'\mathbf{H}=\mathbf{T}$. *V2* indicates the stability of group prediction.

$$RV2 = \frac{\mathrm{tr}\ \mathbf{V}'\mathbf{P}'\mathbf{H}_\mathrm{B}\mathbf{H}_\mathrm{B}'\mathbf{P}\mathbf{V}}{\phi_{\mathrm{B}1}^2 + \phi_{\mathrm{B}2}^2}, \tag{4.30}$$

denotes the variance of the true between-group configuration $\mathbf{P}_\mathrm{B}\Phi_\mathrm{B}$ accounted for by the latent variates, divided by their upper bound for a 2 dimensional solution. $\phi_{\mathrm{B}1}^2$ and $\phi_{\mathrm{B}2}^2$ denote the largest two eigenvalues of $\mathbf{H}_\mathrm{B}'\mathbf{H}_\mathrm{B}=\mathbf{B}$. *RV2* is a simple integrated measure for the exactness and stability of group prediction.

For each constructed configuration $\mathbf{H}_\mathrm{art}$ (4.26) we computed a DA solution, a PC-DA solution and a RDA solution. The three measures mentioned above are computed for each solution. The reduced rank matrix for PC-DA was made by skipping all components of $\mathbf{H}_\mathrm{art}$ with eigenvalues smaller than one eighth of the total variance of the true object configuration $(\mathbf{P}_\mathrm{B}\Phi_\mathrm{B},\mathbf{P}_\mathrm{W}\Phi_\mathrm{W})$, which is $\mathrm{tr}(\Phi_\mathrm{B}^2 + \Phi_\mathrm{W}^2)$.

4.4.3 Results

In the following three tables we present the mean values of the recovery measures *DA2* (4.28), *V2* (4.29) and *RV2* (4.30) over 150 constructed configurations for each combination of *BT* and error level. The DA, PC-DA and RDA solutions are computed using the same 150 constructed configurations.

Table 4.3 *DA2: recovery of the orthonormal basis* $\mathbf{P}_B$ *of the true between-group configuration* $\mathbf{P}_B\mathbf{\Phi}_B$.

n=150		Between-to-total ratio *BT*			
Error level		1	0.4	0.2	0.1
0	DA:	1	1	1	1
	PC-DA:	1	1	0	0
	RDA:	1	1	1	1
0.1	DA:	0.94	0.95	0.95	0.95
	PC-DA:	0.93	0.88	0.03	0.02
	RDA:	0.92	0.93	0.93	0.93
0.4	DA:	0.82	0.84	0.84	0.84
	PC-DA:	0.80	0.73	0.19	0.05
	RDA:	0.77	0.79	0.79	0.80
0.7	DA:	0.73	0.76	0.76	0.76
	PC-DA:	0.72	0.66	0.33	0.09
	RDA:	0.69	0.71	0.72	0.71

The results for *DA2* in table 4.3 show that, as expected from section 4.3.6, the solutions of PC-DA degenerate if the proportion of within variance becomes to large. In that case the solution capitalizes on the within variance in the dimension reduction step. The differences between DA and RDA are never larger than 5% with respect to the *DA2* function and of course always higher for DA because it maximizes (4.13).

Table 4.4 *V2: recovery of the true configuration* $\mathbf{P}_B\mathbf{\Phi}_B,\mathbf{P}_W\mathbf{\Phi}_W$.

n=150		Between-to-total ratio *BT*			
Error level		1	0.4	0.2	0.1
0	DA:	–	–	–	–
	PC-DA:	1	0.67	0.94	0.84
	RDA:	1	0.67	0.25	0.11
0.1	DA:	0.61	0.42	0.16	0.07
	PC-DA:	0.74	0.61	0.70	0.73
	RDA:	0.91	0.61	0.23	0.10
0.4	DA:	0.54	0.38	0.14	0.07
	PC-DA:	0.62	0.50	0.49	0.57
	RDA:	0.72	0.50	0.19	0.08
0.7	DA:	0.48	0.34	0.14	0.07
	PC-DA:	0.50	0.43	0.32	0.43
	RDA:	0.60	0.43	0.17	0.08

In table 4.4 it is shown without doubt that the PC-DA solution capitalizes on the within variance in the dimension reduction step. As for the comparison of DA and RDA we find, as expected from section 4.3.6, that the variance accounted for is remarkably higher for the RDA solution, although the DA function is not much lower (Table 4.3). This could be expected because RDA avoids spurious regions, whereas DA is indifferent with respect to variance accounted for. This is also the reason why the DA solutions for error level zero are not uniquely defined. There are many solutions possible with a perfect fit for one specific constructed configuration.

Table 4.5 *RV2: recovery of the true between-group configuration* $\mathbf{P}_{\mathbf{B}}\mathbf{\Phi}_{\mathbf{B}}$.

n=150		Between-to-total ratio *BT*			
Error level		1	0.4	0.2	0.1
	DA:	–	–	–	–
0	PC-DA:	1	1	0	0
	RDA:	1	1	1	1
	DA:	0.61	0.63	0.62	0.63
0.1	PC-DA:	0.74	0.88	0.03	0.01
	RDA:	0.91	0.92	0.92	0.92
	DA:	0.54	0.56	0.56	0.55
0.4	PC-DA:	0.62	0.71	0.18	0.03
	RDA:	0.72	0.74	0.73	0.70
	DA:	0.48	0.50	0.51	0.50
0.7	PC-DA:	0.50	0.61	0.32	0.06
	RDA:	0.60	0.63	0.62	0.58

In RDA the between variance is predicted (4.22). In table 4.5 RDA clearly predicts the $\mathbf{B}$ matrix better than both DA and PC-DA. The expected improvement of group prediction can be verified by calculating a measure of misclassification of group prediction independent of the discriminant method. We have chosen for this purpose the Leaving-One-Out (L-O-O) error rate for the reduced rank group prediction with two dimensional discriminant space

$$LOO2 = \frac{\text{misclassified objects}}{\text{total number of objects}}, \tag{4.31}$$

The L-O-O error rate is calculated by omitting one object from the raw data prior to the discriminant analysis, projecting the object into the resulting discriminant space, computing the distances to the group centroids, and finally classifying the object with respect to the minimum distance. This is repeated for all objects in the raw data, and the L-O-O error rate is given by the fraction of these objects that are misclassified. The discriminant space is scaled dimensionwise with the square root of the eigenvalues maximizing (4.11) and for RDA with the square root of the eigenvalues maximizing (4.21) (see Maxwell, 1977, p.99). The *LOO2* error rates (4.31) are computed for the same 150 constructed configurations as for table 4.3, 4.4 and 4.5. For each cell in table 4.6 we give the *mean value* of the 150 computed *LOO2* error rates.

Table 4.6 *Mean LOO2: measure of misclassification of group prediction.*

n=150		Between-to-total ratio *BT*			
Error level		1	0.4	0.2	0.1
	DA:	0.10	0.12	0.13	0.11
0.1	PC-DA:	0.09	0.12	0.96	0.97
	RDA:	0.05	0.06	0.05	0.05
	DA:	0.29	0.33	0.35	0.34
0.4	PC-DA:	0.28	0.34	0.85	0.92
	RDA:	0.26	0.31	0.31	0.30
	DA:	0.40	0.45	0.44	0.44
0.7	PC-DA:	0.39	0.43	0.75	0.89
	RDA:	0.40	0.44	0.43	0.46

Generally prediction is better for low error levels. The most striking differences between DA and RDA in table 4.6 are found on the 0.1 error level. RDA group prediction is better than PC-DA and DA prediction, even for *BT*=0.1, were the within matrix W is far from singular. We give the standard deviations for the 0.1 error level.

Table 4.7 *Standard deviations of LOO2 for 0.1 error level.*

n=150		Between-to-total ratio *BT*			
Error level		1	0.4	0.2	0.1
	DA:	0.08	0.08	0.09	0.08
0.1	PC-DA:	0.07	0.11	0.03	0.03
	RDA:	0.05	0.04	0.04	0.04

To give a more detailed impression of the differences between DA and RDA, we show in figure 4.4 the frequency distribution of *LOO2* (4.31) for the 150 constructed configurations with *BT*=1 and with 0.1 error level. The error rates have a discrete distribution, because the artificial data have only 40 objects.

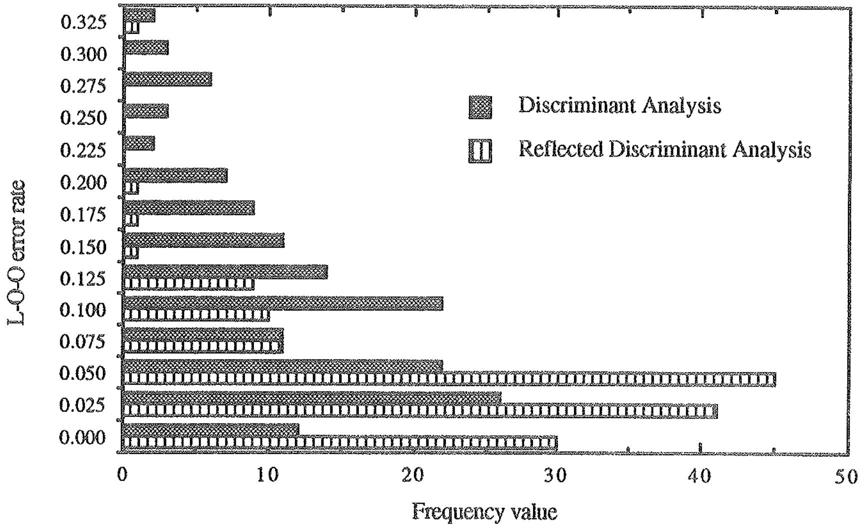


Figure 4.4 *Frequency distribution of LOO2 for DA and RDA.*

To investigate on the pairwise difference in group prediction between DA and RDA for the same 150 constructed configurations with *BT*=1 and with 0.1 error level, we show in figure 4.5 the frequency distribution of the error rate of DA minus the error rate of RDA.

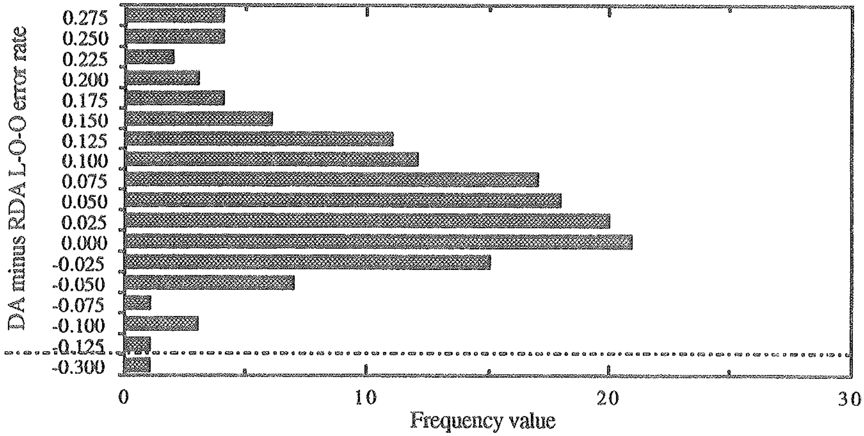


Figure 4.5 *Frequency distribution of LOO2 for DA minus RDA.*

LOO2 difference values of DA minus RDA in figure 4.5 greater than zero indicate that RDA predicts better than DA. The reverse is true for values smaller than zero. There is only one constructed configuration of the 150 where DA performs clearly better than RDA. In this case the correctly classified percentage is 30% higher for DA compared with RDA. From the total distribution in figure 4.5 it is obvious that RDA generally gave a higher percentage correct classification than DA.

Summarising the results of our simulation study we rely most on the differences found with the Leaving-One-Out method. The L-O-O error rates presented above indicate that RDA gives a better group prediction than PC-DA and DA for 5 groups and 2 dimensions, especially for small amounts of random error.

4.5 Some variations on reflecting variance

The principle of reflecting variance can be used to formulate a variety of new methods, but it is not yet clear which extensions have practical use. In section 4.5.1 we offer a two sets example and in 4.5.2 some multiset examples. A promising extension seems to be the introduction of nonlinear transformations for the variables in the reflected variance methods. In section 4.5.3 nonlinear extensions in the line of Gifi (1990) are discussed and illustrated with Reflected Discriminant Analysis.

4.5.1 *Reflected Redundancy Analysis (RRA)*

Reflected Redundancy Analysis for predictor set *c* and criterion set *k*:

$$\text{RRA:} \qquad \text{Fit}(\mathbb{X}) = \text{tr } \mathbb{X}'\mathbb{P}_c\mathbb{P}_k\mathbb{S}_c\mathbb{P}_k\mathbb{P}_c\mathbb{X} + \mathbb{X}'\mathbb{P}_c\mathbb{S}_k\mathbb{P}_c\mathbb{X}, \tag{4.32}$$

with $\mathbb{X}'\mathbb{X}=\mathbb{I}$. The second term on the right-hand side of equation (4.32) gives the Redundancy Analysis fit function as discussed in section 2.3.1 with $\mathbb{S}_k=\mathbb{H}_k\mathbb{H}_k'$ for criterion set *k* and $\mathbb{P}_c\mathbb{X} = \mathbb{Z}_c = \mathbb{H}_c\mathbb{T}_c$ for predictor set *c*. The left part of (4.32) gives the RCA fit function as given in (4.3) with $\mathbb{S}_c=\mathbb{S}$, $\mathbb{P}_k=\mathbb{U}$ and $\mathbb{P}_c=\mathbb{P}$. This part ensures that the predictor space $\mathbb{X}$ is indirectly related to the variance of the predictor set *c* and thereby can stabilize the solution.

4.5.2 *Multiset Reflected Variance*

We give two examples of multiset reflected variance. The first one is Multiset Reflected Image Analysis:

$$\text{MRIA:} \quad \mathrm{Fit}(\mathbb{X}) = \mathrm{tr} \sum_{k=1}^{K} w_k^{-1} \, \mathbb{X}'\mathbb{P}_{-k}\mathbb{S}_k'\mathbb{P}_{-k}\mathbb{X}, \tag{4.33}$$

with $\mathbb{X}'\mathbb{X} = \mathbb{I}$,

where ${}_n\mathbb{X}_p = (\mathbb{x}_1,\ldots,\mathbb{x}_s,\ldots,\mathbb{x}_p)$ denote the common latent variables,
$w_1,\ldots,w_k,\ldots,w_K$ denote fixed balancing constants for set k.
$\mathbb{P}_{-k}$ denotes the projector on to the space spanned by all sets with exception of k.

This method is related to Generalized Image Analysis (GIA) proposed by Van de Geer (1986) with respect to the definition of unique variances for each set. In GIA the variables $\mathbb{H}_k$ of each set k are decomposed externally in a unique part $(\mathbb{I}-\mathbb{P}_{-k})\mathbb{H}_k$ and a non unique part $\mathbb{P}_{-k}\mathbb{H}_k$. (For external decomposition, see (4.7) and further.) GIA maximizes the variance of all $\mathbb{H}_k$ accounted for by $\mathbb{x}$, divided by the variance of the unique parts $(\mathbb{I}-\mathbb{P}_{-k})\mathbb{H}_k$ accounted for by $\mathbb{x}$. MRIA maximizes the variance of the non unique parts $\mathbb{P}_{-k}\mathbb{H}_k$ accounted for by $\mathbb{x}$. The balancing constants w_k emphasize the necessity of an appropriate normalization of the sets (see chapter 2). The second example of multiset reflected variance in the same notation is Multiset Reflected Component Analysis:

$$\text{MRCA:} \quad \mathrm{Fit}(\mathbb{X},\mathbb{Z}_k) = \mathrm{tr} \sum_{k=1}^{K} w_k^{-1} \, \mathbb{Z}_k'\mathbb{X}\mathbb{X}'\mathbb{S}_k'\mathbb{X}\mathbb{X}'\mathbb{Z}_k, \tag{4.34}$$

with $\mathbb{X}'\mathbb{X} = \mathbb{I}$ and $\mathbb{Z}_k'\mathbb{Z}_k = \mathbb{I}$, $\forall k$,

where $\mathbb{Z}_k = (\mathbb{z}_{(k)1},..,\mathbb{z}_{(k)s},..,\mathbb{z}_{(k)p})$ denote the unit orthonormalized variates for set k and dimension s, so $\mathbb{Z}_k = \mathbb{H}_k\mathbb{A}_k$.

In some practical applications it might be interesting to take the dimensionality of $\mathbb{X}$ somewhat higher than the dimensionality of $\mathbb{Z}_k$.

4.5.3 *Nonlinear Reflected Variance*

All reflected variance methods presented in this chapter can be reformulated in such a way that nonlinear transformations of the variables are optimized. In Gifi (1990) a general framework is given for such an operation. Three general types of discrete nonlinear transformations are distinguished: no transformation, monotone transformation and preserving category membership transformation. The type of transformation is indicated by the scaling level of the variable, respectively numerical, ordinal and nominal.

> The term 'scaling level' is replacing the misleading term 'measurement level' and is proposed by Van der Lans (1992) as more appropriate.

Usually the scaling level can be chosen for each variable separately, which gives the researcher sometimes a small classification problem prior to the analysis. Continuous nonlinear transformations can for instance be realized by the appliance of fuzzy coding (see also Van Rijckevorsel, 1987 or Ramsay, 1988). We will show in this section that the extension of reflected variance methods with nonlinear transformations opens new fields of application, hitherto not easy to explore with nonlinear MVA techniques.

We discuss and illustrate the implementation of nonlinear transformations in reflected variance methods on the basis of the most elaborated method of this chapter: Reflected Discriminant Analysis. Nonlinear Reflected Discriminant Analysis (NRDA) will be compared with the nonlinear version of Discriminant Analysis proposed in Gifi (1990) with the acronym CRIMINALS. We give a slightly deviating definition of Nonlinear Discriminant Analysis (NDA) to link up with the subsequent definition of NRDA. In our notation NDA maximizes $\mathrm{CVA}_{\mathrm{BT}}$ (4.13) with a different specification for $\mathbb{P}_{\mathbb{B}}$.

NDA: $$\mathrm{Fit}(\mathbb{V}) = \mathrm{tr}\ \mathbb{V}'\mathbb{P}'\mathbb{P}_{\mathbb{B}}\mathbb{P}_{\mathbb{B}}'\mathbb{P}\mathbb{V}, \tag{4.35}$$

with $\mathbb{V}'\mathbb{P}'\mathbb{P}\mathbb{V}=\mathbb{V}'\mathbb{V}=\mathbb{I}$, for the unit orthonormalized discriminant space $\mathbb{P}\mathbb{V}=\mathbb{F}(\mathbb{H})\mathbb{A}$,

where $\mathbb{F}(\mathbb{H})$ denotes the nonlinear transformed values of $\mathbb{H}$,

$\mathbb{P}$ denotes the orthonormal basis of $\mathbb{F}(\mathbb{H})$, with $\mathbb{F}(\mathbb{H})=\mathbb{P}\Phi\mathbb{Q}'$,

and $\mathbb{P}_{\mathbb{B}}$ denotes the orthonormal basis of $\mathbb{G}\mathbb{F}(\mathbb{H})=\mathbb{F}(\mathbb{H})_{\mathbb{B}}$, see (4.8), with $\mathbb{F}(\mathbb{H})_{\mathbb{B}} = \mathbb{P}_{\mathbb{B}}\Phi_{\mathbb{B}}\mathbb{Q}_{\mathbb{B}}'$.

In other words the variables $\mathbb{H}$ are optimally transformed on their user specified scaling level in such a way that Discriminant Analysis of the transformed variables $\mathbb{F}(\mathbb{H})$ gives maximal discrimination with (4.13). The definition of (4.35) seems to allow only single transformations for the variables $\mathbb{H}=(\mathbb{h}_1,\ldots,\mathbb{h}_k,\ldots,\mathbb{h}_K)$, i.e. the transformations are equal for all p dimensions of $\mathbb{P}\mathbb{V}$. Nevertheless we can incorporate, for instance, a multiple nominal scaling level for variable k in the analysis by expanding the datamatrix $\mathbb{H}$ in the following way:

$$\mathbb{H} = (\mathbb{H}_1,\ldots,\mathbb{H}_k,\ldots,\mathbb{H}_K), \tag{4.36}$$

with $\mathbb{H}_k = \mathbb{h}_k$ for single variables,
and $\mathbb{H}_k = \mathbb{J}\mathbb{G}_k\mathbb{D}_k^{-1/2}$ for multiple nominal variables,

where $\mathbb{G}_k$ denotes an orthogonal category indicator matrix
$\mathbb{J} = \mathbb{I} - \mathbb{1}(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}'$ denotes a centring operator.

The orthonormal matrix $\mathbb{H}_k$ is in deviations from the mean by the centring operator $\mathbb{J}$. For the 'transformation' of the multiple nominal variables we define $\mathbb{F}(\mathbb{H}_k)=\mathbb{H}_k$. This non standard incorporation of multiple nominal variables in NDA (4.35) is necessary to make the step towards a comparable reflected variance method less complicated with respect to the definition of an orthonormal basis for $\mathbb{F}(\mathbb{H})$. With NDA we have only one transformed datamatrix $\mathbb{F}(\mathbb{H})$ for all dimensions, whereas the Gifi CRIMINALS definition would give different transformed datamatrices for each dimension in the case of multiple variables.

The nonlinear version of RDA (NRDA) maximizes RDA (4.21) with another specification of the between variables.

$$\text{NRDA:}\quad \text{Fit}(\mathbb{V}) = \text{tr } \mathbb{V}'\mathbb{P}'\mathbb{F}(\mathbb{H})_{\mathbb{B}}\mathbb{F}(\mathbb{H})_{\mathbb{B}}'\mathbb{P}\mathbb{V} \tag{4.37}$$

with $\mathbb{V}'\mathbb{V}=\mathbb{I}$, for the unit orthonormalized discriminant space $\mathbb{P}\mathbb{V} = \mathbb{F}(\mathbb{H})\mathbb{A}$,

where $\mathbb{F}(\mathbb{H})_{\mathbb{B}}$ denotes $\mathbb{G}\mathbb{F}(\mathbb{H})$, which are the between-variables of $\mathbb{F}(\mathbb{H})$.

In other words the variables $\mathbb{H}$ are optimally transformed on their user specified scaling level in such a way that Reflected Discriminant Analysis of the transformed variables $\mathbb{F}(\mathbb{H})$ gives maximal discrimination with the RDA fit function (4.21). The above mentioned incorporation of multiple nominal variables (4.36) simplifies the definition of the orthonormal basis $\mathbb{P}$ of the transformed variables $\mathbb{F}(\mathbb{H})$.

The merits of NRDA compared with NDA can be assessed in the theoretical framework of this chapter. In the complete rank case described in section 4.3.7 the DA solutions are not uniquely defined. This situation for instance occurs when the datamatrix $\mathbb{H}$ has many more columns than rows. With NDA the non unique solutions will occur even more frequently due to the increase in degrees of freedom. The category expansion for multiple nominal scaling level (4.36) illustrates the possibility of a drastic increase in rank. A nonlinear version of PC-DA for the complete rank case would be maximizing in the first step a non adequate fit function as outlined in section 4.3.6. NRDA on the contrary is able to handle the complete rank case without the above mentioned disadvantages. It can find optimal transformations of the variables for separating different groups in many practical situations where NDA breaks down. In chapter 6 we provide an algorithm for computing an optimal NRDA solution. In chapter 7 we give a real-life application of NRDA.

Summarising this chapter we developed a theoretical framework for reflected variance methods. We have shown that RDA can improve group prediction compared to DA and PC-DA. Nonlinear prediction in the Gifi (1990) framework with a relatively large number of variables is now possible with NRDA. The same properties are expected for other reflected variance methods like RCA and nonlinear RCA. This has to be explored in future research.

Chapter 5

DIRECTED CORRELATIONS AND PARTIAL LEAST SQUARES

In this chapter a new multiplicative hybrid method is formulated that maximizes the product of two complementary fit functions, a local and a global MVA function. The local function gives a multiset alternative for maximizing variance accounted for. The global function maximizes correlations as formulated in chapter 3. These adjusted correlations are called *directed correlations* and are embedded in a multiset path analysis framework utilizing *primary* and *secondary* predictions. The product function that globally maximizes directed correlations and locally increases set variance as much as possible is called Lifted Directed Correlations (LDC). LDC is able to describe many existing MVA methods, hybrid and adjusted methods. It gives one fit function for cyclic hybrid methods like the basic and extended Partial Least Squares (PLS) method of path modelling, Consensus PLS and PLS Hierarchical Components.

Introduction

Defining a product of two functions is also applied in a method called projection pursuit. By definition, projection pursuit searches an optimal projection by maximizing (or minimizing) a certain objective function or projection index. For an overview of projection pursuit see Huber, 1985. Friedman & Tukey (1974) describe a projection index, which is a product of a local and a global function. The discrimination of local versus global is formulated by defining local functions sensitive for local groups of objects versus a global function influenced by all objects in an equal way. In this monograph the concept of a global and local function is always conceived within the context of multiset analysis. A *global* fit function is maximized over all sets *interrelated* and the (sub-)solution for other sets can change if one of the sets is changed, while a *local* fit function gives in principle a maximum for *each set separately*, invariant under changes of other sets.

In section 5.1 two different formulations are given of multiset Local Reciprocal PCA (LRPCO and LRPCV) together with Global Reciprocal PCA (GRPCA). The properties of the global version give an impression of the properties of the local versions in a hybrid context. In GRPCA the solution can never be dominated by one set with a very high variance accounted for because each set is required to have some substantial contribution to the solution for each successive dimension. In section 5.2 we describe the global fit function of Directed Correlations (DC). Therefore we first give an introduction to the SUMCOR fit function and to the concept of a condensed or tangent variate.

Section 5.3 combines the local function of section 5.1 (LRPCV) and the global function of 5.2 (DC) in one product function called Lifted Directed Correlations (LDC). The properties of LRPCV incorporated in LDC are discussed. In order to extend the range of methods that can be described with LDC, we introduce the concept of primary prediction and regression variate. We show how LDC can be used to fit path models with primary and secondary predictions. A general algorithm is presented for LDC in section 5.3.6. With this algorithm we can show the relations of LDC with many other methods. These relations are established by comparing the algorithmic flow. Especially for PLS methods this is a necessary approach, because they are usually only defined by linear equations and not by an overall fit function.

In section 5.4.1 relations of LDC with many other methods are discussed. At the same time we give for all methods LDC path diagrams and show how to turn these path diagrams into specific LDC fit functions. In this way we offer a criterion for Wold's basic PLS method of Soft Modelling (Wold, 1982) and the extended basic PLS method proposed by Lohmöller (1989) as Latent Variable Path modelling. Despite the lack of a 'hard' scalar criterion the PLS system of path modelling has been used for many years, especially by chemometricians. It offers many statistical advantages compared with other path modelling systems like Lisrel (see Fornell & Bookstein, 1982). With the LDC fit function the 'soft' PLS system is made 'hard' and maybe this will give the method a greater impact. We think this will add a valuable instrument to the data-analysis tool-box for many researchers.

Apart from the general PLS system of path modelling we elaborate on some specific PLS methods in more detail. We give the LDC formulation of Consensus PLS proposed by Geladi & Martens (1988), the PLS Hierarchical Components method (Wold, 1982), the PLS1 regression method with a PLS1 continuum extension of Lorber, Wangen & Kowalski (1987) and the relation with Continuum Regression proposed by Stone & Brooks (1990). The reflected variance methods of chapter 4 can also be brought within the LDC framework and from the corresponding LDC path models arises an interesting two sets PLS method. Last but not least we show that Set Component Analysis of chapter 3 is an example of a DC path model. We conclude this chapter in section 5.5 with a theoretical comparison of LDC and DC.

5.1 Global and local formulation of reciprocal PCA

We define a global version of reciprocal PCA followed by a local version. In that manner this section provides us with a simple example of a global and local MVA fit function. At the same time the global version gives an impression of the properties of the local version in a hybrid context. Furthermore, we apply deflation as introduced in section 2.2.4 in order to guarantee a substantial contribution to the solution for *each* dimension.

A global formulation of reciprocal PCA for multiple sets is described by maximizing the following global reciprocal PCA fit function

$$\text{GRPCA: } \mathrm{Fit}(\mathbf{x}) = \frac{1}{\sum_{k=1}^{K} \frac{\mathbf{x}'\mathbf{x}}{\mathbf{x}'\mathbf{S}_k\mathbf{x}}}, \tag{5.1}$$

where $\mathbf{x}$ denotes the common latent variable.

Successive dimensions can be computed by deflating the matrices $\mathbf{S}_k$ according to

$$\mathbf{S}_{(k)s} = \mathbf{S}_k \qquad \text{for } s = 1, \qquad \forall k$$

$$\mathbf{S}_{(k)s} = (\mathbf{I}-\mathbf{x}_{s-1}\mathbf{x}_{s-1}')\mathbf{S}_{(k)s\text{-}1}(\mathbf{I}-\mathbf{x}_{s-1}\mathbf{x}_{s-1}') \qquad \text{for } s = 2,\ldots,p. \qquad \forall k$$

The emphasis of the GRPCA solution is not so much on maximizing the total variance accounted for, but on avoiding for each set k a very low variance accounted for. Each set k is required to have some substantial contribution to the solution.

A local formulation of reciprocal PCA for multiple sets is achieved by introducing in (5.1) for each set the local unit normalized linear combinations $\mathbf{z}_k = \mathbf{H}_k\mathbf{t}_k$ instead of $\mathbf{x}$. The local reciprocal PCA fit function should then be

$$\text{LRPCO:}\quad \text{Fit}(\mathbb{Z}) = \frac{1}{\sum_{k=1}^{K} \frac{\mathbf{z}_k'\mathbf{z}_k}{\mathbf{z}_k'\mathbf{S}_k\mathbf{z}_k}}\ , \tag{5.2}$$

where $\mathbb{Z} = (\mathbf{z}_1,\ldots,\mathbf{z}_k,\ldots,\mathbf{z}_K)$ denote the unit normalized set variates,
with $\mathbf{z}_k = \mathbf{H}_k\mathbf{t}_k$ and $\mathbf{z}_k'\mathbf{z}_k = 1$, $\forall k$
and $\mathbf{t}' = (\mathbf{t}_1',\ldots,\mathbf{t}_k',\ldots,\mathbf{t}_K')$ denotes a vector with $\sum_k m_k$ variable weights,

For reasons to be explained in section 5.3.4 we describe another slightly different version of local reciprocal PCA. In (5.2) we have in fact an object-wise LRPC<u>O</u> formulation when we look at the variance of $\mathbf{H}_k$ accounted for by $\mathbf{z}_k$. In the following variable-wise LRPC<u>V</u> formulation we look at the variance of $\mathbf{H}_k$ accounted for by $\mathbf{t}_k$.

$$\text{LRPCV:}\quad \text{Fit}(\mathbf{t}) = \frac{1}{\sum_{k=1}^{K} \frac{\mathbf{t}_k'\mathbf{t}_k}{\mathbf{t}_k'\mathbf{H}_k'\mathbf{H}_k\mathbf{t}_k}} = \frac{1}{\mathbf{t}'\mathbf{t}}, \tag{5.3}$$

with the same notation and normalization as for (5.2). Therefore $\mathbf{z}_k'\mathbf{H}_k'\mathbf{H}_k\mathbf{z}_k = 1$, $\forall k$.

Successive dimensions in (5.2) and (5.3) can be computed by locally deflating the matrices $\mathbf{H}_k$ according to

$$\mathbf{H}_{(k)s} = \mathbf{H}_k \qquad \text{for } s = 1,$$

$$\mathbf{H}_{(k)s} = (\mathbf{I} - \mathbf{z}_{(k)s-1}\mathbf{z}_{(k)s-1}')\mathbf{H}_{(k)s-1} \qquad \text{for } s = 2,\ldots,p. \qquad \forall k \tag{5.4}$$

The *deflation* is local, because $\mathbf{z}_{(k)s-1}$ is specific for each set separately, while $\mathbf{x}_{s-1}$ in the deflation for (5.1) is the same for all sets. Due to (5.4) we will always find orthogonal variates in successive dimensions.

The *solution* of LRPCO and LRPCV is also local for $\mathbf{t}$, because $\mathbf{t}$ defines different variates $\mathbf{z}_k$ for each set, whereas the solution of this variates is <u>not</u> related to the content of the other sets. Analogous to the properties of GRPCA we can expect that the LRPCO and LRPCV fit functions, put in a global context in a hybrid method, introduce for the hybrid solution the tendency never to be dominated by one set with a very high variance accounted for and to have some substantial contribution of each set to the solution for each dimension. Maximization of (5.2) and (5.3) with (5.4) makes

$\mathbb{z}_{(k)s}$ equal to the eigenvectors of $\mathbb{S}_k$ with the largest p eigenvalues in descending order. By combining the local MVA functions with a global fit function we can obtain less trivial solutions. In section 5.2 we discuss a very good candidate for making such a hybrid method.

Finally, regarding the difference between global and local MVA functions, it is important to notice that having in principle different variates for each set is a necessary, but not sufficient, condition for a MVA function to be local. Crucial for a MVA function to be local is the property of having in principle independent solutions for each set. If there is some slight interrelation we are already dealing with a global MVA function.

5.2 The adjusted method of Directed Correlations

The global fit function to be combined in this chapter with LRPCV is the Directed Correlations (DC) fit function. The adjusted method of Directed Correlations is build up analogous to the SCA method in chapter 3. We integrate set correlation and set variance. Therefore we maximize a weighted sum of set correlations between pairs of adjusted set variates. The sum of set correlations is the SUMCOR fit function (Horst, 1961) and the set variates with improved variance accounted for are called *condensed variates*. In section 5.2.1 and 5.2.2 we introduce respectively the SUMCOR fit function and condensed variates. In section 5.2.3 we construct the Directed Correlations fit function by joining the preceding two sections.

5.2.1 The SUMCOR fit function

The SUMCOR fit function as described by Horst (1961) maximizes the sum of the correlations between all possible combinations of set variates

$$\text{SUMCOR: Fit}(\mathbb{Z}) = \sum_{k=1}^{K} \sum_{l=1}^{K} \mathbb{z}_k{}'\mathbb{z}_l = \mathbb{1}'\mathbb{R}\,\mathbb{1}, \qquad (5.5)$$

where $\mathbb{Z} = (\mathbb{z}_1,\ldots,\mathbb{z}_k,\ldots,\mathbb{z}_K)$ denote the unit normalized set variates,
with $\mathbb{z}_k = \mathbb{H}_k\mathbb{t}_k$, $\forall k$

	$\mathbb{1}$	denotes a column vector of appropriate size with elements 1,
and	$\mathbb{R} = \mathbb{Z}'\mathbb{Z}$	denotes a $(K{\times}K)$ symmetric correlation matrix.

Local deflation is according to (5.4). SUMCOR($\mathbb{t}$) would be an alternative formulation for indicating the unknown parameters of (5.5). Although we have different variates for each set and local deflation, SUMCOR is a global fit function, because (5.5) is maximized over all sets *interrelated* and the solutions for variates of other sets can change if one of the sets is changed. Before constructing the Directed Correlations fit function by multiplying the elements of $\mathbb{R}$ in (5.5) with weights we need to explain the concepts of condensed variates and secondary prediction.

5.2.2 Condensed variate for secondary prediction

In principle the condensed variate is equal to the improved set variate defined by (3.4). Without the indices for the dimensions we obtain $\mathbb{z}_k=\mathbb{S}_k\mathrm{x}(\mathrm{x}'\mathbb{S}_k\mathbb{S}_k\mathrm{x})^{-1/2}$. Applied to the correlations in (5.5) we have three *adjusted* correlations,

$$\mathbb{z}_l'\mathbb{z}_k=\mathbb{z}_l'\mathbb{S}_k\mathbb{z}_l(\mathbb{z}_l'\mathbb{S}_k\mathbb{S}_k\mathbb{z}_l)^{-1/2},$$

$$\mathbb{z}_k'\mathbb{z}_l=\mathbb{z}_k'\mathbb{S}_l\mathbb{z}_k(\mathbb{z}_k'\mathbb{S}_l\mathbb{S}_l\mathbb{z}_k)^{-1/2},$$

or

$$(\mathbb{z}_k'\mathbb{S}_l\mathbb{S}_l\mathbb{z}_k)^{-1/2}\mathbb{z}_k'\mathbb{S}_l\mathbb{S}_k\mathbb{z}_l(\mathbb{z}_l'\mathbb{S}_k\mathbb{S}_k\mathbb{z}_l)^{-1/2}. \tag{5.6}$$

The adjusted correlations are called *directed correlations*. The adjusting set variate we call the *pivot variate* and the adjusted variate we call the *condensed variate*. The third directed correlation in (5.6) is in fact adjusting in two directions. Therefore the set variates are in this case both pivot and condensed variates. In that case we refer to the set variates as condensed variates. The pivot and condensed variates are linear combinations of respectively the *pivot set* of variables and the *condensed set* of variables. The condensed variate $\mathbb{z}_k$ and matrices $\mathbb{R}$, which contain directed correlations, will always be indicated in outline. The properties of the condensed variate are now elaborated.

A condensed variate is a linear combination of a set of variables that can condense the set variance information in such a way that it can replace the whole set with respect to some pivot variable. The spatial position of the condensed variate is such that the variance of all the variables of the condensed set accounted for by some pivot variate is exactly the same as the variance of the condensed variate accounted for by the same pivot variable. For this equality to be valid we require that the variable weights of the condensed variate are unit normalized. The process of predicting variables through an condensed variate we call *secondary prediction.*

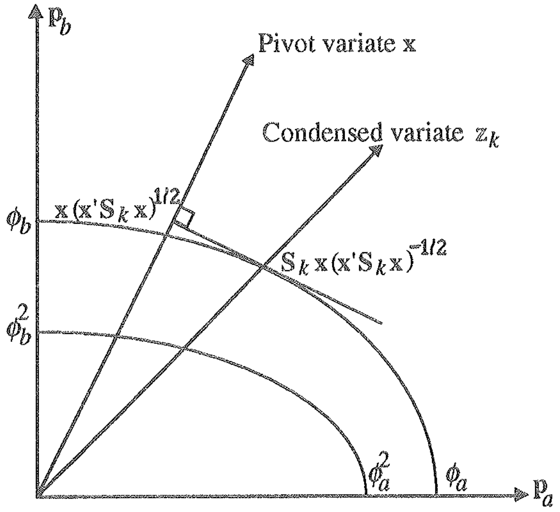


Figure 5.1 *Pivot variate* x *and condensed variate* z_k.

In figure 5.1 we give a geometric example of an pivot variate x and a condensed variate z_k for a set H_k with rank two. In this example the pivot variate x is located in the plane of the singular vectors p_a and p_b of $H_k=(p_a,p_b)\Phi_k Q_k'$ with singular values $\phi_a > \phi_b$. The construction of the condensed variate is made on top of figure 3.1 of chapter 3, where an ellipse through the eigenvalues ϕ_a^2 and ϕ_b^2 was drawn. To facilitate the introduction of condensed and pivot variables at a later stage we extend matrices and vectors with subscripts *k*. We call a variate $H_k t_k$ with unit normalized variable weights, $t_k' t_k=1$ $\forall k$, a *unit weights variate*. All possible unit weights variates form a hyperellipse and in our simple rank two example this is an ordinary ellipse through the singular values ϕ_a and ϕ_b. If we take some fixed pivot variate x in any direction then the variance of H_k accounted for by x is given by the squared length of the largest projection of the hyperellipse on to this pivot variate x. There is exactly one point where the hyperplane orthogonal to the pivot variate x touches the hyperellipse.

The line from the origin through this tangent point determines the unit normalized condensed variate z_k. In other words, geometrically the condensed variate can also be called the tangent variate. We now show the relation between the variance of the unit weights condensed variate accounted for by the pivot variate x and the variance of H_k accounted for by x.

> We emphasize that the variance accounted for is defined by the squared sum of the projections of the variables or variates on to x and not by taking the squared length of some vector on the outer ellipse in figure 5.1.

Definition 5.1. The variance of condensed (or tangent) variate $H_k t_k$ with unit normalized variable weights accounted for by the pivot variate x and the variance of H_k accounted for by x are exactly the same.

Existence and uniqueness. First we find the largest projection of the hyperellipse described by the unit weights variates $H_k t_k$ on to the pivot variate x. This implies that we have to maximize $x'H_k t_k$ for fixed pivot variate x with $x'x=1$ and free parameters t_k with restriction $t_k't_k=1$ $\forall k$. By applying the Cauchy-Schwarz inequality on the non fixed parameters of $x'H_k t_k$ we know that $(x'H_k t_k)^2 \leq (x'S_k x)(t_k't_k)=(x'S_k x)$. The maximum of $x'H_k t_k$ is reached if $(x'H_k t_k)^2=(x'S_k x)$ and therefore the optimal value for $t_k=H_k'x(x'S_k x)^{-1/2}$. The unit weights condensed (tangent) variate is known by substituting the optimal value for t_k in the unit weights variate $H_k t_k$ and we obtain $S_k x(x'S_k x)^{-1/2}$. In figure 5.1 we see that the unit weights condensed variate $S_k x(x'S_k x)^{-1/2}$ is really a tangent variate. The variance of this variate accounted for by the pivot variate x is $(x'S_k x(x'S_k x)^{-1/2})^2=x'S_k x$, which is equal to the variance of H_k accounted for by x. The direction of the condensed variate is uniquely defined by the projection of x on to H_k, which is $P_k P_k'x$, and the set variance structure of H_k. □

For fixed pivot variate x the condensed variate with unit weights normalization is a good candidate for replacing all variables of set *k*. The unit normalized condensed variate $z_k=S_k x(x'S_k S_k x)^{-1/2}$ with $z_k'z_k=1$ $\forall k$, (see figure 5.1), has also a unique direction related to the projection of the pivot variate x and the set variance structure of H_k. The condensing property of the condensed variate is invariant of the sign of vector $S_k x(x'S_k S_k x)^{-1/2}$ and therefore $z_k=\pm S_k x(x'S_k S_k x)^{-1/2}$ is to be preferred in this respect. In the next section we introduce the condensed variates $\mathbf{z}_k$ in the SUMCOR fit function in order to obtain Directed Correlations.

5.2.3 Directed Correlations

By combining the theory of section 5.2.1 and 5.2.2 we change the ordinary Pearson correlations in the SUMCOR fit function into 'directed' correlations. A directed correlation is a Pearson correlation between pivot and condensed variates as specified in section 5.2.2. The correlations are called 'directed', because the condensed variate is an intermediate variate of some set of variables in such a way that the pivot variate can predict the variables by predicting a properly normalized condensed variate. We already referred to this as secondary prediction. There is a prediction path possible from pivot variate passing through the condensed variate and ending with the set of variables. In section 5.3.3 on path diagrams we will give more details.

The dual nature of directed correlations implies that each variate can have two roles. The same variate can be an pivot variate and it can be a condensed variate.

Or a regression variate, which is a condensed variate of an orthonormal set of variables, see section 5.3.2.

We refer to this phenomenon by saying that a variate can be in pivot or in condensed (c.q. regression) mode. Another aspect of the dual nature of directed correlations is that a variate can be the pivot variate of many different condensed variates, but that a variate can be in principle only the condensed variate of several pivot variates if the projections of the pivot variates on the condensed set have exactly the same direction. The last statement is true because the condensed variate of a certain set of condensed variables is uniquely defined by the by the direction of these projected pivot variates. In practice this will usually not occur. If one wishes to have a condensed variate based on many sets, this can be achieved by taking a linear combination of several set variates and using this linear combination as pivot variate. As a result the condensed variate is indirectly determined by several (sub-) pivot variates. With this slightly extended concept of pivot variates we can now construct the one dimensional Directed Correlations fit function by adapting (5.5) as follows

$$\text{DC:} \qquad \text{Fit}(\mathbb{Z},\mathbb{W}) = \sum_{k=1}^{K} \sum_{l=1}^{K} w_{kl} \mathbf{z}_k{}' \mathbf{z}_l = \mathbb{1}'(\mathbb{R} * \mathbb{W})\mathbb{1} \qquad (5.7)$$

where $\mathbb{R} = \mathbf{Z}'\mathbf{Z}$ denotes a $(K{\times}K)$ symmetric correlation matrix with directed correlations,

$\mathbf{Z} = (\mathbf{z}_1,\ldots,\mathbf{z}_k,\ldots,\mathbf{z}_K)$ denote unit normalized pivot variates, $\mathbf{z}_k = \mathbb{H}_k\mathfrak{t}_k$, with a subset $\forall k$

$\mathbb{Z} = (\mathbb{z}_1,\ldots,\mathbb{z}_k,\ldots,\mathbb{z}_K)$ denoting condensed variates,

$\mathbb{z}_k = \pm\mathbb{S}_k\mathbf{z}_{k*}((\mathbb{S}_k\mathbf{z}_{k*})'\mathbb{S}_k\mathbf{z}_{k*})^{-1/2}$, $\forall k\in J_k$,

$\mathbf{z}_{k*} = \sum_{l=1}^{K} w_{kl}\mathbf{z}_l$, and J_k the index set of $\mathbb{Z}$ with condensed variates,

where $\mathbb{W}$ denotes a matrix with weights or function values w_{kl} and with diagonal elements equal to zero, $\mathbb{W}_{\text{Diag}}$=0.

$\mathbb{W}_{\text{Design}}$ denotes a binary matrix with the design pattern of non-zero (1) and zero (0) weights of $\mathbb{W}$,

$*$ denotes the Hadamard (elementwise) product.

If appropriate the sum of squares of the weights $\mathbb{W}$ have to be normalized to a constant value, tr $\mathbb{W}'\mathbb{W}{=}c$. Successive dimensions can be computed by locally deflating the matrices $\mathbb{H}_k$ according to (5.4) or by user specification. We have to emphasize that $\mathbf{z}_k$ and $\mathbb{z}_k$ are just the same variates in different modes. The variate $\mathbf{z}_k$ is in pivot mode and $\mathbb{z}_k$ is in condensed mode. Therefore DC($\mathfrak{t}$,$\mathbb{W}$) would be a more efficient, but less clear formulation for indicating the unknown parameters of (5.7). The essential mathematical difference between $\mathbf{z}_k$ and $\mathbb{z}_k$ is found in the weighted determination of the condensed variate $\mathbb{z}_k$ by the pivot variates $\mathbf{z}_1,\ldots,\mathbf{z}_l,\ldots,\mathbf{z}_K$

$$\mathbb{z}_k = \pm\mathbb{S}_k\mathbf{z}_{k*}((\mathbb{S}_k\mathbf{z}_{k*})'\mathbb{S}_k\mathbf{z}_{k*})^{-1/2}, \qquad \forall k\in J_k,$$

$$\text{with} \quad \mathbf{z}_{k*} = \sum_{l=1}^{K} w_{kl}\mathbf{z}_l, \qquad \text{and } J_k \text{ the index set of condensed variates,} \tag{5.8}$$

If $w_{kl}{\neq}w_{lk}$ for one combination of (k,l), then reversion of the role of $\mathbf{z}_k$ and $\mathbb{z}_k$ can already give a different result for the determination of the condensed variate in (5.8). Restriction ($\forall k\in J_k$, the index set of condensed variates) implies that the condensed variate k is only well defined if there is at least one non-zero element in row k of the weight matrix $\mathbb{W}$. Using the more efficient parameter formulation with DC($\mathfrak{t}$,$\mathbb{W}$) as suggested above and the same notation, DC in (5.7) can be written as

DC: $\mathrm{Fit}(\mathbf{t},\mathbf{W}) = \mathbf{1}'(\mathbb{R}*\mathbf{W})\mathbf{1}$, (5.9)

where $\mathbf{t}' = (\mathbf{t}_1',\ldots,\mathbf{t}_k',\ldots,\mathbf{t}_K')$ denotes a vector with $\sum_k m_k$ variable weights.

By now we have formulated the Directed Correlations fit function. The function is equal to a weighted sum of directed correlations. DC is a local fit function if we require the matrix W to be diagonal. We always choose the matrix W with $\mathrm{W}_{\mathrm{Diag}}$=0, otherwise than diagonal and therefore can conceive and apply DC as a global fit function.

5.3 Lifted Directed Correlations

After all the preparations in the preceding sections we can now elaborate on a product function that can describe a wide variety of methods. The function is Lifted Directed Correlations and it is constructed in section 5.3.1 by globally maximizing the directed correlations of (5.9) and locally raising or lifting the variance as much as possible with LRPCV (5.3). The properties of LRPCV incorporated in LDC are discussed in section 5.3.4.

In section 5.3.2 we introduce the regression mode as a third mode next to condensed and pivot mode, because it extends the range of methods that can be described with LDC (5.10). We also introduce the concept of primary prediction. In section 5.3.3 we show how LDC can be used to fit path models with primary and secondary predictions and many weighting modes for the variates. We describe these weighting modes consisting of different types of weights (like proportional function weights) and different weighting functions.

Although we discuss all algorithms of this monograph in chapter 6, we make an exception for LDC in section 5.3.6. The reason for this special treatment is that we want to discuss in section 5.4 the relations of LDC with other methods. These relations are given by comparing the algorithmic flow. Especially for most PLS methods this is a necessary approach, because they are usually only defined by linear equations and not by an overall fit function. In order to derive linear equations and an algorithm for LDC we first reformulate LDC in section 5.3.5.

5.3.1 Combining global DC with local LRPCV

The Lifted Directed Correlations fit function is given by the product of DC (5.9) and LRPCV (5.3)

$$\text{LDC:} \qquad \text{Fit}(\mathbb{t},\mathbb{W}) = \frac{\mathbb{1}'(\mathbb{R}*\mathbb{W})\mathbb{1}}{\mathbb{t}'\mathbb{t}K^{-1}}, \tag{5.10}$$

where $\mathbb{R} = \mathbb{Z}'\mathbb{Z}$ denotes a $(K{\times}K)$ symmetric correlation matrix with directed correlations,

$\mathbb{Z} = (\mathbb{z}_1,\ldots,\mathbb{z}_k,\ldots,\mathbb{z}_K)$ denote unit normalized pivot variates, $\mathbb{z}_k = \mathbb{H}\,{}^{\alpha}_{k}\mathbb{t}_k$, with a subset $\forall k$

$\mathbb{Z} = (\mathbb{z}_1,\ldots,\mathbb{z}_k,\ldots,\mathbb{z}_K)$ denoting condensed variates,

$\mathbb{z}_k = \pm\mathbb{S}\,{}^{\alpha}_{k}\mathbb{z}_{k*}((\mathbb{S}\,{}^{\alpha}_{k}\mathbb{z}_{k*})'\mathbb{S}\,{}^{\alpha}_{k}\mathbb{z}_{k*})^{-1/2}$, $\forall k\in J_k$,

$\mathbb{z}_{k*} = \sum_{l=1}^{K} w_{kl}\mathbb{z}_l$, and J_k the index set of $\mathbb{Z}$ with condensed variates,

where $\mathbb{W}$ denotes a matrix with weights or function values w_{kl} and with diagonal elements equal to zero, $\mathbb{W}_{\text{Diag}}=0$,

$\mathbb{W}_{\text{Design}}$ denotes the binary design pattern of $\mathbb{W}$.

where $\mathbb{t}' = (\mathbb{t}_1',\ldots,\mathbb{t}_k',\ldots,\mathbb{t}_K')$ denotes a vector with $\sum_k m_k$ variable weights,

$\mathbb{H}^{\alpha}_{k} = \mathbb{P}_k\Phi^{\alpha}_{k}\mathbb{Q}_k'$, where α in this context is short for α_k $\forall k$

If appropriate the sum of squares of the weights $\mathbb{W}$ have to be normalized to a constant value, tr $\mathbb{W}'\mathbb{W}=c$. Local deflation is according to (5.4) or user specified. Constant K adjusts the normalization of LDC to the normalization of DC, as will be explained in section 5.3.2.

In fact we introduced the superscript α_k for the condensed and pivot mode variables $\mathbb{H}^{\alpha}_{k} = \mathbb{P}_k\Phi^{\alpha}_{k}\mathbb{Q}_k'$, with $\mathbb{H}_k = \mathbb{P}_k\Phi_k\mathbb{Q}_k'$ as usual. For convenience and without loss of generality we defined α_k also for set k with only pivot mode variables. For convenience, because it offers an uniform treatment of pivot and condensed mode variables. Without loss of generality, because the pivot variates are in principle only restricted to be in the space of the corresponding variables and therefore invariant under nonsingular transformations within sets. In section 5.3.2 we will explain how α_k gives the possibility of introducing a special mode, namely the regression mode.

5.3.2 The regression mode for primary prediction

The regression mode is introduced next to the condensed and pivot mode, because it extends the range of methods that can be described with LDC (5.10). We also introduce the concept of primary prediction. For α_k=1 we obtain the condensed variates $\mathbf{z}_k=\pm\mathbf{S}_k\mathbf{x}(\mathbf{x}'\mathbf{S}_k\mathbf{S}_k\mathbf{x})^{-1/2}$, as described in section 5.2.2 and the condensed variates in (5.8) specified for multiple pivot variables. For α_k=0 we have the regression mode and regression variates. In that case $\mathbf{H}^0_k=\mathbf{P}_k\mathbf{Q}_k'$ and $\mathbf{S}^0_k=\mathbf{P}_k\mathbf{P}_k'$. So the matrix $\mathbf{H}_k$ is replaced by an orthonormal basis $\mathbf{H}^0_k$ and $\mathbf{S}_k$ is replaced by the projector $\mathbf{P}_k\mathbf{Q}_k'\mathbf{Q}_k\mathbf{P}_k'=\mathbf{P}_k\mathbf{P}_k'$. Any other orthonormal basis would also be fine, but $\mathbf{H}^0_k$ simplifies the notation compared to $\mathbf{H}^1_k$. Geometrically this means that in figure 5.1 the ellipse is replaced by a circle and in the general case the hyperellipse is replaced by a hypersphere. The tangent variate is now found by simply projecting $\mathbf{x}$ on to the space of $\mathbf{H}_k$. The multiple regression weights of $\mathbf{P}_k$ for predicting $\mathbf{x}$ are $(\mathbf{P}_k'\mathbf{P}_k)^{-1}\mathbf{P}_k'\mathbf{x}=\mathbf{P}_k'\mathbf{x}$. The resulting regression variates are $\mathbf{z}_k=\pm\mathbf{P}_k\mathbf{P}_k'\mathbf{x}(\mathbf{x}'\mathbf{P}_k\mathbf{P}_k'\mathbf{x})^{-1/2}$ and for multiple pivot variates

$$\mathbf{z}_k = \pm\mathbf{S}^0_k\mathbf{z}_{k*}((\mathbf{S}^0_k\mathbf{z}_{k*})'\mathbf{S}^0_k\mathbf{z}_{k*})^{-1/2}, \qquad \forall k\in J_k,$$

with $\mathbf{z}_{k*} = \sum_{l=1}^{K} w_{kl}\mathbf{z}_l$, and J_k the index set of condensed variates, (5.11)

In PLS literature (see section 5.4.1) the condensed mode (α_k=1) is called ModeA, outwards directed, or factor mode and the regression mode (α_k=0) is called ModeB, inwards directed, or regression mode. In section 5.2.2 we called the process of predicting variables through an intermediate variate *secondary prediction*. The condensed mode in LDC (5.10) results in secondary prediction, because the variance of the sets accounted for by the pivot variates is lifted locally as much as possible by LRPCV. This is shown in the section 5.3.4. The prediction is directed from the pivot variates towards the variables of the condensed set. The regression mode in LDC (5.11) results in what we call *primary prediction*. In this mode the prediction direction is reversed. The regression variates predict as well as possible the (weighted sums of the) pivot variates. If we fit a path model with only primary predictions for all variates, then $\mathbf{t}'\mathbf{t}=K$ and consequently we obtain the same results for DC and LDC.

5.3.3 Path models and weighting mode

Recursive and non-recursive path models can be fitted with Lifted Directed Correlations. In section 5.4 we will give many examples. In this section we give some general principles and possibilities. For fitting a path model with LDC we need a path diagram which specifies the (hypothetical) design and mode of the relations between the set variables and latent variates, and the design and mode of the weighted sum of pivot variates.

For each *relation between variates and their corresponding set variables* we must first specify which variables are linked to which variates (PLS outer design matrix). Secondly the kind of prediction has to be chosen by specifying the mode of the variates. For primary prediction we have the regression mode with α_k=0 (PLS mode A) and for secondary prediction the condensed mode with α_k=1 (PLS mode B). A condensed variate in LDC predicts the set variables and is at the same time being predicted by an pivot variate.

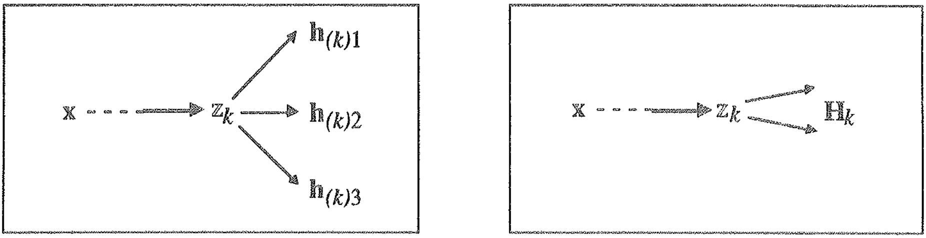


Figure 5.2 *LDC path diagram for pivot variate* $\mathbf{x}$ *and condensed variate* $\mathbf{z}_k$.

Figure 5.2 gives a prototype of the arrow configuration around a condensed variate $\mathbf{z}_k$ for set $\mathbf{H}_k$ with three variables $\mathbf{h}_{(k)1}$, $\mathbf{h}_{(k)2}$ and $\mathbf{h}_{(k)1}$. The right hand diagram gives a more abstract contracted illustration of the left hand diagram.

A regression variate in LDC predicts the pivot variate and is a linear sum of the variables of the corresponding set.

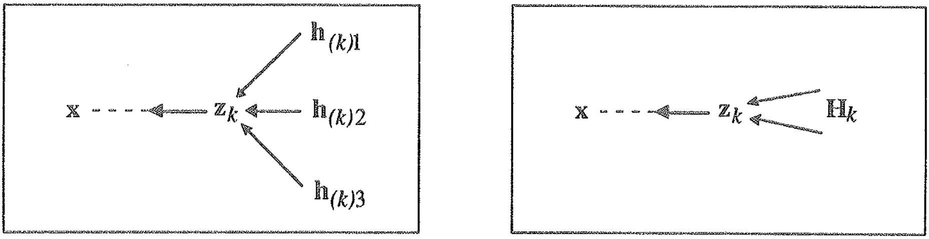


Figure 5.3 *LDC path diagram for pivot variate* $\mathbf{x}$ *and regression variate* $\mathbf{z}_k$.

Figure 5.3 gives a prototype of the arrows around a regression variate $\mathbf{z}_k$ with a contracted illustration on the analogy of figure 5.2. From this point on we will use only diagrams with contracted illustrations.

The weight matrix $\mathbf{W}$ is a combination of a weight matrix $\mathbf{W}_{\text{Design}}$ and a weight mode. The design for *the weighted sums of the pivot variates* is summarised in the rows of matrix $\mathbf{W}_{\text{Design}}$. Figure 5.4 gives an example of the LDC arrow configuration around a condensed variate with two pivot variates $\mathbf{x}_1$ and $\mathbf{x}_2$.

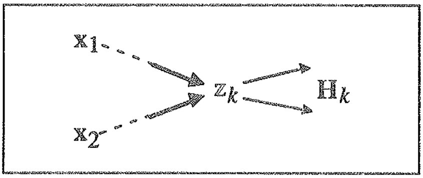


Figure 5.4 *Condensed variate* $\mathbf{z}_k$ *with pivot variates* $\mathbf{x}_1$ *and* $\mathbf{x}_2$.

The binary matrix $\mathbf{W}_{\text{Design}}$ is more or less equivalent with the command design matrix in PLS and it specifies in principle the pattern of adjacent latent variables in the path model with non-zero (1) and zero (0) weights. As for the kind of prediction between the variates special modes are defined for the weighted sums of the pivot variates. The mode of the weights is defined by the type of weight and its function. We discern three type of weights: the fixed weights, the proportional function weights and the function weights. The *fixed weights* are parameters that always remain constant. The *proportional function weights* are proportional to the values of a multivariate weight function. The optimal proportional function weights remain proportional to their corresponding weight function value if the fit function is maximized with these weights fixed. The optimal solution can be found by normalizing the sum of squares of the weights, (tr W'W) to some constant value. The *function weights* are equal to their weight function value.

Several functions can be chosen for the proportional function weights and function weights. The combinations of weight function and type of weight result in many different weighting modes. All PLS methods we have studied used proportional function weights. In the Basic PLS method described by Wold only one weight function and therefore one weighting mode is used for W. In the Extended PLS method Lohmöller adds two other weight functions and discerns three 'inner weighting modes'. They will be discussed in section 5.4.1.

Usually the weight design matrix $\mathbf{W}_{\text{Design}}$ is the same for all p successive dimensions, but in principle a different design can be chosen for the respective dimensions. Finally we mention the possibility of defining an ancillary set of latent variates as if they are manifest variables. For this ancillary set a condensed or regression variate can be established.

5.3.4 Properties of LRPCV incorporated in LDC

In section 5.1 we presented an object-wise LRPCO and a variable-wise LRPCV formulation of local reciprocal PCA. The properties of the local function LRPCV are influenced by the symbiosis with global DC in LDC. We will now examine the properties of *incorporated* LRPCV. In LRPCO we looked at the variance of $\mathbf{H}_k$ accounted for by unit normalized variates $\mathbf{z}_k$. In LRPCV we looked at the variance of $\mathbf{H}_k$ accounted for by variable weights $\mathbf{t}_k$. In incorporated LRPCV we add to this last property that we look at the variance of $\mathbf{H}_k$ accounted for by variates somewhere between the condensed variates $\mathbf{z}_k$ and the weighted sums $\mathbf{z}_{k*}$ (5.8) of the pivot variates $\mathbf{z}_1,\ldots,\mathbf{z}_l,\ldots,\mathbf{z}_K$. By maximizing locally the variance of $\mathbf{H}_k$ accounted for by variable weights $\mathbf{t}_k$ we also enlarge the variance accounted for by the weighted sum of the pivot variates. Formulated mathematically we state that

$$\text{Fit}_{\text{LRPCO}}(\mathbf{z}_{k*}) \leq \text{Fit}_{\text{LRPCV}}(\mathbf{z}_{k*}) \leq \text{Fit}_{\text{LRPCO}}(\mathbf{z}_k), \qquad (5.13)$$

where $\text{Fit}_{\text{LRPCO}}(\mathbf{z}_{k*})$ stands for

$$\text{LRPCO:}\ \ \mathrm{Fit}(\mathbf{z}_{k*}) = \frac{1}{\sum_{k=1}^{K} \frac{\mathbf{z}_{k*}'\mathbf{z}_{k*}}{\mathbf{z}_{k*}'\mathbf{S}_k\mathbf{z}_{k*}}}, \tag{5.12}$$

with notation as usual. In fact we already make (5.2) global by replacing the local variates $\mathbf{z}_k$ by the weighted sums $\mathbf{z}_{k*}$. The definition of LRPCO($\mathbf{z}_k$) is similar. Equation (5.13) needs some further exploration.

In our investigation of the properties of incorporated LRPCV we assume that all variates are of dual nature. This means that all variates are in pivot mode *and* in condensed (or regression) mode. This assumption applies for all methods discussed in this monograph. The restriction $\mathbf{z}_k=\pm\mathbf{S}_k\mathbf{z}_{k*}((\mathbf{S}_k\mathbf{z}_{k*})'\mathbf{S}_k\mathbf{z}_{k*})^{-1/2}$ in (5.10) implies for the variable weights $\mathbf{t}_{k*}=\pm\mathbf{H}_k'\mathbf{z}_{k*}((\mathbf{S}_k\mathbf{z}_{k*})'\mathbf{S}_k\mathbf{z}_{k*})^{-1/2}$. Substitution of these weights in LRPCV (5.3) gives an impression of incorporated LRPCV

$$\text{LRPCV:}\ \ \mathrm{Fit}(\mathbf{z}_{k*}) = \frac{1}{\sum_{k=1}^{K} \frac{\mathbf{z}_{k*}'\mathbf{S}_k\mathbf{z}_{k*}}{\mathbf{z}_{k*}'\mathbf{S}_k\mathbf{S}_k\mathbf{z}_{k*}}}, \tag{5.14}$$

with notation as usual. By applying the Cauchy-Schwarz inequality on $\mathbf{z}_{k*}'\mathbf{S}_k\mathbf{z}_{k*}$ we know that $(\mathbf{z}_{k*}'\mathbf{S}_k\mathbf{z}_{k*})^2\leq(\mathbf{z}_{k*}'\mathbf{S}_k\mathbf{S}_k\mathbf{z}_{k*})(\mathbf{z}_{k*}'\mathbf{z}_{k*})$ and therefore

$$\frac{\mathbf{z}_{k*}'\mathbf{S}_k\mathbf{z}_{k*}}{\mathbf{z}_{k*}'\mathbf{S}_k\mathbf{S}_k\mathbf{z}_{k*}} \leq \frac{\mathbf{z}_{k*}'\mathbf{z}_{k*}}{\mathbf{z}_{k*}'\mathbf{S}_k\mathbf{z}_{k*}}. \tag{5.15}$$

Combining inequality (5.15) with (5.12) and (5.14) we conclude that

$$\mathrm{Fit}_{\mathrm{LRPCO}}(\mathbf{z}_{k*}) \leq \mathrm{Fit}_{\mathrm{LRPCV}}(\mathbf{z}_{k*}). \tag{5.16}$$

A parallel procedure for LRPCO (5.2), with insertion of $\mathbf{H}_k\mathbf{t}_k$ for $\mathbf{z}_k$, and LRPCV (5.3) by applying the Cauchy-Schwarz inequality on $\mathbf{t}_{k*}'\mathbf{H}_k'\mathbf{H}_k\mathbf{t}_{k*}$ leads to the inequality

$$\mathrm{Fit}_{\mathrm{LRPCV}}(\mathbf{z}_{k*}) \leq \mathrm{Fit}_{\mathrm{LRPCO}}(\mathbf{z}_{k}). \tag{5.17}$$

Joining (5.16) with (5.17) we obtain the surplus property of incorporated LRPCV described in (5.13).

5.3.5 LDC revised

We confine ourselves to find a fitting procedure for LDC where all pivot variates are also condensed variates. The index set of condensed variates $J_k = 1,\ldots,k,\ldots,K$. Almost all methods discussed before in this monograph and all basic and many extended PLS methods can be fitted with the derived algorithm. The one dimensional LDC fit function (5.10) is first reformulated for technical reasons into

$$\text{LDC:} \qquad \mathrm{Fit}(\mathbf{t},\mathbb{W},\mathbb{D}_v) = \frac{\mathbf{t}'\mathbb{B}\,\mathbf{t}}{\mathbf{t}'\mathbf{t}} = \psi, \tag{5.18}$$

with $\mathbb{B}\mathbf{t} = \psi\mathbb{D}_v\mathbf{t}$,

$\mathbf{t}'\mathbb{D}_v\mathbf{t} = \mathbf{t}'\mathbf{t}$

$\mathbf{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbf{t}_k = 1,$ $\qquad \forall k$

where $\mathbf{t}' = (\mathbf{t}_1',\ldots,\mathbf{t}_k',\ldots,\mathbf{t}_K')$ denotes the partitioned variable weights of $\mathbb{B}$,

$\mathbb{B} = K(\underset{\cdots}{\mathbb{H}}'\underset{\cdots}{\mathbb{H}}) * \mathbb{W}^{\mathrm{ext}}$ denotes a weighted variance-covariance matrix,

$\underset{\cdots}{\mathbb{H}} = (\mathbb{H}\,{}^{\alpha}_{1},\ldots,\mathbb{H}{}^{\alpha}_{k},\ldots,\mathbb{H}{}^{\alpha}_{K})$ comprising all involved variables for K sets,

$\underset{\cdots}{\mathbb{H}}_k = \mathbb{H}\,{}^{\alpha}_{k} = \mathbb{P}_k\Phi\,{}^{\alpha}_{k}\mathbb{Q}_k'$, where α in this context is short for mode α_k $\qquad \forall k$

$\mathbb{W}^{\mathrm{ext}}$ denotes $\mathbb{W}$ extended blockwise in such a way that

$\mathbb{B}_{kl} = \underset{\cdots}{\mathbb{H}}_k'\underset{\cdots}{\mathbb{H}}_l w_{kl}$, for row block k and column block l, $\qquad \forall k,l$

$\mathbb{W}$ denotes a matrix with weights or function values w_{kl} and with diagonal elements equal to zero, $\mathbb{W}_{\mathrm{Diag}}=0$,

and

$$\mathbb{D}_v \quad = \quad \begin{array}{|c|c|c|} \hline v_1\mathbb{I}_{m_1} & 0 & 0 \\ \hline 0 & v_k\mathbb{I}_{m_k} & 0 \\ \hline 0 & 0 & v_K\mathbb{I}_{m_K} \\ \hline \end{array} \; .$$

If appropriate the sum of squares of the weights $\mathbb{W}$ have to be normalized to a constant value, tr $\mathbb{W}'\mathbb{W}=c$. Local deflation is according to (5.4) or user specified. As for the normalization of $\mathbf{t}$ it is important to notice that $\mathbf{t}$ has an explicit strong normalization $\mathbf{t}_k'\mathbb{H}_k'\underset{\cdots}{\mathbb{H}}_k\mathbf{t}_k=1$, $\forall k$, and at the same time an implicit weak normalization $(\mathbf{t}'\mathbf{t})^{-1}$. We call this a strong-weak normalization. The introduction of auxiliary matrix $\mathbb{D}_v$ with weights v_k in the LDC restrictions makes it always possible to find an optimal solution that satisfies this rigid strong-weak normalization. Without matrix $\mathbb{D}_v$

this is usually not possible. The relation of (5.18) with (5.10) can be made explicit by realizing that $\mathbb{t}_k'\mathbb{H}_k'\mathbb{H}_l\mathbb{t}_l = \mathbb{z}_k'\mathbb{z}_l$ denotes the directed correlation between unit normalized condensed variate $\mathbb{z}_k$ and pivot variate $\mathbb{z}_l$. Constant K in (5.10) is included in matrix $\mathbb{B}$ in (5.18).

Equation $\mathbb{z}_k = \pm\mathbb{S}_k^{\alpha}\mathbb{z}_{k*}((\mathbb{S}_k^{\alpha}\mathbb{z}_{k*})'\mathbb{S}_k^{\alpha}\mathbb{z}_{k*})^{-1/2} = \pm\mathbb{S}_k\mathbb{z}_{k*}((\mathbb{S}_k\mathbb{z}_{k*})'\mathbb{S}_k\mathbb{z}_{k*})^{-1/2}$, in (5.10) for the condensed variates, is in (5.18) incorporated in the restrictions
$\mathbb{B}\mathbb{t}=\psi\mathbb{D}_v\mathbb{t}$ and $\mathbb{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbb{t}_k=1$, and is redundant.

- Restriction $\mathbb{t}=\psi^{-1}\mathbb{D}_v^{-1}\mathbb{B}\mathbb{t}$ implies $\mathbb{z}_k = \mathbb{H}_k\mathbb{t}_k = \mathbb{H}_k\mathbb{B}_{k*}\mathbb{t}(v_k\psi)^{-1}$, with $\mathbb{B}_{k*} = (\mathbb{B}_{k1},\ldots,\mathbb{B}_{kl},\ldots,\mathbb{B}_{kK})$.
- Restriction $\mathbb{t}_l'\mathbb{H}_l'\mathbb{H}_l\mathbb{t}_l=1$ for the column blocks implies $\mathbb{z}_k = \mathbb{S}_k\mathbb{z}_{k*}(v_k\psi)^{-1}$, and finally restriction $\mathbb{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbb{t}_k=1$ for the row blocks implies equation $\mathbb{z}_k = \pm\mathbb{S}_k\mathbb{z}_{k*}((\mathbb{S}_k\mathbb{z}_{k*})'\mathbb{S}_k\mathbb{z}_{k*})^{-1/2}$ with the ± dependent on the sign of $(v_k\psi)$. Of course the restrictions for the row and column blocks are one and the same. They are presented sequentially only to simplify the derivation.

The condition $(\forall k\in J_k)$ for the weighted sum $\mathbb{z}_{k*}$ in (5.10) is not found in (5.18), because we confined ourselves to find a fitting procedure for path models where all pivot variates are also condensed variates.

5.3.6 Algorithm

In this section we elaborate an algorithm for LDC in several optimization steps. First we derive optimization steps for *fixed weights* $\mathbb{W}$ followed by additional equations for *proportional function weights* W. We end up with some remarks on *function weights*.

For *fixed weights* $\mathbb{W}$ we derive steps to find optimal $\mathbb{t}$ and $\mathbb{D}_v$ for fit function (5.18). If there were no strong normalizations ($\mathbb{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbb{t}_k=1$, $\forall k$) on $\mathbb{t}$ and no auxiliary matrix $\mathbb{D}_v$, the optimization problem would be to find a maximum for $(\mathbb{t}'\mathbb{B}\mathbb{t})(\mathbb{t}'\mathbb{t})^{-1}$, with restriction $\mathbb{B}\mathbb{t}=\psi\mathbb{t}$. For fixed weights this maximum attained if $\mathbb{t}$ is the right-hand eigenvector of matrix $\mathbb{B}$ with the largest eigenvalue. We explain this statement with a short intermezzo on Eigenvalue Decomposition of some nondefective square asymmetric matrix $\mathbb{A}$.

The Eigenvalue Decomposition of some nondefective square asymmetric matrix $\mathbb{A}$ (Golub & Van Loan, 1990, p.338) is given by

$$\mathbb{A} = \mathbb{U}\Lambda\mathbb{U}^{-1}, \tag{5.19}$$

where	$\mathbb{U}$	denote the right-hand eigenvectors,
with	$\mathrm{diag}(\mathbb{U}'\mathbb{U})=\mathbb{I}$,	
	$(\mathbb{U}^{-1})'$	denote the left-hand eigenvectors,
and	Λ	denotes a diagonal matrix with eigenvalues.

The right-hand and left-hand eigenvectors of $\mathbb{A}$ are usually not orthogonal (Wilkinson, 1965). If $\mathbb{A}$ is symmetric we have $\mathbb{U}'\mathbb{U}=\mathbb{I}$, and $\mathbb{U}^{-1}=\mathbb{U}'$. Linear equations for the right-hand eigenvectors of matrix $\mathbb{A}$ are

$$\mathbb{A}\mathbf{t}_{\mathrm{right}} = \psi\mathbf{t}_{\mathrm{right}},$$

and for the left-hand eigenvectors

$$\mathbf{t}_{\mathrm{left}}'\mathbb{A} = \psi\mathbf{t}_{\mathrm{left}}'.$$

The eigenvalues are given at stationary points by

$$\psi = \lambda_{(\mathbb{A})} = \frac{\mathbf{t}_{\mathrm{right}}'\mathbb{A}\mathbf{t}_{\mathrm{right}}}{\mathbf{t}_{\mathrm{right}}'\mathbf{t}_{\mathrm{right}}} = \frac{\mathbf{t}_{\mathrm{left}}'\mathbb{A}\mathbf{t}_{\mathrm{left}}}{\mathbf{t}_{\mathrm{left}}'\mathbf{t}_{\mathrm{left}}},$$

and they satisfy the equation $|\mathbb{A}-\lambda_{(\mathbb{A})}\mathbb{I}|=0$. Maximization of $\psi=(\mathbf{t}'\mathbb{A}\mathbf{t})(\mathbf{t}'\mathbf{t})^{-1}$, with restriction $\mathbb{A}\mathbf{t}=\psi\mathbf{t}$ is another formulation for finding the largest eigenvalue of $\mathbb{A}$ with corresponding right-hand eigenvector $\mathbf{t}$. Restriction $\mathbb{A}\mathbf{t}=\psi\mathbf{t}$ can be omitted if we maximize

$$\psi = \lambda_{(\mathbb{A})} = \frac{\mathbf{t}'\mathbb{C}_{\mathbb{A}}\mathbb{A}\mathbf{t}}{\mathbf{t}'\mathbb{C}_{\mathbb{A}}\mathbf{t}}, \tag{5.20}$$

where $\mathbb{C}_{\mathbb{A}} = (\mathbb{U}^{-1})'(\mathbb{U}^{-1})$ denotes the variance-covariance matrix of the left-hand eigenvectors $(\mathbb{U}^{-1})'$ of $\mathbb{A}$.

Matrix $\mathbb{C}_{\mathbb{A}}$ is symmetric and matrix $\mathbb{C}_{\mathbb{A}}\mathbb{A}$ is also symmetric. Restriction $\mathbb{A}\mathbf{t}=\psi\mathbf{t}$ is incorporated implicitly in (5.20), because the stationary equations are equal to this restriction. With substitution $\mathbb{A}=\mathbb{B}$ in the previous exposition on eigenvalue decomposition it is obvious why the optimization problem to find a maximum for $(\mathbf{t}'\mathbb{B}\mathbf{t})(\mathbf{t}'\mathbf{t})^{-1}$, with only restriction $\mathbb{B}\mathbf{t}=\psi\mathbf{t}$, is solved by taking for $\mathbf{t}$ the right-hand eigenvector $\mathbf{u}_{\mathrm{max}}$ of matrix $\mathbb{B}$ with the largest eigenvalue $\lambda_{\mathrm{max}(\mathbb{B})}$.

If $\mathbb{A}$ is p.s.d. the right-hand eigenvector $\mathbf{u}_{\mathrm{max}}$ with largest eigenvalue λ_{max} can be found with the Power Method by

$$\mathbb{A}\mathbf{t}^{i}\,(\,\mathbf{t}^{i\prime}\mathbb{A}'\mathbb{A}\mathbf{t}^{i}\,)^{-1/2} = \mathbf{t}^{i+1}. \tag{5.21}$$

If we want to apply the Power Method generally we have to substitute $\mathbb{A}=\mathbb{B}+c\mathbb{I}$ in (5.21), with some estimate for $c\geq-\lambda_{\mathrm{min}(\mathbb{B})}$ in order to make matrix $\mathbb{A}$ positive (semi-)

definite. This method works if all eigenvalues of $\mathbf{B}$ are distinct. It still works if matrix $\mathbf{B}$ has a large multiple eigenvalue and is similar to a diagonal matrix, although in that case the solution is not unique. If the Jordan canonical form of matrix $\mathbf{B}$ is not diagonal the Power Method does not work and other methods have to be used. For algorithm (5.21) the implicit weak normalization $(\mathbf{t}'\mathbf{t})^{-1}$ is temporarily changed into the explicit weak normalization $\mathbf{t}'\mathbf{t}=1$. In general with procedure (5.21) $\mathbf{t}$ converges to $\mathbf{u}_{\max}$ and ψ to $\lambda_{\max}$. We must emphasize that although the Power Method is monotone convergent with respect to (5.20), it is not always *monotone* convergent with respect to $\psi=(\mathbf{t}^{i\prime}\mathbf{A}\mathbf{t}^{i})(\mathbf{t}^{i\prime}\mathbf{t}^{i})^{-1}$. The reason for this phenomenon is that restriction $\mathbf{A}\mathbf{t}=\psi\mathbf{t}$ is violated during the iteration process and that it is only satisfied after convergence is reached. Therefore intermediate values of $\psi=(\mathbf{t}^{i\prime}\mathbf{A}\mathbf{t}^{i})(\mathbf{t}^{i\prime}\mathbf{t}^{i})^{-1}$ are only feasible if $\mathbf{t}$ is an eigenvector of $\mathbf{A}$.

We proceed further with the maximization of (5.18) for fixed weights $\mathbf{W}$ *with* strong normalizations and auxiliary matrix $\mathbf{D}_v$. In other words we have the optimization problem to find a maximum for $\psi=(\mathbf{t}'\mathbf{B}\mathbf{t})(\mathbf{t}'\mathbf{t})^{-1}$, with restrictions $\mathbf{D}_v^{-1}\mathbf{B}\mathbf{t}=\psi\mathbf{t}$, $\mathbf{t}'\mathbf{D}_v\mathbf{t}=\mathbf{t}'\mathbf{t}$ and $(\mathbf{t}_k'\mathbf{H}_k'\mathbf{H}_k\mathbf{t}_k=1, \forall k)$. The parameters v_k in $\mathbf{D}_v$ give extra freedom in order to be able to satisfy the strong restrictions on $\mathbf{t}$.

According to (5.20) we can also maximize

$$\psi = \lambda_{(\mathbf{D}_v^{-1}\mathbf{B})} = \frac{\mathbf{t}'\mathbf{C}_{(\mathbf{D}_v^{-1}\mathbf{B})}\mathbf{D}_v^{-1}\mathbf{B}\mathbf{t}}{\mathbf{t}'\mathbf{C}_{(\mathbf{D}_v^{-1}\mathbf{B})}\mathbf{t}}, \qquad (5.22)$$

with restrictions $\mathbf{t}'\mathbf{D}_v\mathbf{t}=\mathbf{t}'\mathbf{t}$ and $(\mathbf{t}_k'\mathbf{H}_k'\mathbf{H}_k\mathbf{t}_k=1, \forall k)$.

Applying principles of alternating least squares we maximize ψ with $\mathbf{t}$ and $\mathbf{D}_v$ fixed in turn. Correspondingly restrictions $\mathbf{t}'\mathbf{D}_v\mathbf{t}=\mathbf{t}'\mathbf{t}$ and $(\mathbf{t}_k'\mathbf{H}_k'\mathbf{H}_k\mathbf{t}_k=1, \forall k)$ also have to be relaxed in turn, because all restrictions can only be satisfied after convergence is reached.

For fixed $\mathbf{W}$ and $\mathbf{D}_v$ we find a maximum for (5.22), by taking for $\mathbf{t}$ the right-hand eigenvector $\mathbf{u}_{\max}$ of matrix $\mathbf{D}_v^{-1}\mathbf{B}$ with the largest eigenvalue $\lambda_{\max(\mathbf{D}_v^{-1}\mathbf{B})}$. For the Power Method we have to substitute $\mathbf{A}=\mathbf{D}_v^{-1}\mathbf{B}+c\mathbf{I}$ in (5.21), with an estimate for c in such away that $c\geq-\lambda_{\min(\mathbf{D}_v^{-1}\mathbf{B})}$ and therefore matrix $\mathbf{A}$ is positive (semi-) definite. A good starting value for $\mathbf{D}_v$ fixed is to take $v_k=1, \forall k$. It gives a global maximum for (5.18) without strong normalizations on $\mathbf{t}$. After the fixed $\mathbf{W}$ and $\mathbf{D}_v$ optimization step the strong normalization of $\mathbf{t}$ has to be assessed in the next step and if necessary adjusted.

For fixed $\mathbb{W}$ and $\mathbf{t}$ we find an optimal $\mathbb{D}_v$ in such an way that restrictions $\mathbf{t}'\mathbf{D}_v\mathbf{t}=\mathbf{t}'\mathbf{t}$ and ($\mathbf{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbf{t}_k=1$, $\forall k$) are satisfied. After the previous optimization step we have satisfied equality $\mathbb{D}_v^{-1}\mathbf{B}\mathbf{t}=\psi\mathbf{t}$. This equality can be partitioned and transformed in K equalities

$$\mathbb{H}_k\mathbf{B}_{k*}\mathbf{t} = \psi v_k\mathbb{H}_k\mathbf{t}_k, \qquad \forall k$$

with $\mathbb{B}_{k*} = (\mathbb{B}_{k1},\ldots,\mathbb{B}_{kl},\ldots,\mathbb{B}_{kK})$.

Applying thereupon restriction $\mathbf{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbf{t}_k = 1$ results in

$$\mathbf{d}_k(\mathbf{t}) = \psi v_k = (\mathbf{t}'\mathbb{B}_{k*}'\mathbb{H}_k'\mathbb{H}_k\mathbf{B}_{k*}\mathbf{t})^{1/2}\text{sign}(\mathbf{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbf{B}_{k*}\mathbf{t}),$$

or $$\mathbb{d}_k(\mathbf{t}) = \psi v_k = \mathbf{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbf{B}_{k*}\mathbf{t}, \qquad \forall k$$

with $\text{sign}(x)$ denoting the sign of x.

In matrix notation we have the equalities

$$\left\{\begin{array}{ccc} \mathbf{D}(\mathbf{t}) & = & \psi\mathbf{D}_v \\ \mathbb{D}(\mathbf{t}) & = & \psi\mathbb{D}_v \end{array}\right\}, \tag{5.23}$$

with $$\mathbf{D}(\mathbf{t}) = \begin{array}{|c|c|c|}\hline \mathbf{d}_1(\mathbf{t})\mathbf{I}_{m_1} & 0 & 0 \\ \hline 0 & \mathbf{d}_k(\mathbf{t})\mathbf{I}_{m_k} & 0 \\ \hline 0 & 0 & \mathbf{d}_K(\mathbf{t})\mathbf{I}_{m_K} \\ \hline \end{array},$$

$$\mathbb{D}(\mathbf{t}) = \begin{array}{|c|c|c|}\hline \mathbb{d}_1(\mathbf{t})\mathbb{I}_{m_1} & 0 & 0 \\ \hline 0 & \mathbb{d}_k(\mathbf{t})\mathbb{I}_{m_k} & 0 \\ \hline 0 & 0 & \mathbb{d}_K(\mathbf{t})\mathbb{I}_{m_K} \\ \hline \end{array},$$

$$\mathbf{d}_k(\mathbf{t}) = (\mathbf{t}'\mathbb{B}_{k*}'\mathbb{H}_k'\mathbb{H}_k\mathbf{B}_{k*}\mathbf{t})^{1/2}\text{sign}(\mathbf{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbb{B}_{k*}\mathbf{t}),$$

and $$\mathbb{d}_k(\mathbf{t}) = \mathbf{t}_k'\mathbb{H}_k'\mathbb{H}_k\mathbf{B}_{k*}\mathbf{t}.$$

Finally by applying subsequently $\mathbf{t}'\mathbf{D}_v\mathbf{t}=\mathbf{t}'\mathbf{D}(\mathbf{t})\mathbf{t}\psi^{-1}=\mathbf{t}'\mathbf{t}$ and $\mathbf{t}'\mathbb{D}_v\mathbf{t}=\mathbf{t}'\mathbb{D}(\mathbf{t})\mathbf{t}\psi^{-1}=\mathbf{t}'\mathbf{t}$, we obtain the required formulas for an optimal $\mathbb{D}_v$

$$\mathbb{D}_\nu = \mathbb{D}(\mathbf{t})\frac{\mathbf{t}'\mathbf{t}}{\mathbf{t}'\mathbb{D}(\mathbf{t})\mathbf{t}}, \tag{5.24}$$

or

$$\mathbb{D}_\nu = \mathbb{D}(\mathbf{t})\frac{\mathbf{t}'\mathbf{t}}{\mathbf{t}'\mathbb{D}(\mathbf{t})\mathbf{t}}. \tag{5.25}$$

In summary, the optimal $\mathbf{t}$ and $\mathbb{D}_\nu$ for maximizing (5.18) with fixed weights $\mathbb{W}$ can be found by alternately taking for $\mathbf{t}$ the right-hand eigenvector $\mathbf{u}_{\max}$ of matrix $\mathbb{D}_\nu^{-1}\mathbb{B}$ with the largest eigenvalue $\lambda_{\max(\mathbb{D}_\nu^{-1}\mathbb{B})}$ and for $\mathbb{D}_\nu$ the function values of (5.24) or (5.25). Another strategy is to perform only one step of the Power Method (5.21) corrected with some constant $c \geq -\lambda_{\min(\mathbb{D}_\nu^{-1}\mathbb{B})}$ and then to update $\mathbb{D}_\nu$. The first procedure is usually converging very fast. The second procedure can only be competitive in consuming CPU time, if an estimate for c is chosen, large enough to make matrix $\mathbb{A}$ positive (semi-)definite during all iterations. A lower bound for c can be found, because restriction $\mathbf{t}'\mathbb{D}_\nu\mathbf{t}=\mathbf{t}'\mathbf{t}$ is satisfied after each step of the Power Method. For $\psi_{\min}$ we have $\mathbb{D}_\nu^{-1}\mathbb{B}\mathbf{t}_{\min}=\psi_{\min}\mathbf{t}_{\min}$, which implies $\mathbf{t}_{\min}'\mathbb{B}\mathbf{t}_{\min}=\psi_{\min}\mathbf{t}_{\min}'\mathbb{D}_\nu\mathbf{t}_{\min}=\psi_{\min}\mathbf{t}_{\min}'\mathbf{t}_{\min}$ and therefore a lower bound for ψ also gives an estimate for c in this context. One possible estimate for c is

$$c = -\lambda_{\min(.5\mathrm{x}\mathbb{B}+\mathbb{B}')} \geq -\psi_{\min} = -\frac{\mathbf{t}_{\min}'\mathbb{B}\mathbf{t}_{\min}}{\mathbf{t}_{\min}'\mathbf{t}_{\min}}. \tag{5.26}$$

A third procedure can be defined by constructing the positive semi-definite matrix $\underset{\cdot\cdot}{\mathbb{B}}=\mathbb{B}+c\mathbb{I}$, with $c \geq -\lambda_{\min(.5\mathrm{x}\mathbb{B}+\mathbb{B}')}$. Substitution of $\underset{\cdot\cdot}{\mathbb{B}}$ in (5.26) always gives a positive or zero value for $\psi_{\min}$. A simple attractive algorithm results if we substitute $\underset{\cdot\cdot}{\mathbb{B}}$ instead of $\mathbb{B}$ in (5.23) and (5.24), and subsequently substitute (5.24) in $\mathbb{A}=\mathbb{D}_\nu^{-1}\underset{\cdot\cdot}{\mathbb{B}}$ and $\mathbb{A}$ in (5.21). For future comparison with the basic PLS algorithm we have chosen to substitute (5.24) and not (5.25). The total iteration process with fixed weights $\mathbb{W}$ reduces now to

$$\underset{\cdot\cdot}{\mathbb{D}}(\mathbf{t}^i)^{-1}\underset{\cdot\cdot}{\mathbb{B}}\mathbf{t}^i = \mathbf{t}^{i+1}, \tag{5.27}$$

where $\underset{\cdot\cdot}{\mathbb{D}}(\mathbf{t})$ denotes $\mathbb{D}(\mathbf{t})$ in (5.23) with $\mathbb{B}$ replaced by $\underset{\cdot\cdot}{\mathbb{B}}=\mathbb{B}+c\mathbb{I}$,
with $c \geq -\lambda_{\min(.5\mathrm{x}\mathbb{B}+\mathbb{B}')}$.

Convergence is reached if $1-\mathbf{t}^{i\prime}\mathbf{t}^{i+1}(\mathbf{t}^{i\prime}\mathbf{t}^{i}\mathbf{t}^{i+1\prime}\mathbf{t}^{i+1})^{-1/2}$ is sufficiently close to zero. Algorithm (5.27) for the strong-weak normalized $\mathbf{t}$ is only slightly modified compared to the Power Method for the weak normalized $\mathbf{t}$. Generally we expect to find a global maximum for (5.18), especially if we take the optimal $\mathbb{t}$ from (5.21) with $\mathbb{A}=\mathbb{B}$ as a starting point $\mathbb{t}^0$ for (5.27). If the Power Method is not feasible we have to rely on other methods for computing the largest eigenvalue and corresponding eigenvectors.

Having established linear equations and an algorithm for (5.18) for *fixed weights* $\mathbb{W}$ we will now consider the necessary modifications for *proportional function weights* $\mathbb{W}$. We simply have to add an extra step in the alternating least squares algorithm described so far. For $\mathbb{t}$ fixed we have to maximize (5.10), rewritten as

$$\text{LDC:} \qquad \mathrm{Fit}(\mathbb{W}) = \frac{\mathbb{1}'(\mathbb{R}*\mathbb{W})\mathbb{1}}{\mathbb{t}'\mathbb{t}K^{-1}} = \mathbb{c}'\mathrm{vec}(\mathbb{W}), \tag{5.28}$$

where $\mathbb{c} = \mathrm{vec}((\mathbb{t}'\mathbb{t})^{-1}K\mathbb{R})$ denotes matrix $(\mathbb{t}'\mathbb{t})^{-1}K\mathbb{R}$ strung out to a vector,

with $\mathrm{vec}(\mathbb{W})'\mathrm{vec}(\mathbb{W})=\mathbb{1}'\mathbb{W}_{\mathrm{Design}}\mathbb{1}$. Maximization of (5.28) is equivalent to minimizing the residual variance $\mathbb{e}'\mathbb{e}$, where $\mathbb{e}=\mathbb{c}-\mathrm{vec}(\mathbb{W})$ for unrestricted $\mathbb{W}$ or some appropriate nonlinear transformation of $\mathbb{W}$. (See Gifi, 1990, page 529 and Kruskal &Carroll, 1969.) For being proportional function weights the weights $\mathbb{W}$ are restricted to be proportional to the values of the multivariate *weight function* $\mathbb{F}(\mathbb{t})$,

$$\mathbb{W} = \beta\mathbb{F}(\mathbb{t}).$$

One example of such a weight function is $\mathbb{F}(\mathbb{t})=\mathbb{R}*\mathbb{W}_{\mathrm{Design}}$, where the weights are proportional to the directed correlations or equal to zero, according to the design of the weights. With this weight function maximization of (5.28) gives the same result as for unrestricted $\mathbb{W}$. Summarising the estimation of the proportional function weights we have to add the following step in the alternating least squares algorithm described so far

$$\mathbb{W} = \mathbb{F}(\mathbb{t})\left(\frac{\mathbb{1}'\mathbb{W}_{\mathrm{Design}}\mathbb{1}}{\mathrm{tr}\mathbb{F}(\mathbb{t})'\mathbb{F}(\mathbb{t})}\right)^{1/2}. \tag{5.29}$$

Equality (5.29) has to be alternated with (5.27) or an equivalent eigenvector step in order to obtain an algorithm for (5.18) with proportional function weights. For this algorithm we do not expect to find always a global maximum.

In the beginning of this section we promised some remarks on *function weights*. Until now we have not established a general algorithm for function weights. The only MVA method we have encountered that fits a path model with function weights is the SCA method of chapter 3. For this special case we have developed an algorithm. All other MVA methods can be described with fixed function weights or proportional function weights.

5.4 Relations of LDC with other methods

In section 5.4.1 a short introduction to Wold's *basic PLS method* of Lohmöller (1989) is translated into our notation. Thereafter we give the corresponding basic PLS algorithm in section 5.4.2 and show the relation with LDC in section 5.4.3. In section 5.4.5 the *extended PLS method* proposed by Lohmöller (1989) is brought within the LDC framework. Section 5.4.6 gives the LDC formulation of *consensus PLS* proposed by Geladi & Martens (1988) and also gives an equivalent 'variance accounted for' criterion that is fitted by the consensus PLS algorithm. Section 5.4.7 elaborates the *PLS1 regression* method and *Continuum Regression* proposed by Stone & Brooks (1990). The reflected variance methods of chapter 4 are discussed in section 5.4.8 and from the corresponding LDC path models arises an interesting two sets PLS method. Last but not least we show in section 5.4.9 that *Set Component Analysis* of chapter 3 is an example of fitting a DC path model with real function weights. Some LDC extensions of the SCA path model are formulated, like the PLS *Hierarchical Components* method.

5.4.1 Wold's basic method of Soft Modelling

Herman Wold (1982) has introduced a type of modelling with latent variables which he calls "Soft Modelling". The name indicates that this sort of model building applies when the theoretical knowledge is scarce and stringent distributional assumptions are not applicable. Lohmöller (1989) calls this method the "basic Partial Least Squares method". As no single criterion had been established 'Partial Least Squares' or 'PLS' refers to the partitioning of parameters in estimable subsets.

In order to avoid an overflow of new symbols for readers used to PLS notation, we will present only the *estimated* PLS models in our notation.

Variables. A soft model involves manifest variables (MV's) and latent variables (LV's) related by linear equations. The MV's (directly observed, observables, indicators) are partitioned into non-overlapping subsets of K blocks $\mathbb{H}_k$ with m_k manifest variables, each block being indicative of one LV or variate $\mathbb{z}_k$. According to Lohmöller all involved variables and variates can be treated as deviations of means without loss of generality.

Inner model or structural model. The variates $\mathbb{Z}=(\mathbb{z}_1,\ldots,\mathbb{z}_k,\ldots,\mathbb{z}_K)$ are assumed to be interconnected by one or more linear relations. The basic method requires the variates to form a recursive path model (a causal chain),

$$\mathbb{Z} = \mathbb{Z}\mathrm{A} + \mathbb{E}_{\mathrm{A}}, \tag{5.30}$$

where $\mathrm{A}=(\mathbf{a}_1,\ldots,\mathbf{a}_k,\ldots,\mathbf{a}_K)$ denote regression weights or path coefficients,
$\mathbb{E}_{\mathrm{A}}$ denote residual variables,
with $(\mathbb{E}_{\mathrm{A}}'\mathbb{Z})_{\mathrm{Offdiag}}=0$, all offdiagonal values of $\mathbb{E}_{\mathrm{A}}'\mathbb{Z}$ equal to zero.

The design matrix $\mathrm{A}_{\mathrm{Design}}$ of a recursive path model is subdiagonal.

Outer model or measurement model. The m_k manifest variables $\mathbb{H}_k$ are assumed to be generated as a linear function of its variate $\mathbb{z}_k$ and the outer residual variables $\mathbb{E}_k$,

$$\mathbb{H}_k = \mathbb{z}_k\mathbf{c}_k' + \mathbb{E}_k, \tag{5.31}$$

where $\mathbf{c}_k$ denotes a vector with m_k loadings for set k,
$\mathbb{E}_k$ denote residual variables for set k,
with $\mathbb{E}_k'\mathbb{z}_k=0$.

Weight relations. As a vehicle for the estimation of the model parameters, the variates $\mathbb{z}_k$ are estimated as weighted aggregates of their indicators,

$$\mathbb{z}_k = \mathbb{H}_k\mathbf{t}_k, \tag{5.32}$$

where $\mathbf{t}_k$ denotes a vector with m_k variable weights,
with $\mathbf{z}_k'\mathbf{z}_k=1$.

The weights are estimated by least squares methods in two different versions. In the first version (called mode A, outwards directed, or factor mode) the manifest variables $\mathbf{H}_k$ are regressed on an instrumental variate $\mathbf{z}_{k*}$ (the so-called inside approximation)

$$\mathbf{H}_k = \mathbf{z}_{k*}\mathbf{t}_{k*}' + \mathbf{E}_{k*}, \tag{5.33}$$

where $\mathbf{t}_{k*}$ denotes a vector with m_k regression weights,
$\mathbf{E}_{k*}$ denote residual variables for set k,
with $\mathbf{E}_{k*}'\mathbf{z}_{k*}=0$.

and the variances of the outer residuals $\mathbf{E}_{k*}$ in (5.33) are minimized for unknown $\mathbf{t}_{k*}$. In the second version (called mode B, inwards directed, or regression mode) the instrumental variate $\mathbf{z}_{k*}$ is regressed on the manifest variables $\mathbf{H}_k$

$$\mathbf{z}_{k*} = \mathbf{H}_k\mathbf{t}_{k*} + \mathbf{e}_{k*}, \tag{5.34}$$

where $\mathbf{t}_{k*}$ denotes a vector with m_k regression weights,
$\mathbf{e}_{k*}$ denote a residual variable for set k,
with $\mathbf{e}_{k*}'\mathbf{z}_{k*}=0$.

The weights $\mathbf{t}_k$ in (5.32) are rescaled versions of the provisional weights $\mathbf{t}_{k*}$ and provide that $\mathbf{z}_k$ in (5.32) is unit normalized.

5.4.2 The basic PLS algorithm

The algorithm for estimating the unknowns of the models proceeds in three stages. In the first two stages the variables $\mathbf{H}_k$ and variates $\mathbf{z}_k$ are centred. In the third stage the variate means and the location parameters are estimated. Stage three will not be discussed here, because we omitted without loss of generality the means and location parameters in the definition of the estimated models in section 5.4.1.

Before specifying the basic PLS algorithm, the subdiagonal path design matrix $\mathbf{A}_{\text{Design}}$ constructed from a recursive path model has to be completed with an upper-diagonal part in a command design matrix $\mathbf{W}_{\text{Design}}$ with a corresponding command diagram. The term command design matrix has no explicit reference to PLS literature, but the command diagram is extensively discussed by Bookstein (1982). The command design matrix is implicitly defined by the choice of several optimization operators. We will show this in section 5.4.4. Lohmöller (1989) calls the command or weight matrix $\mathbb{W}$ the 'inner weight matrix', but he does not clearly emphasize the difference between the command design matrix $\mathbb{W}_{\text{Design}}$ and the subdiagonal path design matrix $\mathbb{A}_{\text{Design}}$. Knowing $\mathbb{W}_{\text{Design}}$ the PLS algorithm for the basic method of soft modelling is given by

Stage1: Iterative estimation of weights $\mathbb{t}_k$ and variates $\mathbb{z}_k$. Starting at Step 4, repeat Steps 1 to 4 until convergence is obtained.

Step 1. Inner weights

Compute $\mathbb{W} = \mathbb{W}_{\text{Design}} * \mathbb{R}_{\text{Sign}}$,

where $\mathbb{X}_{\text{Sign}}$ denotes a matrix with the signs of the elements of $\mathbb{X}$,

and $\mathbb{R} = \mathbb{Z}'\mathbb{Z}$ correlations between the variates $\mathbb{Z}=(\mathbb{z}_1,\ldots,\mathbb{z}_k,\ldots,\mathbb{z}_K)$.

Step 2. Inside approximation

Compute $\mathbb{z}_{k*} = \sum_{l=1}^{K} w_{kl}\mathbb{z}_l$.

Step 3. Outer weights. Solve for $\mathbb{t}_{k*}$ in (5.33) or (5.34)

$\mathbb{H}_k = \mathbb{z}_{k*}\mathbb{t}_{k*}' + \mathbb{E}_{k*}$, for set k in mode A

$\mathbb{z}_{k*} = \mathbb{H}_k\mathbb{t}_{k*} + \mathbb{e}_{k*}$, for set k in mode B

Step 4. Outside approximation

Compute $\mathbb{z}_k = \mathbb{H}_k\mathbb{t}_k = \mathbb{H}_k\mathbb{t}_{k*}((\mathbb{H}_k\mathbb{t}_{k*})'\mathbb{H}_k\mathbb{t}_{k*})^{-1/2}$.

Stage2: Estimation of path coefficients $\mathbb{A}$ and loadings $\mathbb{c}_k$ by minimizing in least squares sense the error of respectively (5.30) and (5.31).

We consider *Stage1* as most essential for the basic PLS algorithm and therefore classify basic PLS as a cyclic hybrid method. Cyclic hybrid methods maximize several fit functions cyclically, while utilizing optimal parameters of previously fitted models, until a stationary phase is reached.

5.4.3 Basic PLS a special case of LDC

Having specified the basic PLS algorithm in section 5.4.2, we can now establish the relation with the LDC algorithm of section 5.3.6. In PLS stage 1 the variates $\mathbf{z}_k$ are computed and all other PLS parameters can easily be derived from this solution. Therefore we will compare the basic PLS algorithm with the LDC algorithm with respect to the computation of the variates $\mathbf{z}_k$ in stage 1.

First we simplify the basic PLS algorithm. The variables $\mathbf{H}_k$ for the regression mode B in Step 3 can be replaced without loss of generality by the orthonormal basis $\mathbf{H}_k^0 = \mathbf{P}_k\mathbf{Q}_k'$. Only the weights $\mathbf{t}_{k*}$ will change, but not the corresponding variate $\mathbf{z}_k$ in Step 4. Step 3 can now be reduced to

Step 3. Outer weights. Solve for $\mathbf{t}_{k*}$

$$\mathbf{H}_k^{\alpha} = \mathbf{z}_{k*}\mathbf{t}_{k*}' + \mathbf{E}_{k*}, \qquad \text{with } \alpha_k = 1 \text{ for mode A, } \alpha_k = 0 \text{ for mode B.}$$

Subsequently Step 2, 3 and 4 can now be reduced to

$$\mathbf{z}_k = \mathbf{S}_k^{\alpha}\mathbf{z}_{k*}((\mathbf{S}_k^{\alpha}\mathbf{z}_{k*})'\mathbf{S}_k^{\alpha}\mathbf{z}_{k*})^{-1/2}, \tag{5.35}$$

with $\mathbf{z}_{k*} = \sum_{l=1}^{K} w_{kl}\mathbf{z}_l$,

$\alpha_k = 1$ for mode A and $\alpha_k = 0$ for mode B.

Substituting in (5.35) equality $\mathbf{S}_k\mathbf{z}_{k*} = \mathbf{H}_k\mathbf{B}_{k*}\mathbf{t}$ from section 5.3.5, with $\mathbf{H}_k = \mathbf{H}_k^{\alpha}$, and premultiplying both sides with $\mathbf{H}_k^{-1}$ we obtain an algorithm for basic PLS for finding an optimal $\mathbf{t}$ and therefore optimal variates $\mathbf{z}_k = \mathbf{H}_k\mathbf{t}_k$. This algorithm consist of two steps

Step 1. Inner weights

Compute $\mathbf{W} = \mathbf{W}_{\text{Design}} * \mathbf{R}_{\text{Sign}}$. (5.36)

Step 2. Variable weights

Compute $\mathbf{D}(\mathbf{t}^i)_{\text{Abs}}^{-1}\mathbf{B}\mathbf{t}^i = \mathbf{t}^{i+1}$, (5.37)

where $\mathbf{X}_{\text{Abs}}$ denotes a matrix with the absolute values of $\mathbf{X}$,

and $\mathbf{D}(\mathbf{t}^i)$ denotes a diagonal matrix as defined in (5.23).

Because the sign of $\mathbb{D}(\mathbb{t}^i)$ in (5.37) is compensated in (5.36) by the sign of the correlations $\mathbb{R}$, we are allowed to replace (5.37) by

$$\text{Compute } \mathbb{D}(\mathbb{t}^i)^{-1}\mathbb{B}\mathbb{t}^i = \mathbb{t}^{i+1}, \tag{5.38}$$

In summary, the algorithm for basic PLS can be reduced to alternating between (5.36) and (5.38). Arriving at this point we can make a comparison with the LDC algorithm. This algorithm consist of alternating between (5.29) and (5.27). Formula (5.29) is equal to (5.36), if we define the weight function $\mathbb{F}(\mathbb{t})$ by

$$\mathbb{F}(\mathbb{t}) = \mathbb{W}_{\text{Design}} * \mathbb{R}_{\text{Sign}}. \tag{5.39}$$

Formula (5.27) is equal to (5.38), if $\lambda_{\min(.5\text{x}\mathbb{B}+\mathbb{B}')} \geq 0$. Then for $c=0$ we have $\underset{\cdot\cdot}{\mathbb{B}}=\mathbb{B}$. Therefore the basic PLS algorithm uses in principle the Power Method on a matrix that can have negative eigenvalues. In this situation the Power Method usually converges to the eigenvector with the largest absolute eigenvalue, positive or negative.

Because the LDC algorithm converges to a maximum for (5.18), the basic PLS algorithm will also converge to a maximum if $\lambda_{\min(.5\text{x}\mathbb{B}+\mathbb{B}')}$ stays greater than or equal to zero during the iteration process. The PLS algorithm will also converge to a maximum if $\lambda_{\max(.5\text{x}\mathbb{B}+\mathbb{B}')}$ stays much larger in absolute value than $\lambda_{\min(.5\text{x}\mathbb{B}+\mathbb{B}')}$ during the iteration process. If $\lambda_{\min(.5\text{x}\mathbb{B}+\mathbb{B}')}$ is more or less equal to $\lambda_{\max(.5\text{x}\mathbb{B}+\mathbb{B}')}$ the basic PLS algorithm might not converge to an optimal solution. If $\lambda_{\min(.5\text{x}\mathbb{B}+\mathbb{B}')}$ is much larger in absolute value than $\lambda_{\max(.5\text{x}\mathbb{B}+\mathbb{B}')}$ the basic PLS algorithm will simply switch the signs of the weights $\mathbb{W}$ in (5.36) and proceed to find a maximum with $\lambda_{\max(.5\text{x}\mathbb{B}+\mathbb{B}')}$ much larger in absolute value than $\lambda_{\min(.5\text{x}\mathbb{B}+\mathbb{B}')}$. Therefore the basic PLS algorithm will generally find a maximum for LDC (5.10) with weight function (5.39) substituted

$$\text{bPLS:} \quad \text{Fit}(\mathbb{t}) = \frac{\mathbb{1}'(\mathbb{R}_{\text{Abs}} * \mathbb{W}_{\text{Design}})\mathbb{1}}{\mathbb{t}'\mathbb{t}K^{-1}}, \tag{5.40}$$

where $\mathbb{R} = \mathbb{Z}'\mathbb{Z}$ denotes a $(K \times K)$ symmetric correlation matrix with directed correlations,

$\mathbb{X}_{\text{Abs}}$ denotes a matrix with the absolute values of $\mathbb{X}$.

In some special cases the PLS algorithm will *not converge*, because it is not using the Power Method in a proper way. Nevertheless the basic PLS algorithm generally finds a global maximum for (5.40) due to the sign switching of the weights $\mathbf{W}$ described above. The LDC algorithm *always* finds a local or global maximum. By computing maximum solutions starting with several feasible non-sign-similar $\mathbf{W}$'s we usually find the global and one or more local maxima for the basic PLS fit function (5.40). Two matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ are sign-similar if they can be made equal by changing the signs of rows and corresponding columns. So $\mathbf{W}_1 = \mathbf{D}\mathbf{W}_2\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with diagonal elements 1 or -1. In other words two matrices defined by (5.36) are sign-similar if they can be made equal by sign transformations of the variates. Therefore a group of g sign similar weight matrices $\mathbf{W}$ leads to g maxima of (5.40), that differ only with respect to the sign of the variates. We call this a sign-similar solution of bPLS. The matrices $\mathbf{W}$ are feasible if they do not violate the restrictions imposed by the weight function, which is (5.39) for bPLS. Observing this weight function we know for instance that weights matrices $\mathbf{W}$ are not feasible, if $w_{ij} = -w_{ji}$ for some elements. The number c of feasible non-sign-similar $\mathbf{W}$'s defines the number c of sign-similar solutions of bPLS. Therefore we need c feasible non-sign-similar starting values for $\mathbf{W}$ in order to be sure to find the global maximum with the LDC algorithm. In a case of three sets, where for instance all non-diagonal elements are non-zero, $c=2$. In the case of two sets PLS ($K=2$) we have $c=1$, and we always find two equal global maxima, which only differ with respect to the sign of the variates $\mathbf{z}_1$ and $\mathbf{z}_2$. Generally we state that we will find with the LDC algorithm only global maxima with different signs of variates, if all feasible $\mathbf{W}$'s are sign-similar ($c=1$). From this statement we deduce on the other hand that the bPLS algorithm will *always* find a global maximum for the LDC fit function, if all possible $\mathbf{W}$'s are sign-similar ($c=1$) and $\mathbf{W}_{\text{Sign}} = \mathbf{W}_{\text{Design}} * \mathbf{R}_{\text{Sign}}$.

5.4.4 The command design matrix

The command design matrix $\mathbf{W}_{\text{Design}}$ is implicitly defined by the choice of several optimization operators. Sometimes ancillary blocks are added in the command diagram. Bookstein (1982) describes six operators, called mode A to F. By applying the basic PLS fit function (5.40) we will show how all these optimization operators can be fitted with LDC path models. In our terminology the optimization operators or *Opt* operators define the construction of the condensed and regression variates. The *Opt* command has the general form of

$$\mathbf{z}_k = Opt_{\mathrm{X}}(\mathbf{H}_k, \mathbf{z}_j, \ldots, \mathbf{z}_l), \tag{5.41}$$

where $\mathbf{z}_k$ denotes the condensed or regression variate with variables $\mathbf{H}_k$ and $\mathbf{z}_j, \ldots, \mathbf{z}_l$ denote one or more pivot variates. X denotes the mode of the *Opt* command. In the case of a two sets path model (K=2) we have only one pivot variate in (5.41). In this case we can choose between two *Opt* commands, Opt_{A} or Opt_{B}. The PLS path diagram in figure 5.5 can therefore be fitted with three command diagrams given in figure 5.6.A to 5.6.C.

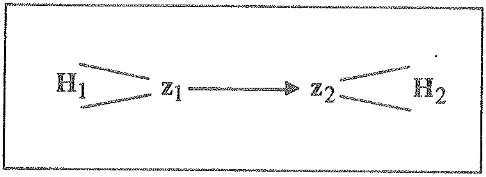


Figure 5.5 *PLS path diagram for two sets.*

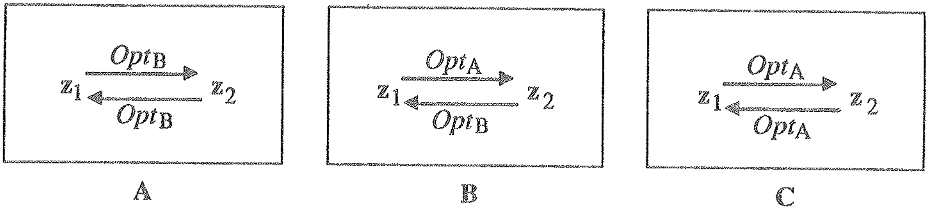


Figure 5.6 *PLS command diagrams for two sets.*

In figure 5.7.A to 5.7.C we give the corresponding LDC path diagrams with primary (α_k=0) and secondary (α_k=1) predictions according to the prototypes in respectively figure 5.3 and 5.2.

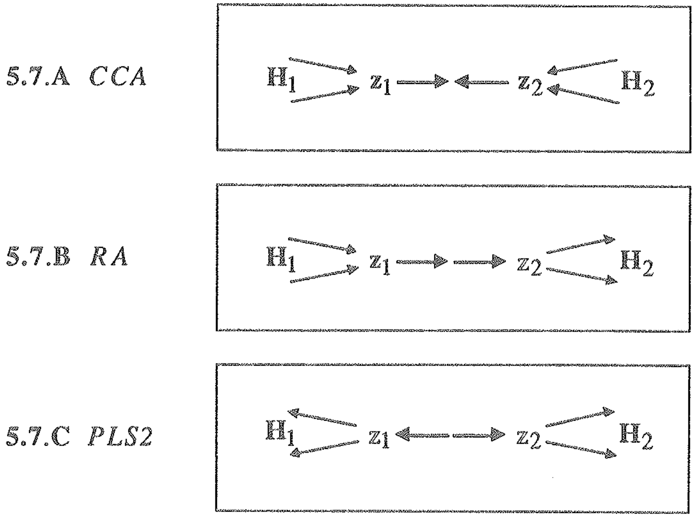


Figure 5.7 *LDC path diagrams for two sets.*

The fourth combination of Opt_A and Opt_B would be to reverse the direction of the arrows in 5.7.B in order to 'predict' the regression variate of set 2 with the condensed variate of set 1. This paradoxical combination is omitted, because at least primary or secondary prediction must be on line with the chosen flow of prediction. The PLS path design matrix A_{Design} (5.30) for the path model in figure 5.5 is

$$A_{Design} = \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline\end{array} , \qquad (5.42)$$

and the command design matrix W_{Design} for all diagrams in figure 5.5, 5.6 and 5.7 is

$$W_{Design} = \begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 0 \\ \hline\end{array} . \qquad (5.43)$$

From the LDC path diagrams we can easily derive directed correlations matrices $\mathbb{R}$. The rows of these matrices always give the condensed or regression variates and the columns the pivot variates. The condensed variates with α_k=1 will always be indicated in outline and the regression variates with α_k=0 in **bold**. For figure 5.7.A the corresponding directed correlations matrix $\mathbb{R}$ with the desired design is

$$\mathbb{R}{*}\mathbb{W}_{\mathrm{Design}} \quad = \quad \begin{array}{|c|c|}\hline 0 & \mathbb{z}_1'\mathbb{z}_2 \\ \hline \mathbb{z}_2'\mathbb{z}_1 & 0 \\ \hline \end{array} \quad . \tag{5.44}$$

For figure 5.7.B the corresponding matrix is

$$\mathbb{R}{*}\mathbb{W}_{\mathrm{Design}} \quad = \quad \begin{array}{|c|c|}\hline 0 & \mathbb{z}_1'\mathbb{z}_2 \\ \hline \mathbb{z}_2'\mathbb{z}_1 & 0 \\ \hline \end{array} \quad , \tag{5.45}$$

and for figure 5.7.C

$$\mathbb{R}{*}\mathbb{W}_{\mathrm{Design}} \quad = \quad \begin{array}{|c|c|}\hline 0 & \mathbb{z}_1'\mathbb{z}_2 \\ \hline \mathbb{z}_2'\mathbb{z}_1 & 0 \\ \hline \end{array} \quad . \tag{5.46}$$

Maximizing the bPLS fit function (5.40) for (5.44), (5.45) and (5.46) we find two equal global maxima, which only differ with respect to the sign of the variates of set 1 and 2. The bPLS solutions for (5.44), (5.45) and (5.46) can be linked to solutions of well-known MVA methods (Lohmöller, 1989, p.110). For (5.44) the bPLS solution is equal to a one dimensional CCA solution (2.23) with canonical variates $\mathbb{z}_1$ and $\mathbb{z}_2$ and canonical correlation $\mathbb{z}_1'\mathbb{z}_2$. For (5.45) the bPLS solution is equal to a one dimensional solution of the Principal Predictor model, where the canonical variate of the predictor set is equal to $\mathbb{z}_1$ and the variate of the criteria is equal to $\mathbb{z}_2$. When $\mathbb{z}_2$ is omitted we have the RA solution (2.26) and also Fortier's simultaneous linear prediction (Fortier, 1966). For (5.46) the bPLS solution is equal to a one dimensional solution of Tucker's (1958) Interbattery Factor model. In chemometrics the corresponding asymmetric deflation algorithm is usually called the PLS2 method (Manne, 1987). The predictor variable $\mathbb{z}_1$ is usually called the column vector of scores ($\mathbb{t}_1$) for the independent block $\mathbb{X}$, and $\mathbb{z}_2$ the scores ($\mathbb{u}_1$) for the dependent block $\mathbb{Y}$ (Geladi & Kowalski, 1986).

In the case of three or more sets path models Bookstein (1982) describes four additional operators, called Opt_C to Opt_F. From his geometrical description, which we found more consistent than the command diagrams, we have distiled the projections for all operators in table 5.1.

Table 5.1 *Optimization operators for PLS.*

$$
\begin{aligned}
\mathbf{z}_k &= \mathrm{Opt}_A(\mathbf{H}_k,\mathbf{z}_l) = \mathbf{S}^1_k\mathbf{z}_{k*}((\mathbf{S}^1_k\mathbf{z}_{k*})'\mathbf{S}^1_k\mathbf{z}_{k*})^{-1/2}, && \text{with } \mathbf{z}_{k*} = \mathbf{z}_l,\\
\mathbf{z}_k &= \mathrm{Opt}_B(\mathbf{H}_k,\mathbf{z}_l) = \mathbf{S}^0_k\mathbf{z}_{k*}((\mathbf{S}^0_k\mathbf{z}_{k*})'\mathbf{S}^0_k\mathbf{z}_{k*})^{-1/2}, && \text{with } \mathbf{z}_{k*} = \mathbf{z}_l,\\
\mathbf{z}_k &= \mathrm{Opt}_C(\mathbf{H}_k,\mathbf{z}_j,\ldots,\mathbf{z}_l) = \mathbf{S}^1_k\mathbf{z}_{k*}((\mathbf{S}^1_k\mathbf{z}_{k*})'\mathbf{S}^1_k\mathbf{z}_{k*})^{-1/2}, && \text{with } \mathbf{z}_{k*} = (\mathbf{z}_j,\ldots,\mathbf{z}_l)\mathbf{1},\\
\mathbf{z}_k &= \mathrm{Opt}_D(\mathbf{H}_k,\mathbf{z}_j,\ldots,\mathbf{z}_l) = \mathbf{S}^0_k\mathbf{z}_{k*}((\mathbf{S}^0_k\mathbf{z}_{k*})'\mathbf{S}^0_k\mathbf{z}_{k*})^{-1/2}, && \text{with } \mathbf{z}_{k*} = (\mathbf{z}_j,\ldots,\mathbf{z}_l)\mathbf{1},\\
\mathbf{z}_k &= \mathrm{Opt}_E(\mathbf{H}_k,\mathbf{z}_j,\ldots,\mathbf{z}_l)\\
&= \mathrm{Opt}_B(\mathbf{H}_k,\mathbf{z}_w), && \text{with } \mathbf{z}_w = \mathrm{Opt}_A(\mathbf{H}_w,\mathbf{z}_k^{\mathrm{old}}), \text{ with } \mathbf{H}_w = (\mathbf{z}_j,\ldots,\mathbf{z}_l),\\
\mathbf{z}_k &= \mathrm{Opt}_F(\mathbf{H}_k,\mathbf{z}_j,\ldots,\mathbf{z}_l)\\
&= \mathrm{Opt}_B(\mathbf{H}_k,\mathbf{z}_w), && \text{with } \mathbf{z}_w = \mathrm{Opt}_B(\mathbf{H}_w,\mathbf{z}_k^{\mathrm{old}}), \text{ with } \mathbf{H}_w = (\mathbf{z}_j,\ldots,\mathbf{z}_l).
\end{aligned}
$$

In fact Bookstein defines for mode C and D not the sums $(\mathbf{z}_j,\ldots,\mathbf{z}_l)\mathbf{1}$, but the means $\mathbf{z}_{k*} = (\mathbf{1}'\mathbf{1})^{-1}(\mathbf{z}_j,\ldots,\mathbf{z}_l)\mathbf{1}$. Due to the normalization of $\mathbf{z}_k$ this makes no difference. For Opt_E and Opt_F an ancillary set of variates is defined as if they are manifest variables $\mathbf{H}_w$. It is clear that all operators can be brought in the general format of (5.35) and (5.40) with a proper definition of $\mathbf{W}_{\mathrm{Design}}$.

We give an example with three sets to show how the mode C to F operators can be incorporated in the bPLS fit function (5.40). The PLS path diagram in figure 5.8 can for instance be fitted with three PLS command diagrams given in figure 5.9.A to 5.9.C, comprising respectively mode C, mode D and mode F.

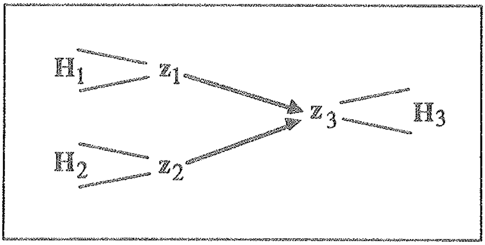
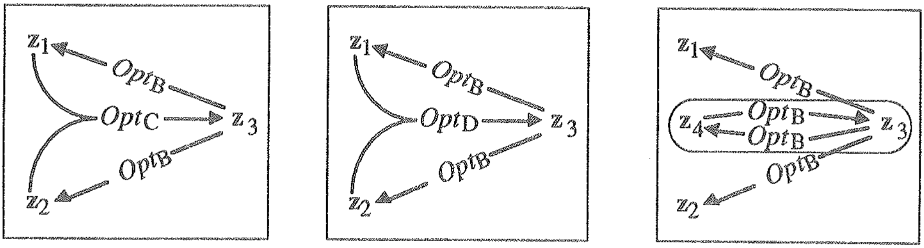


Figure 5.8 *PLS path diagram for three sets.*



A *PLS mode C operator* B *PLS mode D operator* C *PLS mode F operator*

Figure 5.9 *PLS command diagrams for three sets.*

The working area of mode F is outlined in figure 5.9.C. In figure 5.10.A to 5.10.C we give the LDC path diagrams with primary and secondary predictions. The LDC path configurations for respectively mode C, mode D and mode F operators are outlined.

5.10.A *Mode C*

5.10.B *Mode D*

5.10.C *Mode F*

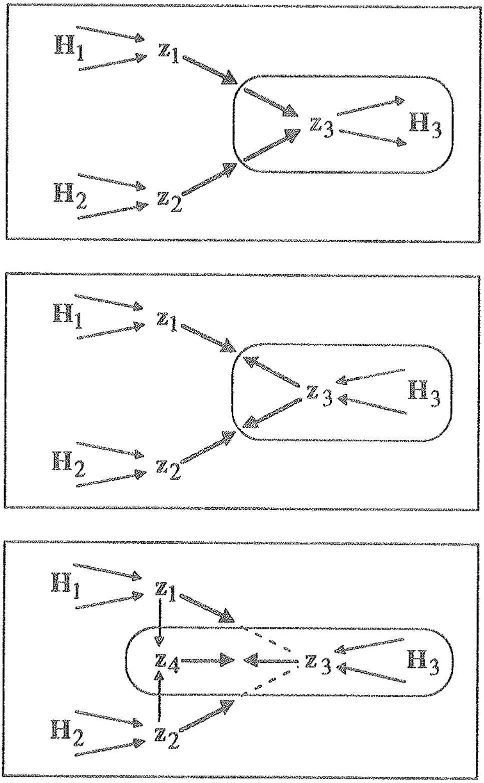


Figure 5.10 *LDC path diagrams for three sets.*

The mode C configuration outlined in figure 5.10.A is equal to figure 5.4, which gave an example of the LDC arrow configuration around a condensed variate with two pivot variates. In figure 5.9.C and 5.10.C we find an ancillary set of variates, $\mathbb{H}_4=(\mathbb{z}_1,\mathbb{z}_2)$, with regression variate $\mathbb{z}_4$. In mode E this would be a condensed variate $\mathbb{z}_4$ (see figure 5.11.C). The PLS path design matrix A_{Design} (5.30) for the path model in figure 5.8 is

$$A_{Design} \quad = \quad \begin{array}{|c|c|c|}\hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 1 & 1 & 0 \\ \hline \end{array} \quad , \qquad (5.47)$$

and the command design matrix W_{Design} for all diagrams in figure 5.9.A, 5.9.B, 5.10.A and 5.10.B is

$$W_{Design} \quad = \quad \begin{array}{|c|c|c|}\hline 0 & 0 & 1 \\ \hline 0 & 0 & 1 \\ \hline 1 & 1 & 0 \\ \hline \end{array} \quad . \qquad (5.48)$$

For figure 5.10.A the corresponding directed correlations matrix $\mathbb{R}$ with the desired design is

$$\mathbb{R}*W_{Design} \quad = \quad \begin{array}{|c|c|c|}\hline 0 & 0 & \mathbb{z}_1'\mathbb{z}_3 \\ \hline 0 & 0 & \mathbb{z}_2'\mathbb{z}_3 \\ \hline \mathbb{z}_3'\mathbb{z}_1 & \mathbb{z}_3'\mathbb{z}_2 & 0 \\ \hline \end{array} \quad . \qquad (5.49)$$

For figure 5.10.B the corresponding matrix is

$$\mathbb{R}*W_{Design} \quad = \quad \begin{array}{|c|c|c|}\hline 0 & 0 & \mathbb{z}_1'\mathbb{z}_3 \\ \hline 0 & 0 & \mathbb{z}_2'\mathbb{z}_3 \\ \hline \mathbb{z}_3'\mathbb{z}_1 & \mathbb{z}_3'\mathbb{z}_2 & 0 \\ \hline \end{array} \quad . \qquad (5.50)$$

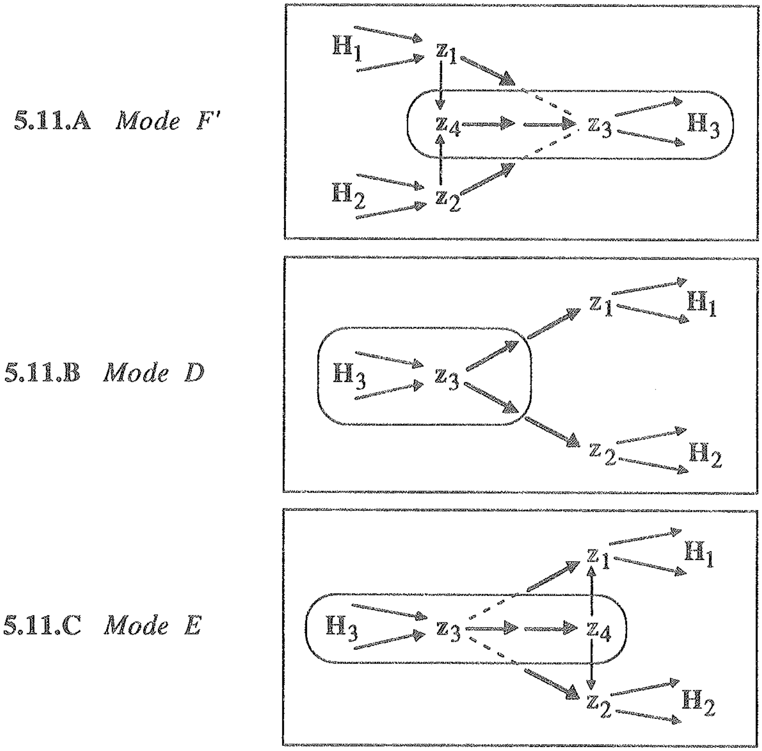
The command design matrix W_{Design} for the diagrams in figure 5.9.C and 5.10.C is

$$W_{Design} \quad = \quad \begin{array}{|c|c|c|c|}\hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 0 \\ \hline \end{array} \quad . \qquad (5.51)$$

and the corresponding directed correlations matrix $\mathbb{R}$ with the desired design for figure 5.10.C is

$$\mathbb{R} * \mathrm{W}_{\mathrm{Design}} = \begin{array}{|c|c|c|c|} \hline 0 & 0 & z_1'z_3 & 0 \\ \hline 0 & 0 & z_2'z_3 & 0 \\ \hline 0 & 0 & 0 & z_3'z_4 \\ \hline 0 & 0 & z_4'z_3 & 0 \\ \hline \end{array} \quad . \qquad (5.52)$$

At convergence of the bPLS solution z_3 and z_4 will be the canonical variates of a one dimensional CCA solution (5.44) with canonical correlation $z_3'z_4$. By studying the arrow diagram of figure 5.10.C we could for instance decide that our LDC path model would be better approximated by changing the primary prediction of z_4 by z_3 into a secondary prediction of z_3 by z_4. In this way we can make a mode F' as drawn in figure 5.11.A.



5.11.A *Mode F'*

5.11.B *Mode D*

5.11.C *Mode E*

Figure 5.11 *LDC path diagrams for three sets, continuation.*

The corresponding directed correlations matrix $\mathbb{R}$ with the desired design would be

$$\mathbb{R}{*}\mathrm{W}_{\mathrm{Design}} = \begin{array}{|c|c|c|c|}\hline 0 & 0 & z_1'z_3 & 0 \\ \hline 0 & 0 & z_2'z_3 & 0 \\ \hline 0 & 0 & 0 & z_3'z_4 \\ \hline 0 & 0 & z_4'z_3 & 0 \\ \hline \end{array} \quad . \qquad (5.53)$$

To demonstrate the properties of the optimization operator Opt_E of table 5.1, we reverse the arrows in the PLS path diagram in figure 5.8. Now z_3 is predicting both z_1 and z_2. In this case usually Opt_D or Opt_E are included in the command diagram. Opt_D is already incorporated in figure 5.9.B and 5.10.B. An intelligible alternative for the LDC configuration in figure 5.10.B could be to change the combination of mode D with the primary prediction of z_3 by z_1 and z_2 into mode D with the secondary prediction of z_1 and z_2 by z_3. This alternative is illustrated in figure 5.11.B. In figure 5.11.C Opt_E is incorporated in a LDC diagram. The corresponding directed correlations matrix $\mathbb{R}$ with the desired design is

$$\mathbb{R}{*}\mathrm{W}_{\mathrm{Design}} = \begin{array}{|c|c|c|c|}\hline 0 & 0 & z_1'z_3 & 0 \\ \hline 0 & 0 & z_2'z_3 & 0 \\ \hline 0 & 0 & 0 & z_3'z_4 \\ \hline 0 & 0 & z_4'z_3 & 0 \\ \hline \end{array} \quad . \qquad (5.54)$$

At convergence of the bPLS solution variate 3 and 4 will be equal to a one dimensional solution of the Principal Predictor model (5.45), where z_3 is canonical variate of the predictor set and z_4 the variate of the criteria.

5.4.5 Latent variable path modelling: extended PLS method

In this section the extended PLS method proposed by Lohmöller (1989) is brought within the LDC framework. Lohmöller designates the methods presented with the general name 'Latent Variable Path' methods (LVP methods).

The extensions of LVP modelling compared to basic PLS path modelling are:

- The relaxation of the basic PLS restriction that the variables must be partitioned into non-*overlapping sets* of variables.

- The possibility to compute methods with more than one dimension for each set, with *three different orthogonality restrictions* on the variates. These restrictions are different from the common deflation procedure used in basic PLS.
- The addition of two *extra weight functions* for the proportional function weights.

As for the *overlapping sets* of variables in LVP modelling, we remark that this overlap is also possible with LDC path models.

As for the *three different orthogonality restrictions*, this option can be realized in LDC path models by copying whole sets of variables and imposing orthogonality restrictions on the variates. For the computation of this orthogonality pattern according to an orthogonality design specified by the user, we can use the 'Pattor' rotation procedure developed by Lohmöller (1989, page 43). Lohmöller remarks that this patterned orthogonalization can interfere with the main PLS procedure by destroying in each iteration cycle the improvement made by the main procedure and thus blocking the convergence. Maybe this interference is due to the improper use of the Power Method in the basic PLS algorithm and will therefore not occur in the LDC algorithm.

The last LVP extension simply implies the specification of the *extra weight functions* in LDC. We repeat that in PLS only proportional function weights are applied, described by multivariate weight functions. The proportional function weights are combined with several weight functions into inner weighting modes (see section 5.3.3). In the basic PLS method described by Wold only one weight function (5.39) and therefore one weighting mode is used for $\mathbb{W}$. In the Extended PLS method Lohmöller (1989, page 42) adds two other weight functions and discerns three 'inner weighting modes' or 'weighting schemes'. These weighting schemes are the *path* weighting scheme, the *centroid* weighting scheme, and the *factor* weighting scheme. The centroid weighting scheme uses the basic PLS weight function (5.39). The factor weighting scheme uses the following weight function

$$\mathbb{F}(\mathbb{t}) = W_{\text{Design}} * \mathbb{R}. \tag{5.55}$$

The path weighting scheme uses in fact a combination of two weight functions. These are weight function (5.55) and the following weight function

$$\mathbb{F}(t) = W_{\mathrm{Design}}*\mathbb{A}, \tag{5.56}$$

with the regression weight A according to (5.30). For the path weighting scheme we have the weight function

$$\mathbb{F}(t) = W_{\mathrm{Design}(\mathbb{R})}*\mathbb{R} + W_{\mathrm{Design}(\mathbb{A})}*\mathbb{A}, \tag{5.57}$$

with $\mathrm{WDesign}(\mathbb{R}) + \mathrm{WDesign}(\mathbb{A}) = \mathrm{WDesign}$,

where $W_{\mathrm{Design}(\mathbb{R})}$, denotes the design for elements weighted according to (5.55),

and $W_{\mathrm{Design}(\mathbb{A})}$, denotes the design for elements weighted according to (5.56).

In PLS terminology the weights of the MVA weight functions (5.39), (5.55), (5.56) and (5.57) are respectively based on the centroid, the principal component, the multiple regression and the MIMIC variable (Lohmöller 1989, page 40). The MIMIC method is described in chapter 2.

Like (5.39) for basic PLS all additional defined weight functions for LVP modelling result in proportional function weights W for LDC, that are normalized according to (5.29). In LVP modelling the weights are always defined as $W = \mathbb{F}(t)$. Although the optimal solution for the variates z_k is the same for LDC and LVP modelling, this definition suggests that the weights are real function weights and obscures the fact that they are proportional. With a simple two sets path model one can easily verify that the two sets LVP solution cannot be influenced by the choice of the weight function, whereas this would definitely be the case if the weights W are real function weights. Furthermore it has been found that the choice of inner weighting modes has only little influence on the results when the model is a realistic one (Noonan & Wold, 1982). These results are not so surprising if one realizes that PLS uses proportional function weights. Real function values might differentiate the results much more with respect to the different weighting modes.

5.4.6 Three-way Consensus PLS

In Martens & Martens (1986), the family of PLS mode A methods is classified in two main groups to suit different analytical situations. These groups are *predictive* and *correlative* PLS. Among the predictive PLS algorithms they mention PLS2 (see formula (5.46) in section 5.4.4 and further) and PLS1 (see section 5.4.7). Concerning the correlative PLS methods they only discuss three-way Consensus PLS (CPLS). We will first give the estimated model and algorithm of CPLS proposed by Geladi, Martens, Martens, Kalvenes & Esbensen (1988), translated into our notation and normalization. The LDC formulation is possible by defining an appropriate weight function. Furthermore we give two alternative 'variance accounted for' criteria that are fitted by the consensus PLS algorithm.

Models used:

Consensus model: The m_k manifest variables $\mathbf{H}_k$ are assumed to be generated as a linear function of the consensus variate $\mathbf{x}$ and the residual variables $\mathbf{E_c}$,

$$\mathbf{H} = \mathbf{x}\mathbf{c}' + \mathbf{E}_\mathbf{c}, \tag{5.58}$$

where $\mathbf{c}' = (\mathbf{c}_1',..,\mathbf{c}_k',..,\mathbf{c}_K')$ denotes a vector with $\sum_k m_k$ loadings,
$\mathbf{c}_k$ denotes a vector with m_k loadings for set k,
$\mathbf{E_c}$ denote residual variables for all sets.

Model for each set: The variables $\mathbf{H}_k$ are assumed to be generated as a linear function of variate $\mathbf{z}_k$ and the residual variables $\mathbf{E}_k$,

$$\mathbf{H}_k = \mathbb{z}_k\mathbf{c}_k' + \mathbf{E}_k, \tag{5.59}$$

where $\mathbb{z}_k$ denotes a non-normalized version of variate $\mathbf{z}_k$,
$\mathbf{E}_k$ denote residual variables for set k.

Contribution of each variate $\mathbf{z}_k$ to consensus scores $\mathbf{x}$,

$$\mathbf{x} = \mathbb{Z}\mathbf{a}, \tag{5.60}$$

where $\mathbb{Z} = (\mathbb{z}_1,..,\mathbb{z}_k,..,\mathbb{z}_K)$,
and $\mathbf{a}$ denote consensus block weights by regression of $\mathbb{Z}$ on the consensus scores $\mathbf{x}$.

The loadings $\mathbf{c}_k$ for (5.58) and (5.59) are equal, as well as the non-normalized variates $\mathbb{z}_k$ in (5.59) and the columns of $\mathbb{Z}$ in (5.60). With our normalization $\mathbf{x}'\mathbf{x}=1$ and $\mathbf{z}_k'\mathbf{z}_k=1\ \forall k$, the CPLS iterative algorithm is as follows:

Select some starting values for consensus scores $\mathbf{x}$. For each factor, perform Steps 1 to 7:

Step 1. Solve for $\mathbf{c}_k$ in (5.58)

Compute $\mathbf{c}_k = \mathbb{H}_k'\mathbf{x}$, $\forall k$

Step 2. Convergence ? go to Step 7

Step 3. Solve for $\mathbb{z}_k$ in (5.59)

Compute $\mathbb{z}_k = \mathbb{H}_k\mathbf{c}_k(\mathbf{c}_k'\mathbf{c}_k)^{-1}$. $\forall k$

Step 4. Solve for $\mathbf{a}$

$\mathbb{Z} = \mathbf{x}\mathbf{a}' + \mathbb{E}_{\mathbf{a}}$,

Step 5. $\mathbf{x} = \mathbb{Z}\mathbf{a}((\mathbb{Z}\mathbf{a})'\mathbb{Z}\mathbf{a})^{-1/2}$. (5.60)

Step 6. Return to Step 1

Step 7. Residuals from (5.58)

$\mathbb{E}_{\mathbf{c}} = \mathbb{H} - \mathbf{x}\mathbf{c}'$,

Use the residuals $\mathbb{E}_{\mathbf{c}}$ as $\mathbb{H}$ in the next consensus dimension.

Step 1 to 3 of the CPLS algorithm can by substitution be shortened to one step:

$$\text{Compute} \quad \mathbb{z}_k = \mathbf{S}_k\mathbf{x}(\mathbf{x}'\mathbf{S}_k\mathbf{x})^{-1}, \qquad \forall k \quad (5.61)$$

After substitution of (5.61) in Step 4 this implies for the consensus block weights $\mathbf{a}$ that $\mathbf{a}=\mathbf{u}$, a vector with elements 1. Now the whole CPLS algorithm can be reduced to:

$$\text{Compute} \quad \mathbf{x} = f\sum_{k=1}^{K}\mathbf{S}_k\mathbf{x}(\mathbf{x}'\mathbf{S}_k\mathbf{x})^{-1} = \mathbf{Z}\mathbf{w}((\mathbf{Z}\mathbf{w})'\mathbf{Z}\mathbf{w})^{-1/2}, \qquad (5.62)$$

where $f = ((\Sigma_k\mathbf{S}_k\mathbf{x}(\mathbf{x}'\mathbf{S}_k\mathbf{x})^{-1})'\Sigma_k\mathbf{S}_k\mathbf{x}(\mathbf{x}'\mathbf{S}_k\mathbf{x})^{-1})^{-1/2}$,

$\mathbf{Z} = (\mathbf{z}_1,\ldots,\mathbf{z}_k,\ldots,\mathbf{z}_K)$ with $\mathbf{z}_k = \mathbf{S}_k\mathbf{x}(\mathbf{x}'\mathbf{S}_k\mathbf{S}_k\mathbf{x})^{-1/2}$,

and $\mathbf{w}' = (w_1,\ldots,w_k,\ldots,w_K)$ with $w_k = (\mathbf{x}'\mathbf{z}_k)^{-1}$.

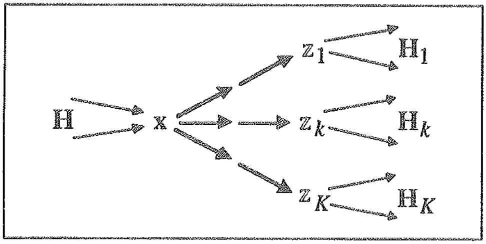


Figure 5.12 *LDC path diagram for Consensus PLS.*

The CPLS algorithm (5.62) is also obtained if we fit the LDC path model in figure 5.12 by substituting proportional function weights (5.29) with the following weight function

$$\mathbb{F}(\mathbf{t}) = \mathbb{W}_{\text{Design}} * \mathbb{R}_{\text{Recip}}. \tag{5.63}$$

where $\mathbb{X}_{\text{Recip}}$ denotes a matrix with the reciprocal values of $\mathbb{X}$, which implies that $\mathbb{X} * \mathbb{X}_{\text{Recip}}$ is a matrix with elements 1.

For figure 5.12 the corresponding directed correlations matrix $\mathbb{R}$ with the desired design is

$$\mathbb{R} * \mathbb{W}_{\text{Design}} = \begin{array}{|c|c|c|c|}\hline 0 & \mathbf{x}'\mathbf{z}_1 & \mathbf{x}'\mathbf{z}_k & \mathbf{x}'\mathbf{z}_K \\ \hline \mathbf{z}_1'\mathbf{x} & 0 & 0 & 0 \\ \hline \mathbf{z}_k'\mathbf{x} & 0 & 0 & 0 \\ \hline \mathbf{z}_K'\mathbf{x} & 0 & 0 & 0 \\ \hline \end{array}\ , \tag{5.64}$$

where $\mathbf{x} = \mathbb{H}_0\mathbf{t}_0 = \mathbb{H}^{\alpha}\mathbf{t}_0$ denotes a regression variate with α_0=0, so $\mathbb{H}^0$=$\mathbb{PQ}'$,

and $\mathbf{z}_k = \mathbb{H}_k\mathbf{t}_k = \mathbb{H}^{\alpha}_k\mathbf{t}_k$, denote condensed variates with α_k=1, $\forall k$

with $\mathbb{H} = (\mathbb{H}_0,\ldots,\mathbb{H}_1,\ldots,\mathbb{H}_k,\ldots,\mathbb{H}_K)$
$= (\mathbb{H}^0,\ldots,\mathbb{H}^1_1,\ldots,\mathbb{H}^1_k,\ldots,\mathbb{H}^1_K) = (\mathbb{H}^0,\mathbb{H}^1) = (\mathbb{PQ}',\mathbb{P\Phi Q}')$.

If we take $\mathbb{R} * \mathbb{W}_{\text{Design}}$ equal to (5.64), with K equal to the total number of condensed variates, the LDC formulation of the CPLS fit function is elaborated to

$$\text{CPLS:}\quad \text{Fit}(\mathbf{x}) = \frac{2K(K+1)}{\mathbf{t}'\mathbf{t}(K^{-1}\sum_{k=1}^{K}(\mathbf{x}'\mathbf{z}_k)^{-2})^{1/2}}$$

$$= \frac{2K(K+1)}{(1+\sum_{k=1}^{K}\frac{(\mathrm{x}'\mathrm{z}_k)^2}{\mathrm{x}'\mathrm{S}_k\mathrm{x}})(K^{-1}\sum_{k=1}^{K}(\mathrm{x}'\mathrm{z}_k)^{-2})^{1/2}}, \tag{5.65}$$

where substitution of (5.63) and (5.29) in (5.10) has given this rather complicated function, despite the simplification due to $\mathbb{R}*\mathbb{F}(\mathfrak{t})=\mathbb{R}*\mathbb{W}_{\mathrm{Design}}*\mathbb{R}_{\mathrm{Recip}}=\mathbb{W}_{\mathrm{Design}}$.

Because all feasible $\mathbb{W}$'s are sign-similar and $\mathbb{W}_{\mathrm{Sign}}=\mathbb{W}_{\mathrm{Design}}*\mathbb{R}_{\mathrm{Sign}}$, ($\mathbb{R}_{\mathrm{Sign}}=(\mathbb{R}_{\mathrm{Recip}})_{\mathrm{Sign}}$, see section 5.4.3), we elaborate an algorithm for CPLS (5.65) along the lines of the bPLS algorithm. The bPLS algorithm can for this purpose be defined by alternating between (5.35) and inner weights (5.36). For the CPLS algorithm we only have to replace (5.36) by (5.63). After substitution of the appropriate matrices in (5.35) we obtain the computation of (5.62) parallel with the computation of $\mathrm{z}_k = \mathrm{S}_k\mathrm{x}(\mathrm{x}'\mathrm{S}_k\mathrm{S}_k\mathrm{x})^{-1/2}$, $\forall k$. The possible solutions with this bPLS like algorithm are two equal global maxima, which only differ with respect to the sign of the variates x and $\mathrm{z}_1,\ldots,\mathrm{z}_k,\ldots,\mathrm{z}_K$ as a group. Therefore the CPLS solution in LDC format and the solution of the reformulated original CPLS algorithm (5.62) are equal. Only sign reversal might occur.

After substitution of the appropriate matrices in the (relatively simple) LDC fit function we obtain a rather complicated CPLS fit function (5.65). We give two other less complicated fit functions, CPLS_2 and CPLS_3, that lead to the same CPLS algorithm and the same solution. The second fit function for consensus PLS uses proportional weights w_k with weight function $\mathfrak{f}(\mathrm{x})$

$$\mathrm{CPLS}_2\text{:}\quad \mathrm{Fit}(\mathrm{x}) = \sum_{k=1}^{K}\mathrm{x}'\mathrm{S}_k\mathrm{x}\,w_k, \tag{5.66}$$

$$\begin{array}{lll} \text{where} & (w_1,\ldots,w_k,\ldots,w_K) & = \mathrm{w}', \\ \text{with} & \mathrm{w} & = \mathfrak{f}(\mathrm{x})\ (\mathfrak{f}(\mathrm{x})'\mathfrak{f}(\mathrm{x}))^{-1/2}, \\ & \mathfrak{f}(\mathrm{x})' & = f_1(\mathrm{x}),\ldots f_k(\mathrm{x}),\ldots,f_K(\mathrm{x}), \\ & f_k(\mathrm{x}) & = (\mathrm{x}'\mathrm{S}_k\mathrm{x})^{-1}. \end{array}$$

The third fit function for consensus PLS is

$$\text{CPLS}_3\text{:}\quad \text{Fit}(\mathbf{x}) = \sum_{k=1}^{K} (\mathbf{x}'\mathbf{S}_k\mathbf{x})^{\varepsilon}, \tag{5.67}$$

with $\varepsilon > 0$ and $\varepsilon \to 0$.

Maximization of (5.66) and (5.67) leads to the same algorithm for CPLS as found in (5.62). We will not elaborate on this in detail. Finally we remark that Geladi, Martens, Martens, Kalvenes & Esbensen (1988) suggest that alternative CPLS algorithms can be envisioned. These algorithms can also be brought in the LDC framework, but they will not be discussed here.

5.4.7 PLS1 regression and extensions

The general PLS1 mode A algorithm is actually a special case of the basic PLS method presented in section 5.4.1 and 5.4.2. Therefore it is also a special case of the LDC method, as we showed in section 5.4.3. Nevertheless we will discuss the LDC formulation of PLS1 regression in detail. It gives the opportunity to incorporate a PLS1 extension of Lorber, Wangen & Kowalski (1987) in the LDC framework. We also indicate relations with Continuum Regression proposed by Stone & Brooks (1990).

The LDC path diagram for PLS1 regression with a rank p decomposition of the predictor set $\mathbb{H}_1$, is given in figure 5.13.

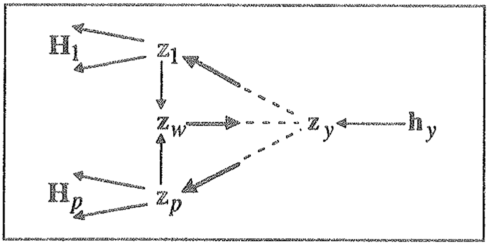


Figure 5.13 *LDC path diagram for PLS1 regression.*

In this figure we find an ancillary set of variates, $\mathbb{H}_w = (\mathbb{z}_1,\ldots,\mathbb{z}_s,\ldots,\mathbb{z}_p)$, with regression variate $\mathbb{z}_w$. Ancillary sets have been introduced in section 5.4.4. The dependent unit normalized variable $\mathbb{h}_y$ is conceived as a set with only one variable. Therefore the regression variate $\mathbb{z}_y$ is always equal to this variable, $\mathbb{z}_y=\mathbb{h}_y$. The predictor set $\mathbb{H}_1$ produces by deflation orthonormal condensed variates $(\mathbb{z}_1,\ldots,\mathbb{z}_s,\ldots,\mathbb{z}_p)$ on the analogy of (5.4)

$$\begin{aligned} \mathbb{H}_s &= \mathbb{H}_1 && \text{for } s = 1, \\ \mathbb{H}_s &= (\mathbb{I}-\mathbb{z}_{s-1}\mathbb{z}_{s-1}')\mathbb{H}_{s-1} && \text{for } s = 2,\ldots,p. \end{aligned} \qquad (5.68)$$

For figure 5.13 the corresponding PLS1 directed correlations matrix $\mathbb{R}$ with the desired design is

$$\mathbb{R}*\mathbb{W}_{\text{Design}} \quad = \quad \begin{array}{|c|c|c|c|} \hline 0 & 0 & \mathbb{z}_1'\mathbb{z}_y & 0 \\ \hline 0 & 0 & \mathbb{z}_p'\mathbb{z}_y & 0 \\ \hline 0 & 0 & 0 & \mathbb{z}_y'\mathbb{z}_w \\ \hline 0 & 0 & \mathbb{z}_w'\mathbb{z}_y & 0 \\ \hline \end{array} \quad , \qquad (5.69)$$

where the sequence of the rows and columns of $\mathbb{R}$ is $\mathbb{z}_1,\ldots,\mathbb{z}_p$, $\mathbb{z}_y$ and $\mathbb{z}_w$.

The correlation $\mathbb{z}_y'\mathbb{z}_w$ in position (3,4) is added in this design in order to satisfy the restriction of section 5.3.5 that all pivot variates have to be condensed (or regression) variates. This implies that $(\mathbb{W}_{\text{Design}}\mathbb{1})_k\neq 0$, $\forall k$. Instead we could have added $\mathbb{z}_1'\mathbb{z}_w$ in position (3,1) or other correlations in row 3. We also could have defined $\mathbb{z}_y$ to be a condensed variate. All these options do not change the final solution, because $\mathbb{h}_y$ contains only one variable and therefore $\mathbb{z}_s'\mathbb{z}_y$ is fixed $\forall s$. If we take $\mathbb{R}*\mathbb{W}_{\text{Design}}$ equal to (5.69), with K equal to the total number of involved variates, the LDC formulation of the PLS1 fit function is elaborated to

$$\text{PLS1:} \quad \text{Fit}(\mathbb{t}) = \frac{2(\mathbb{z}_w'\mathbb{z}_y)_{\text{Abs}} + \sum_{s=1}^{p} (\mathbb{z}_s'\mathbb{z}_y)_{\text{Abs}}}{\mathbb{t}'\mathbb{t}K^{-1}}. \qquad (5.70)$$

Because all feasible W's are sign-similar and $\mathbb{W}_{\text{Sign}}=\mathbb{W}_{\text{Design}}*\mathbb{R}_{\text{Sign}}$, see section 5.4.3), we will always find a maximum for (5.70) with the bPLS algorithm. The global maximum is reached after one iteration, starting with $\mathbb{z}_y=\mathbb{h}_y$. The PLS1 algorithm resulting from (5.70) is:

Start with $\mathbb{z}_y=\mathbb{h}_y$

Step 1. Compute $\mathbf{z}_s = \mathbf{S}_s\mathbf{z}_y(\mathbf{z}_y'\mathbf{S}_s\mathbf{S}_s\mathbf{z}_y)^{-1/2}$, $\forall s$
with deflation according to (5.68).

Step 2. Compute $\mathbf{z}_w = \mathbf{S}_w\mathbf{z}_y(\mathbf{z}_y'\mathbf{S}_w\mathbf{S}_w\mathbf{z}_y)^{-1/2}$,
with orthonormal $\mathbf{H}_w = (\mathbf{z}_1,\ldots,\mathbf{z}_s,\ldots,\mathbf{z}_p)$.

The optimal rank of p is usually assessed via some data-based statistical procedure. One can for instance use cross-validation to calculate a predicted residual error sum of squares (PRESS).

The extension of Lorber, Wangen & Kowalski (1987) can now be defined through a minor adaptation of the LDC fit function. In LDC the constant α_k is restricted to have two values, α_k=1 for condensed variates and α_k=0 for regression variates. Lorber c.s. restrict the α_1 of the predictor variables $\mathbf{H}_1^{\alpha}$ to be in a continuum ranging from 0 to ∞. (In their notation α_1 is n.) They show that for α_1=0, we have ordinary least squares regression, for α_1=1, we have PLS1 regression and for α_1=∞, we have Principal Component Regression (PCR). In an example the optimal combination of p and α_1 is assessed with PRESS.

In Continuum Regression proposed by Stone & Brooks (1990) a similar idea is elaborated. They also describe a continuum from OLS regression, PLS1 regression to PCR for respectively γ=0, 1 and ∞, see page 243 of Stone & Brooks, 1990. A selection function is maximized to find the successive variates of $\mathbf{H}_w = (\mathbf{z}_1,\ldots,\mathbf{z}_s,\ldots,\mathbf{z}_p)$ as an alternative for Step 1 in the previous PLS1 algorithm. These variates are *not* condensed variates as in PLS1. In our notation and normalization the direction of the variates $\mathbf{z}_s$ is found by maximizing

$$\text{CR:} \qquad \text{Fit}(\mathbf{t}_s) = \frac{(\mathbf{z}_s'\mathbf{z}_y)^2}{(\mathbf{t}_s'\mathbf{t}_s)^{\gamma}}, \tag{5.71}$$

where $\mathbf{z}_s = \mathbf{H}_s\mathbf{t}_s$, $\forall s$
with $\mathbf{z}_s'\mathbf{z}_s = 1$.

Successive variates of $\mathbf{H}_w$ are found by deflating the predictor set $\mathbf{H}_1$ according to (5.68), with $\mathbf{z}_s$ replaced by $\mathbf{z}_s$. In a final step one has to perform Step 2 of the PLS1 algorithm, with $\mathbf{z}_s$ replaced by $\mathbf{z}_s$. In the normalization of Stone & Brooks (1990)

$\mathbf{t}_s'\mathbf{t}_s=1$, and selection function (5.71) is defined as $T(\mathbf{t}_s)=(\mathbf{z}_s'\mathbf{z}_y)^2(\mathbf{t}_s'\mathbf{H}_s'\mathbf{H}_s\mathbf{t}_s)^{\gamma-1}$. Originally $T(\mathbf{t}_s)$ is multiplied with the constant $(\mathbf{z}_y'\mathbf{z}_y)^{1/2}$, but this term can be left out without loss of generality, because the selection function (5.71) is maximized. The equality $T(\mathbf{t}_s)=CR(\mathbf{t}_s)$ can be shown by making the explicit normalization of $\mathbf{t}_s$ in $T(\mathbf{t}_s)$ implicit, followed by shifting to the explicit normalization $\mathbf{z}_s'\mathbf{z}_s=1$.

If we try to describe the complete CR method with the LDC fit function (5.10), we see that CR cannot exactly be formulated as a LDC method. Nevertheless the selection function of CR (5.71) is also a product of a global correlation fit function and a local reciprocal PCA fit function, LRPCV (5.3). The constant γ regulates the relative importance of the correlation and the PCA part. The CR selection function can therefore be classified as a Lifted Correlation fit function.

In summary, we have presented two continuum regression methods, continuum PLS1 proposed by Lorber, Wangen & Kowalski (1987) and CR proposed by Stone & Brooks (1990). If the continuum parameters α_1 and γ are equal to 0, 1 and ∞, both methods produce respectively a solution equal to OLS regression, PLS1 regression and PCR. Continuum PLS1 can be fitted with a LDC fit function with relaxed α. CR is closely related to LDC with respect to the CR selection function.

5.4.8 *Reflected variance methods and PLS2*

The reflected variance methods RCA and RDA of chapter 4 are fitted with a LDC path model. First we will give an exposition of the LDC formulation of these reflected variance methods and subsequently we will show that the corresponding LDC algorithm leads to the same solutions. Some minor changes to the LDC reflected variance path models produce an interesting alternative for RCA, RDA and for PLS2.

In LDC notation the one dimensional RCA fit function $RCA(\mathbb{X}) = \mathrm{tr}\ \mathbb{X}'\mathbb{PUSUPX}$, (4.3), is given by

$$\text{RCA:}\qquad Fit(\mathbf{z}_1) = \mathbf{z}_1'\mathbf{S}_2^0\mathbf{S}_1\mathbf{S}_2^0\mathbf{z}_1, \qquad (5.72)$$

where	$\mathbf{H}_1$	denote the predictor variables $\mathbf{H}$ in (4.3),
with	$\mathbf{H}_1\mathbf{H}_1' = \mathbf{S}_1$	$= \mathbf{S}$, in (4.3),

$\mathbb{z}_1 = \mathbb{H}_1^0\mathbb{t}_1$ denotes the latent variable $\mathbb{x}$,

with $\mathbb{H}_1^0\mathbb{t}_1 = \mathbb{S}_1^0\mathbb{H}_1^0\mathbb{t}_1$ $= \mathbb{P}\mathbb{x}$, in (4.3),

and $\mathbb{z}_1'\mathbb{z}_1 = 1$,

$\mathbb{H}_2$ denote the external variables $\mathbb{H}_\mathbb{U}$ in (4.3),

$\mathbb{H}_2^0$ denotes the orthonormal mirror matrix $\mathbb{U}$,

with $\mathbb{H}_2^0\mathbb{H}_2^{0\prime} = \mathbb{S}_2^0$ $= \mathbb{U}$, in (4.3).

For the RDA fit function we can also use (5.72) by changing the definition of $\mathbb{S}_2^0 = \mathbb{U}$, into $\mathbb{S}_2^0 = \mathbb{G} = \mathbb{G}\mathbb{D}^{-1}\mathbb{G}'$, (4.8). With the Power Method we define an iterative algorithm for finding the optimal $\mathbb{z}_1$ in (5.72).

Starting with some arbitrary $\mathbb{z}_1$ iterate until convergence:

$$\text{Compute } \mathbb{z}_1 = f\mathbb{S}_1^0\mathbb{S}_2^0\mathbb{S}_1\mathbb{S}_2^0\mathbb{z}_1 = f\mathbb{S}_1^0\mathbb{S}_2^0\mathbb{S}_3\mathbb{S}_4^0\mathbb{z}_1, \tag{5.73}$$

where $f = ((\mathbb{S}_1^0\mathbb{S}_2^0\mathbb{S}_1\mathbb{S}_2^0\mathbb{z}_1)'\mathbb{S}_1^0\mathbb{S}_2^0\mathbb{S}_1\mathbb{S}_2^0\mathbb{z}_1)^{-1/2}$,

$\mathbb{H}_3 = \mathbb{H}_1$,

and $\mathbb{H}_4 = \mathbb{H}_2$.

The LDC counterpart of (5.72) we call RCA$_{\text{LDC}}$. As stated before RCA$_{\text{LDC}}$ automatically comprises RDA$_{\text{LDC}}$. Consistent with the notation in (5.73) with duplicate sets 3 and 4 we draw a LDC path diagram for RCA$_{\text{LDC}}$ in figure 5.14.A.

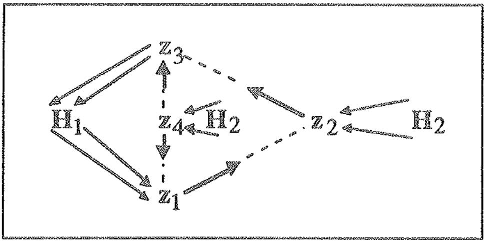


Figure 5.14.A *LDC path diagram for Reflected Component Analysis.*

The corresponding RCA$_{\text{LDC}}$ directed correlations matrix $\mathbb{R}$ with the desired design is

$$\mathbb{R}*\mathbb{W}_{\text{Design}} = \begin{array}{|c|c|c|c|} \hline 0 & \mathbb{z}_1'\mathbb{z}_2 & 0 & 0 \\ \hline 0 & 0 & \mathbb{z}_2'\mathbb{z}_3 & 0 \\ \hline 0 & 0 & 0 & \mathbb{z}_3'\mathbb{z}_4 \\ \hline \mathbb{z}_4'\mathbb{z}_1 & 0 & 0 & 0 \\ \hline \end{array} \;. \tag{5.74}$$

Substituting (5.74) and the bPLS weight function (5.39) in (5.10) and taking into account the proportionality restrictions for condensed and regression variates, we obtain the $\mathrm{RCA}_{\mathrm{LDC}}$ fit function

$$\mathrm{RCA}_{\mathrm{LDC}}\text{: Fit}(\mathbf{t}) = \frac{(\mathbf{z}_2'\mathrm{S}_1^0\mathbf{z}_2)^{1/2} + (\mathbf{z}_3'\mathrm{S}_2^0\mathbf{z}_3)^{1/2} + (\mathbf{z}_3'\mathbf{z}_4)_{\mathrm{Abs}} + (\mathbf{z}_1'\mathrm{S}_4^0\mathbf{z}_1)^{1/2}}{\mathbf{t}'\mathbf{t}K^{-1}}. \qquad (5.75)$$

We emphasize that $(\mathbf{z}_2'\mathrm{S}_1^0\mathbf{z}_2)^{1/2}$ is equal to $(\mathbf{z}_1'\mathbf{z}_2)_{\mathrm{Abs}}$ with *restriction*

$$\mathbf{z}_1 = \mathrm{S}_1^0\mathbf{z}_2(\mathbf{z}_2'\mathrm{S}_1^0\mathbf{z}_2)^{-1/2}.$$

If we omit this restriction and take $(\mathbf{z}_1'\mathbf{z}_2)_{\mathrm{Abs}}$ instead of $(\mathbf{z}_2'\mathrm{S}_1^0\mathbf{z}_2)^{1/2}$, we are fitting another path model. With the LDC algorithm we find a global maximum for (5.75) with optimal $\mathbf{t}$. Knowing the optimal $\mathbf{z}_1 = \mathbf{H}_1\mathbf{t}_1$ we can derive the optimal values for the other three variates with the updating equations of the bPLS algorithm. These equations are for (5.75)

$$\left\{\begin{array}{lcl} \mathbf{z}_4 & = & \mathrm{S}_4^0\mathbf{z}_1(\mathbf{z}_1'\mathrm{S}_4^0\mathbf{z}_1)^{-1/2}(\mathbf{z}_4'\mathbf{z}_1)_{\mathrm{Sign}} \\ \mathbf{z}_3 & = & \mathrm{S}_3\mathbf{z}_4(\mathbf{z}_4'\mathrm{S}_3\mathrm{S}_3\mathbf{z}_4)^{-1/2}(\mathbf{z}_3'\mathbf{z}_4)_{\mathrm{Sign}} \\ \mathbf{z}_2 & = & \mathrm{S}_2^0\mathbf{z}_3(\mathbf{z}_3'\mathrm{S}_2^0\mathbf{z}_3)^{-1/2}(\mathbf{z}_2'\mathbf{z}_3)_{\mathrm{Sign}} \\ \mathbf{z}_1 & = & \mathrm{S}_1^0\mathbf{z}_2(\mathbf{z}_2'\mathrm{S}_1^0\mathbf{z}_2)^{-1/2}(\mathbf{z}_1'\mathbf{z}_2)_{\mathrm{Sign}} \end{array}\right\} \qquad (5.76)$$

By subsequent substitution of all equations in (5.76) we have for $\mathbf{z}_1$ (5.73), but now with

$$f = \pm((\mathrm{S}_1^0\mathrm{S}_2^0\mathrm{S}_1\mathrm{S}_2^0\mathbf{z}_1)'\mathrm{S}_1^0\mathrm{S}_2^0\mathrm{S}_1\mathrm{S}_2^0\mathbf{z}_1)^{-1/2}.$$

This elaboration of the $\mathrm{RCA}_{\mathrm{LDC}}$ algorithm implies that maximization of $\mathrm{RCA}(\mathbf{z}_1)$ (5.72) and $\mathrm{RCA}_{\mathrm{LDC}}$ (5.75) leads to the same optimal solution for $\mathbf{z}_1$ apart from sign reversion.

An intelligible LDC alternative for RCA (and RDA) is to change the LDC path diagram for $\mathrm{RCA}_{\mathrm{LDC}}$ in figure 5.14.A into the path diagram in figure 5.14.B.

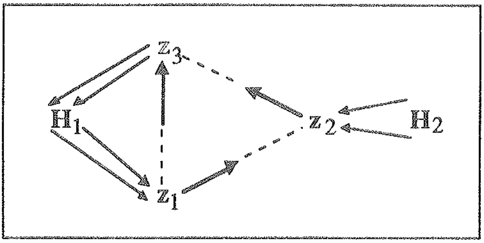


Figure 5.14.B *LDC path diagram for RCA^{sim}.*

In this way z_1 still predicts a weighted sum of the variables H_2 related to the variance structure of $H_1=H_3$. We call this simplified RCA method RCA^{sim}. If we additionally want the RCA^{sim} solution of z_1 to be related to the variance structure of H_2, we change the RCA^{sim} diagram in figure 5.14.B further into the path diagram in figure 5.14.C.

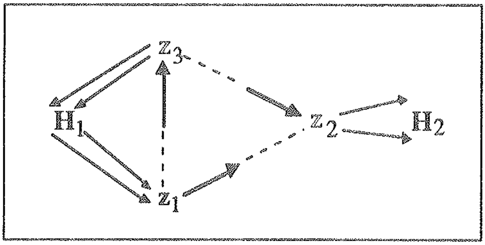


Figure 5.14.C *LDC path diagram for PLS2^{multi}.*

We call this method PLS2^{multi}. PLS2, because it is the goal of the PLS2 method (Manne, 1987) to predict with a predictor variable z_1 a weighted sum of dependent variables H_2 related to both the variance structure of H_2 and the variance structure of the independent variables H_1, (see figure 5.7.C and further in section 5.4.4). The superscript '^{multi}' we add, because z_1 can only be related to the variance of H_1 if H_2 has more than one variable and a rank higher than one. For the rank one case PLS2^{multi} is equal to ordinary least squares regression. We expect PLS2^{multi} to have better predictive properties than PLS2 in a multivariate setting, because the secondary prediction of z_2 in PLS2 is replaced by a primary prediction in PLS2^{multi}. Yet PLS2^{multi} remains stable for essential multivariate problems, because then it is also related to the variance structure of both sets. The PLS2^{multi} solution can easily be derived from the ordinary PLS2 solution. By subsequent substitution of all PLS2 and

PLS2$^{\text{multi}}$ algorithmic equations, as we did in (5.76) for RCA$_{\text{LDC}}$, we have respectively for $\mathbb{z}_1$ (PLS2)

$$\mathbb{z}_1 = f\,\mathbb{S}_1\mathbb{S}_2\mathbb{z}_1, \tag{5.77}$$

where $f = \pm((\mathbb{S}_1\mathbb{S}_2\mathbb{z}_1)'\mathbb{S}_1\mathbb{S}_2\mathbb{z}_1)^{-1/2}$.

and for $\mathbb{z}_1{}^{\text{multi}}$ (PLS2$^{\text{multi}}$)

$$\mathbb{z}_1{}^{\text{multi}} = f\,\mathbb{S}_1^0\mathbb{S}_2\mathbb{S}_3\mathbb{z}_1{}^{\text{multi}} = f\mathbb{S}_1^0\mathbb{S}_2\mathbb{S}_1\mathbb{z}_1{}^{\text{multi}}, \tag{5.78}$$

where $f = \pm((\mathbb{S}_1^0\mathbb{S}_2\mathbb{S}_1\mathbb{z}_1{}^{\text{multi}})'\mathbb{S}_1^0\mathbb{S}_2\mathbb{S}_1\mathbb{z}_1{}^{\text{multi}})^{-1/2}$.

The optimal $\mathbb{z}_1$ and $\mathbb{z}_1{}^{\text{multi}}$ are found by repeating (5.77) and (5.78) iteratively until convergence is reached. Because $\mathbb{S}_1\mathbb{S}_1^0{=}\mathbb{S}_1$ we know by combining (5.77) and (5.78) that $\mathbb{z}_1{=}\pm\mathbb{S}_1\mathbb{z}_1{}^{\text{multi}}((\mathbb{S}_1\mathbb{z}_1{}^{\text{multi}})'\mathbb{S}_1\mathbb{z}_1{}^{\text{multi}})^{-1/2}$. The PLS2$^{\text{multi}}$ solution is related to the optimal PLS2 solution by $\mathbb{z}_1{}^{\text{multi}}{=}\mathbb{S}_1^0\mathbb{S}_2\mathbb{z}_1((\mathbb{S}_1^0\mathbb{S}_2\mathbb{z}_1)'\mathbb{S}_1^0\mathbb{S}_2\mathbb{z}_1)^{-1/2}$.

5.4.9 Set Component Analysis and PLS Hierarchical Components

Last but not least in section 5.4 about relations of LDC with other methods, we discuss the SCA method of chapter 3 as an example of a DC path model with real function weights. Some LDC extensions of the SCA method are formulated, like the PLS Hierarchical Components.

The one dimensional SCA fit function (3.5) of chapter 3 translated into the terminology of this chapter is defined by the sum of the squared directed correlations between pivot variate $\mathbb{x}$ and condensed variates $\mathbb{z}_k$, where $\mathbb{x}$ is the pivot variate for all K sets

SCA: $\quad \text{Fit}(\mathbb{x}) = \mathbb{x}'\mathbb{Z}\mathbb{Z}'\mathbb{x}$,

where $\mathbb{Z} = (\mathbb{z}_1,\ldots,\mathbb{z}_k,\ldots,\mathbb{z}_K)$ denote the unit normalized condensed variates,
with $\mathbb{z}_k = \mathbb{S}_k\mathbb{x}(\mathbb{x}'\mathbb{S}_k\mathbb{S}_k\mathbb{x})^{-1/2}$.

The corresponding DC path diagram with secondary predictions is given in figure 5.15.A.

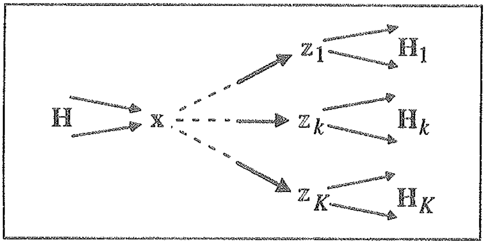


Figure 5.15.A *DC path diagram for Set Component Analysis.*

The path diagram for SCA has to be fitted with the DC fit function (5.7) using real function weights and weight function (5.55). After substitution of $\mathbb{W}=\mathbb{F}(\mathbf{t})=\mathbb{W}_{\text{Design}}*\mathbb{R}$ in (5.7), we have the equality SCA(x) =

$$\text{SCA}_{\text{DC}}:\ \text{Fit}(\mathbf{t},\mathbb{W}) = \mathbf{u}'(\mathbb{R}*\mathbb{R}*\mathbb{W}_{\text{Design}})\mathbf{u}, \qquad (5.79)$$

with

$$\mathbb{R}*\mathbb{W}_{\text{Design}} \quad = \quad \begin{array}{|c|c|c|c|} \hline 0 & 0 & 0 & 0 \\ \hline \mathbf{z}_1'\mathbf{x} & 0 & 0 & 0 \\ \hline \mathbf{z}_k'\mathbf{x} & 0 & 0 & 0 \\ \hline \mathbf{z}_K'\mathbf{x} & 0 & 0 & 0 \\ \hline \end{array} \qquad (5.80)$$

where $\mathbf{x} = \mathbb{H}_0\mathbf{t}_0 = \mathbb{H}^{\alpha}\mathbf{t}_0$ denotes a regression variate with α_0=0, so $\mathbb{H}^0$=PQ',

and $\mathbf{z}_k = \mathbb{H}_k\mathbf{t}_k = \mathbb{H}_k^{\alpha}\mathbf{t}_k$, denote condensed variates with α_k=1, $\forall k$
$= \mathbf{S}_k\mathbf{x}(\mathbf{x}'\mathbf{S}_k\mathbf{S}_k\mathbf{x})^{-1/2}$,

with $\mathbb{H} = (\mathbb{H}_0,\ldots,\mathbb{H}_1,\ldots,\mathbb{H}_k,\ldots,\mathbb{H}_K)$
$= (\mathbb{H}^0,\ldots,\mathbb{H}_1^1,\ldots,\mathbb{H}_k^1,\ldots,\mathbb{H}_K^1) = (\mathbb{H}^0,\mathbb{H}^1) = (\mathbf{PQ}',\mathbf{P\Phi Q}')$.

The condensed variates $\mathbf{z}_k$ are exactly as introduced in section 5.2.2. For more dimensional solutions, there are no orthogonal restrictions on $\mathbf{z}_k$, only on $\mathbf{x}$. If all prediction arrows in figure 5.15.A are primary the condensed variates $\mathbf{z}_k$ in (5.80) are replaced by the regression variates $\mathbf{z}_k$ and we obtain in (5.79) a DC fit function for MCCA (see section 2.2.6).

The optimal SCA$_{\text{DC}}$ parameters can be found with an algorithm described in chapter 6. The SCA path model can also be fitted with the LDC fit function. We cannot use the LDC algorithm of section 5.3.6 for this purpose, because we confined ourselves to find a fitting procedure for LDC with fixed or proportional function weights and with $(\mathbf{W}_{\text{Design}}\mathbf{1})_{l}\neq\mathbf{0}$, $\forall l$. By adding simple extensions to the SCA method we can use the LDC algorithm for proportional function weights.

One simple extension to the SCA method would be to impose a subspace restriction on $\mathbf{x}$ (see section 2.1.1). Figure 5.15.B gives a path diagram of this extended path model.

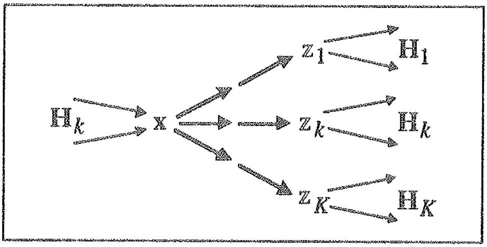


Figure 5.15.B *DC path diagram for SCA with subspace restriction.*

The subspace restriction $\mathbf{P}_k\mathbf{P}_k'\mathbf{x}$ can easily be imposed on $\mathbf{x}$, by the following specification of $\mathbb{R}*\mathbf{W}_{\text{Design}}$

$$\mathbb{R}*\mathbf{W}_{\text{Design}} \quad = \quad \begin{array}{|c|c|c|c|}\hline 0 & 0 & \mathbf{x}'\mathbf{z}_k & 0 \\ \hline \mathbf{z}_1'\mathbf{x} & 0 & 0 & 0 \\ \hline \mathbf{z}_k'\mathbf{x} & 0 & 0 & 0 \\ \hline \mathbf{z}_K'\mathbf{x} & 0 & 0 & 0 \\ \hline \end{array} \qquad (5.81)$$

For (5.81) we can compute a LDC solution with proportional function weights with the LDC algorithm of section 5.3.6, because we have $(\mathbf{W}_{\text{Design}}\mathbf{1})_{l}\neq\mathbf{0}$, $\forall l$. For a more dimensional solution we can apply the Pattor rotation procedure developed by Lohmöller (see section 5.4.5).

Another simple extension of the SCA method would be a symmetric formulation of the design $\mathbf{W}_{\text{Design}}$. The symmetric formulation is visualised in the path diagram in figure 5.12. In this figure primary prediction arrows are added compared to figure 5.15.A. Fitting this path diagram with LDC or with DC (5.79) would imply for $\mathbb{R}*\mathbf{W}_{\text{Design}}$ a specification according to (5.64) and for $\mathbf{x}$ the restriction

$$\mathbf{x} = \mathbb{Z}\mathbf{w}((\mathbb{Z}\mathbf{w})'\mathbb{Z}\mathbf{w})^{-1/2} = \mathbb{Z}\mathbb{Z}'\mathbf{x}((\mathbf{x}'\mathbb{Z}\mathbb{Z}'\mathbb{Z}\mathbb{Z}'\mathbf{x})^{-1/2}, \tag{5.82}$$

where $\mathbb{Z} = (\mathbf{z}_1,\ldots,\mathbf{z}_k,\ldots,\mathbf{z}_K)$ with $\mathbf{z}_k = \mathbf{S}_k\mathbf{x}(\mathbf{x}'\mathbf{S}_k\mathbf{S}_k\mathbf{x})^{-1/2}$,
and $\mathbf{w}' = (w_1,\ldots,w_k,\ldots,w_K)$ with $w_k = \mathbf{z}_k'\mathbf{x}$.

The additional restriction on $\mathbf{x}$ will in principle lead to a different solution for the symmetric formulation of SCA$_{DC}$. If we fit the modified path diagram in figure 5.12 with proportional function weights and weight function (5.55) the LDC formulation gives a PLS Hierarchical Components (HC) algorithm with factor weighting scheme, general factor $\mathbf{x}$ in mode B and special factors $\mathbf{z}_k$ in mode A (Lohmöller, 1989, page 131). We refer to this algorithm as the HC$_{SCA}$ algorithm. By changing the directions of the arrows in figure 5.12 all other modes of Hierarchical Components methods can be specified with LDC path diagram and fitted with the LDC fit function. This can be proven along the same lines as we did for Consensus PLS in section 5.4.6 and will not be elaborated here. The HC$_{SCA}$ algorithm is given by iteratively repeating (5.82). The same algorithm is obtained by fitting with the bPLS fit function (5.40) the hierarchical path diagram in figure 5.15.C.

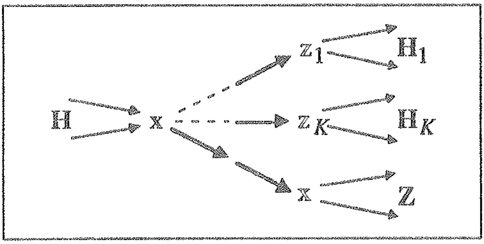


Figure 5.15.C *Alternative path diagram for HC$_{SCA}$ method.*

Pivot variate $\mathbf{x}$ predicts primary and secondary the condensed variate $\mathbf{x}$ of the ancillary set of condensed variates $\mathbb{Z}$. At convergence the optimal $\mathbf{x}$ is equal to $\mathbf{x}$ and equal to the first principal component of $\mathbb{Z}$.

Summarising, we formulated in this section SCA as a DC path model with real function weights and some LDC extensions of the SCA method, like the PLS Hierarchical Components method. Finally we remark that this HC method has an interesting relation with another PLS method: The only difference in the LDC formulation of the HC method and Consensus PLS is the definition of their respective weight functions (5.55) and (5.63).

5.5 Comparison of LDC and DC

We expect that fitting path models with LDC and DC very often gives the same results, because they both have the same severe restrictions on the directions of the variates z_k. If the solutions are different it is possible that the DC solution gives a better prediction of the variables than the LDC solution and is still acceptably stable. The directed correlations will probably be higher and therefore so will the prediction. The LRPCV fit (5.3) will probably be not so much lower that it seriously affects stability. The SCA method in chapter 3 gives an indication for this tendency of DC to maintain stability. It would be interesting to investigate DC variants of LDC methods like PLS2 and PLS1. On the other hand we expect that the predictive power of path models can be increased not so much by the choice between LDC and DC, but more drastically by the formulation of adequate path models, like for instance the path model of PLS2$^{\mathrm{multi}}$ instead of ordinary PLS2 in section 5.4.8. From a practical point of view the choice for the LDC fit function is more likely, because we developed an algorithm for LDC that can handle a wide variety of path models. For DC such a general algorithm is not yet available.

Chapter 6

ALGORITHMS

We present two algorithms for non eigenvalue-eigenvector problems. First a simultaneous and successive monotone convergent algorithm for Set Component Analysis (chapter 3) is developed, where an interesting general algorithmic subproblem is to maximize the variance of different matrices accounted for by corresponding orthogonal latent variables. Secondly we elaborate a monotone convergent algorithm for Nonlinear Reflected Discriminant Analysis (chapter 4).

Introduction

The optimal parameters for almost all methods in this monograph can be estimated by solving an eigenvalue-eigenvector problem. In computing practice numerous algorithms are available to the researcher for executing the job. The methods previously presented which cannot be estimated in this way are SCA (chapter 3), NRDA (chapter 4) , DC and LDC (chapter 5). The algorithmic aspects of DC and LDC are already treated in chapter 5 for reasons mentioned there. An algorithm for SCA is elaborated in section 6.1 and for NRDA in section 6.2.

6.1 Computation of the SCA method

For the maximization of the SCA fit function (3.5) we use the reformulation of this function as given in (3.7)

$$\text{SCA:} \qquad \text{Fit}(\mathbf{x}_s, w_{(k)s}) = \sum_{s=1}^{p}\sum_{k=1}^{K} \mathbf{x}_s'\mathbf{x}_s - (\mathbf{x}_s - \mathbf{S}_k\mathbf{x}_s w_{(k)s}^{-1})'(\mathbf{x}_s - \mathbf{S}_k\mathbf{x}_s w_{(k)s}^{-1})$$

$$= \sum_{s=1}^{p}\sum_{k=1}^{K} \mathbf{x}_s'\mathbf{P}_k\{\mathbf{I} - (\mathbf{I} - \mathbf{\Phi}_k^2 w_{(k)s}^{-1})^2\}\mathbf{P}_k'\mathbf{x}_s,$$

where $w_{(1)s},\ldots,w_{(k)s},\ldots,w_{(K)s}$ denote free balancing factors for set k and dimension s and the remaining parameters are defined as usual.

The maximization of (3.7) is simpler than (3.5), because there is no complicated function of $\mathbf{x}_s$ in the denominator.

6.1.1 *Simultaneous SCA solution*

For the *simultaneous* SCA solution we maximize (3.7). The qualification simultaneous is needed to distinguish this solution from the *successive* solution. This solution first maximizes (3.7) for one dimension $\mathbb{x}_1$. A second dimension $\mathbb{x}_2$ then must be determined such that it maximizes (3.7) with $\mathbb{X}_{1,2}'\mathbb{X}_{1,2}=\mathbb{I}$ and $\mathbb{x}_1$ fixed. We proceed this way until p dimensions of $\mathbb{X}$ are computed, while keeping all previous dimensions fixed. By this procedure we introduce a hierarchical ordering of the successive dimensions in terms of maximizing the SCA fit function. The simultaneous solution has in principle no restrictions in terms of fixing previous dimensions. Therefore the simultaneous SCA solution cannot have a maximum less than the maximum of the successive solution. On the other hand the fit of the first dimension of the successive SCA solution is always greater than or equal to the fit of any separate dimension of the simultaneous solution.

The iterative ALS algorithm for simultaneous SCA consist of two alternating main steps. In the first main step the balancing factors $w_{(k)s}$, $\forall k,s$, are updated for given $\mathbb{X}$. In the second main step the $\mathbb{X}$ are updated for given $w_{(k)s}$ by applying an iterative sub-algorithm. This algorithm is obtained by modifying a procedure described by Ten Berge (1986, 1988) for maximizing the Maxbet function. The first step is specified in section 6.1.2 and the second step in 6.1.3.

6.1.2 *Balancing factors*

The optimal balancing factors $w_{(k)s}$ for all sets are updated in the first main step. By fixing the $\mathbb{X}$ in (3.7) and setting the first derivative equal to zero we find *suboptimal* balancing factors, which are a function of $\mathbb{X}$. We denote these suboptimal balancing factors with $\hat{w}_{(k)s}$. The suboptimal balancing factors are given in (3.9)

$$\hat{w}_{(k)s} = \frac{\mathbb{x}_s'\mathbb{S}_k\mathbb{S}_k\mathbb{x}_s}{\mathbb{x}_s'\mathbb{S}_k\mathbb{x}_s} = \frac{\mathbb{t}_{(k)s}'\mathbb{H}_k'\mathbb{H}_k\mathbb{t}_{(k)s}}{\mathbb{t}_{(k)s}'\mathbb{t}_{(k)s}} = \frac{\mathbb{t}_{(k)s}'\mathbb{Q}_k\Phi_k^2\mathbb{Q}_k'\mathbb{t}_{(k)s}}{\mathbb{t}_{(k)s}'\mathbb{t}_{(k)s}}, \qquad \forall k,s$$

with $\mathbb{t}_{(k)s} = \mathbb{H}_k'\mathbb{x}_s, \quad \forall k,s.$

For the first main step the updates for the optimal balancing factors $w_{(k)s}$ with $\mathbb{X}$ fixed are specified in (3.9). The updates for $\mathbb{X}$ with fixed optimal balancing factors $w_{(k)s}$ are specified in the next section.

6.1.3 The variance of different matrices accounted for simultaneously

The general problem we have to solve in the second main step is to maximize the variance of different matrices accounted for by corresponding orthogonal latent variables. In this particular case we maximize (3.7) with $w_{(k)s} = \hat{w}_{(k)s}$ (3.9). The resulting function is $f(\mathbb{X}) - \Sigma_s c_s$, with $\Sigma_s c_s$ constant and

$$f(\mathbb{X}) = \sum_{s=1}^{p} \mathbb{x}_s{}'\mathbb{B}_s\mathbb{x}_s \tag{6.1}$$

where $\quad {}_n\mathbb{X}_p = (\mathbb{x}_1,\ldots,\mathbb{x}_s,\ldots,\mathbb{x}_p)$ denote the common latent variables with $\mathbb{X}'\mathbb{X}=\mathbb{I}$

$$\mathbb{B}_s = c_s\mathbb{I} + \sum_{k=1}^{K} \mathbb{P}_k\{\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2\}\mathbb{P}_k{}' \qquad \forall s$$

$c_1,\ldots,c_s,\ldots,c_p$ denote constant scalars.

Maximization of $f(\mathbb{X}) - \Sigma_s c_s$ gives the same results for $\mathbb{X}$ as maximization of $f(\mathbb{X})$. We maximize $f(\mathbb{X})$, because the c_s are chosen in such a way that the corresponding matrices $\mathbb{B}_s$ are positive semi-definite as will be explained later. An appropriate choice for each c_s is the negative of the smallest eigenvalue of
$\Sigma_k\mathbb{P}_k\{\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2\}\mathbb{P}_k{}'$, in another notation written as
$-\lambda_{\min}(\Sigma_k\mathbb{P}_k\{\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2\}\mathbb{P}_k{}')$.

We developed an iterative sub-algorithm for increasing $f(\mathbb{X})$ in (6.1) monotonely, with matrices $\mathbb{B}_s$ symmetric positive semi-definite $\forall s$. This algorithm, which is an adaptation of a procedure described by Ten Berge (1986), can be constructed in the following way.

Theorem 6.1. For arbitrary starting matrix $\mathbb{X}$ satisfying $\mathbb{X}'\mathbb{X}=\mathbb{I}$, consider the SVD

$$(\mathbb{B}_1\mathbb{x}_1,\ldots,\mathbb{B}_s\mathbb{x}_s,\ldots,\mathbb{B}_p\mathbb{x}_p) = \mathbb{M}\Psi\mathbb{N}', \tag{6.2}$$

where $\mathbf{M}$ ($n\times p$) and $\mathbf{N}$ ($p\times p$) denote orthonormal singular vector matrices and Ψ denotes a diagonal matrix with p singular values. Let the matrix $\mathbf{X}$ be updated by setting

$$\mathbf{X}^u = \mathbf{M}\mathbf{N}'. \tag{6.3}$$

Following Ten Berge (1986), which is analogous to the case $K=1$, we now have

$$f(\mathbf{X}^u) \geq f(\mathbf{X}), \tag{6.4}$$

which implies that (6.3) increases $f(\mathbf{X})$ monotonely.

Proof. Verification of (6.4) is possible in two steps by defining the auxiliary function $\hat{f}(\mathbf{Z})$ as

$$\hat{f}(\mathbf{Z}) = \sum_{s=1}^{p} \mathbf{z}_s'\mathbf{B}_s\mathbf{x}_s = \operatorname{tr} \mathbf{Z}'(\mathbf{B}_1\mathbf{x}_1,\ldots,\mathbf{B}_s\mathbf{x}_s,\ldots,\mathbf{B}_p\mathbf{x}_p), \tag{6.5}$$

with $\mathbf{x}_s$ fixed $\forall s$ and $\mathbf{Z}'\mathbf{Z}=\mathbf{I}$. The constrained maximum of (6.5) is attained for $\mathbf{Z} = \mathbf{X}^u$ given by (6.3), cf. Green (1969). Hence we have

$$\hat{f}(\mathbf{X}^u) \geq f(\mathbf{X}). \tag{6.6}$$

In the second step we apply the Cauchy-Schwarz inequality

$$\begin{aligned}\hat{f}(\mathbf{X}^u) &= \sum_{s=1}^{p} \mathbf{x}_s^{u\prime}\mathbf{B}_s\mathbf{x}_s = \operatorname{tr} \mathbf{Y}^{u\prime}\mathbf{Y} \;\leq\; (\operatorname{tr} \mathbf{Y}^{u\prime}\mathbf{Y}^u)^{1/2}(\operatorname{tr} \mathbf{Y}'\mathbf{Y})^{1/2} = \\ &= (\sum_{s=1}^{p} \mathbf{x}_s'\mathbf{B}_s\mathbf{x}_s\,)^{1/2}(\sum_{s=1}^{p} \mathbf{x}_s^{u\prime}\mathbf{B}_s\mathbf{x}_s^u)^{1/2} = f(\mathbf{X}^u)^{1/2}f(\mathbf{X})^{1/2},\end{aligned} \tag{6.7}$$

where
$$\begin{aligned}\mathbf{Y}^u &= (\mathbf{B}_1^{1/2}\mathbf{x}_1^u,\ldots,\mathbf{B}_s^{1/2}\mathbf{x}_s^u,\ldots,\mathbf{B}_p^{1/2}\mathbf{x}_p^u)\\ \mathbf{Y} &= (\mathbf{B}_1^{1/2}\mathbf{x}_1,\ldots,\mathbf{B}_s^{1/2}\mathbf{x}_s,\ldots,\mathbf{B}_p^{1/2}\mathbf{x}_p).\end{aligned}$$

We can apply the Cauchy-Schwarz inequality, because the matrices $\mathbf{B}_s$ are positive semi-definite and therefore can be written as $\mathbf{B}_s = \mathbf{B}_s^{1/2}\mathbf{B}_s^{1/2}$, where $\mathbf{B}_s^{1/2}$ is the unique positive semi-definite square root of $\mathbf{B}_s$. This is why convergence of the algorithm presented here is guaranteed only if the matrices $\mathbf{B}_s$ are positive semi-definite.

Finally we combine the inequalities (6.6) and (6.7) into one sequence of connected inequalities $f(\mathbb{X}) \leq \hat{f}(\mathbb{X}^u) \leq f(\mathbb{X}^u)^{1/2} f(\mathbb{X})^{1/2}$, which implies $f(\mathbb{X})^{1/2} \leq f(\mathbb{X}^u)^{1/2}$ and completes the proof of (6.4). □

A necessary and sufficient condition for convergence can be derived. That is, $\mathbb{X}$ cannot be improved if and only if (6.4) holds as an equality, and therefore by combining (6.2) and (6.3)

$$(\mathbb{B}_1\mathbb{x}_1,\ldots,\mathbb{B}_s\mathbb{x}_s,\ldots,\mathbb{B}_p\mathbb{x}_p) = \mathbb{X}(\mathbb{N}\Psi\mathbb{N}') = \mathbb{X}\Delta, \qquad (6.8)$$

for certain positive semi-definite matrix Δ. It follows that (6.8) is a necessary condition for a global maximum of $f(\mathbb{X})$.

6.1.4 The algorithm for simultaneous SCA

In the initialization steps of the algorithm for simultaneous SCA we can choose any arbitrary starting matrix $\mathbb{X}^0$. Nevertheless convergence is faster if we start with a reasonable guess. We fix the balancing factors (3.9) for all dimensions to their maximum value ϕ_{1k}^2, which is the largest eigenvalue of $\mathbb{H}_k'\mathbb{H}_k$. After substitution in (3.7) we compute the optimal $\mathbb{X}^0$. For convenience, in the final step we rearrange the dimensions of $\mathbb{X}$ in such a way that the SCA function is decreasing. The complete algorithm for simultaneous SCA can be summarized as follows.

Initialization:

Step 1. Compute SVD $\mathbb{H}_k = \mathbb{P}_k\Phi_k\mathbb{Q}_k'$ $\qquad \forall k$

Step 2. Compute EigenVD $\sum_{k=1}^{K} \mathbb{P}_k\{\mathbb{I} - (\mathbb{I} - \Phi_k^2/\phi_{1k}^2)^2\}\mathbb{P}_k' = \mathbb{K}\Lambda\mathbb{K}'$

Step 3. Set $\mathbb{X}^0 = {}_n\mathbb{K}_p = (\mathbb{k}_1,\ldots,\mathbb{k}_s,\ldots,\mathbb{k}_p)$

Iterations:

Step 4. Compute $w_{(k)s}^i = \dfrac{\mathbb{x}_s^{i}{}'\mathbb{P}_k\Phi_k^4\mathbb{P}_k'\mathbb{x}_s^i}{\mathbb{x}_s^{i}{}'\mathbb{P}_k\Phi_k^2\mathbb{P}_k'\mathbb{x}_s^i}$ $\qquad \forall k,s$

Step 5. Compute $\underline{\mathbb{B}}_s^i = \sum_{k=1}^{K} \mathbb{P}_k\{\mathbb{I} - (\mathbb{I} - \Phi_k^2(w_{(k)s}^i)^{-1})^2\}\mathbb{P}_k'$ $\forall s$

Step 6. Compute $c_s^i = -\lambda_{\min}(\underline{\mathbb{B}}_s^i)$ $\forall s$

Step 7. Compute $\mathbb{B}_s^i = c_s^i\mathbb{I} + \underline{\mathbb{B}}_s^i$ $\forall s$

Step 8. Compute $\mathbb{B}^i = (\mathbb{B}_1^i\mathbf{x}_1^i,\ldots,\mathbb{B}_s^i\mathbf{x}_s^i,\ldots,\mathbb{B}_p^i\mathbf{x}_p^i)$

Step 9. Compute $\mathbb{X}^{i+1} = \mathbb{B}^i(\mathbb{B}^{i\prime}\mathbb{B}^i)^{-1/2}$

Step 10. Evaluate $\mathrm{SCA}(\mathbb{X}^{i+1}) = \sum_{s=1}^{p} (\mathbf{x}_s^{i+1\prime}\mathbb{B}_s^i\mathbf{x}_s^{i+1} - c_s^i)$

If $\mathrm{SCA}(\mathbb{X}^{i+1}) - \mathrm{SCA}(\mathbb{X}^i) > \varepsilon$, for some small value ε, then go to Step 4.

Termination:

Step 11. Rearrange dimensions $\mathrm{SCA}(\mathbf{x}_1^{i+1}) \geq \mathrm{SCA}(\mathbf{x}_s^{i+1}) \geq \mathrm{SCA}(\mathbf{x}_p^{i+1})$

Step 8 and 9 can be repeated in inner iterations many times for updating $\mathbb{X}$ as shown in section 6.1.3. We do not have general recommendations for optimal tuning. For the simultaneous SCA solution we have in general no rotational freedom as can be found in simultaneous formulations of methods like PCA, CCA or Multiset CCA (Carroll, 1968). Rotational freedom is guaranteed if $\Phi_k^2 = \mathbb{I}$, $\forall k$, because in that special case $w_{(k)s}=1$, $\forall k,s$ and SCA comes down to the same thing as MCCA.

6.1.5 Computational short cuts

Although an appropriate choice for the constant scalar c_s in Step 6 is the negative of the smallest eigenvalue of $\underline{\mathbb{B}}_s^i$, we can define another estimate of c_s that is computationally less demanding. We call this estimate $\hat{c}_s$ and it has c_s as an lower bound. In this way the matrices $\mathbb{B}_s$ in Step 7 will always be positive semi-definite.

Theorem 6.2. The estimate

$$\hat{c}_s = \sum_{k=1}^{K} -(0,\{1-(1-\phi_{1k}^2\hat{w}_{(k)s}^{-1})^2\})_{\min}, \tag{6.9}$$

where $()_{\min}$ gives the minimum value of the two elements between the brackets, has $c_s = -\lambda_{\min}(\underline{\mathbb{B}}_s)$ as a lower bound.

In order to verify Theorem 6.2, $c_s \leq \hat{c}_s$, $\forall s$, we check the validity of the following three equations

$$c_s = -\lambda_{\min}(\underline{\mathbb{B}}_s) \leq \sum_{k=1}^{K} -\lambda_{\min}(\underline{\mathbb{B}}_{(k)s}), \qquad \forall s \quad (6.10)$$

$$-\lambda_{\min}(\underline{\mathbb{B}}_{(k)s}) \leq -(\mathbb{I} - (\mathbb{I} - \Phi_k^2 \hat{w}_{(k)s}^{-1})^2)_{\min}, \qquad \forall k,s \quad (6.11)$$

and $$(\mathbb{I} - (\mathbb{I} - \Phi_k^2 \hat{w}_{(k)s}^{-1})^2)_{\min} = (0, \{1 - (1 - \phi_{1k}^2 \hat{w}_{(k)s}^{-1})^2\})_{\min}. \qquad \forall k,s \quad (6.12)$$

where $\lambda_{\min}()$ gives the minimum eigenvalue of the matrix between the brackets,

$\underline{\mathbb{B}}_{(k)s}$ $= \mathbb{P}_k\{\mathbb{I} - (\mathbb{I} - \Phi_k^2 \hat{w}_{(k)s}^{-1})^2\}\mathbb{P}_k'$, $\forall k,s$

$\underline{\mathbb{B}}_s$ $= \Sigma_k \underline{\mathbb{B}}_{(k)s}$, $\forall s$

and $()_{\min}$ gives the minimum value of *all* the elements of the matrix or string between the brackets.

Proof (6.10). By expanding (6.10) as $c_s = -\lambda_{\min}(\underline{\mathbb{B}}_s) = -\lambda_{\min}(\Sigma_k \underline{\mathbb{B}}_{(k)s}) \leq \Sigma_k - \lambda_{\min}(\underline{\mathbb{B}}_{(k)s})$, we have to prove that $\lambda_{\min}(\Sigma_k \underline{\mathbb{B}}_{(k)s}) \geq \Sigma_k \lambda_{\min}(\underline{\mathbb{B}}_{(k)s})$. Therefore we introduce in the following equation the functions g(x) and g_k(x)

$$g(\mathrm{x}) = \mathrm{x}'\underline{\mathbb{B}}_s\mathrm{x} = \mathrm{x}'(\sum_{k=1}^{K} \underline{\mathbb{B}}_{(k)s})\mathrm{x} = \sum_{k=1}^{K} \mathrm{x}'\underline{\mathbb{B}}_{(k)s}\mathrm{x} = \sum_{k=1}^{K} g_k(\mathrm{x}), \qquad (6.13)$$

with x'x = 1. Obviously we have

$$g(\mathrm{x}) \geq \lambda_{\min}(\underline{\mathbb{B}}_s) = g(\mathrm{y})$$

and $$g_k(\mathrm{x}) \geq \lambda_{\min}(\underline{\mathbb{B}}_{(k)s}) = g_k(\mathrm{y}_k), \qquad \forall k \quad (6.14)$$

with $\mathrm{x}'\mathrm{x} = \mathrm{y}'\mathrm{y} = \mathrm{y}_k'\mathrm{y}_k = 1$, $\forall k$. The vectors y and y_k denote the eigenvectors corresponding to the smallest eigenvalues of respectively $\underline{\mathbb{B}}_s$ and $\underline{\mathbb{B}}_{(k)s}$. We substitute y in (6.13) and in the second term of (6.14) and obtain respectively $g(\mathrm{y}) = \Sigma_k g_k(\mathrm{y})$ and $g_k(\mathrm{y}) \geq g_k(\mathrm{y}_k)$. Combining these equations into $g(\mathrm{y}) \geq \Sigma_k g_k(\mathrm{y}_k)$ gives $\lambda_{\min}(\Sigma_k \underline{\mathbb{B}}_{(k)s}) \geq \Sigma_k \lambda_{\min}(\underline{\mathbb{B}}_{(k)s})$ and therefore completes the proof of (6.10). □

Proof (6.11). In order to verify (6.11) we substitute $g_k(\mathbf{y}_k)$ as defined in (6.13) and (6.14)

$$\begin{aligned}\lambda_{\min}(\underline{\mathbb{B}}_{(k)s}) &= g_k(\mathbf{y}_k) = \mathbf{y}_k'\underline{\mathbb{B}}_{(k)s}\mathbf{y}_k \\ &= \mathbf{y}_k'\mathbb{P}_k\{\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2\}\mathbb{P}_k'\mathbf{y}_k + \mathbf{y}_k'\underline{\mathbb{P}}_k\Lambda_0\underline{\mathbb{P}}_k'\mathbf{y}_k \\ &= \mathbf{y}_k'(\mathbb{P}_k,\underline{\mathbb{P}}_k)\Lambda_{(k)s}^2(\mathbb{P}_k,\underline{\mathbb{P}}_k)'\mathbf{y}_k, \qquad \forall k,s \qquad (6.15)\end{aligned}$$

where Λ_0 denotes a matrix of appropriate size with only zero elements,

$\underline{\mathbb{P}}_k$ is the orthonormal complement of $\mathbb{P}_k$, so that $(\mathbb{P}_k,\underline{\mathbb{P}}_k)'(\mathbb{P}_k,\underline{\mathbb{P}}_k) = (\mathbb{P}_k,\underline{\mathbb{P}}_k)(\mathbb{P}_k,\underline{\mathbb{P}}_k)' = \mathbb{I}_n$, $\forall k$

and $\Lambda_{(k)s}$ denotes a diagonal $(n\times n)$ matrix containing the $(p_k\times p_k)$ matrix $\{\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2\}$ in its upper left corner and zeros elsewhere. $\forall k,s$

In fact (6.15) entails a full eigenvalue decomposition of $\underline{\mathbb{B}}_{(k)s}$ with all eigenvectors $(\mathbb{P}_k,\underline{\mathbb{P}}_k)$ and corresponding eigenvalues on the diagonal of $\Lambda_{(k)s}$. This implies that the smallest eigenvalue of $\underline{\mathbb{B}}_{(k)s}$ is equal to the minimum diagonal value of $\Lambda_{(k)s}$. We summarize our results as $\lambda_{\min}(\underline{\mathbb{B}}_{(k)s}) = (\text{diag}(\Lambda_{(k)s}))_{\min} \geq (\Lambda_{(k)s})_{\min}$. The inequality in the previous sequence only occurs if all diagonal elements of $\Lambda_{(k)s}$ are greater than zero. In that special case we have $(\text{diag}(\Lambda_{(k)s}))_{\min} > 0 = (\Lambda_{(k)s})_{\min}$ and we know that the number of columns of $\mathbb{P}_k$ is at least equal to the number of rows, $(p_k \geq n)$. The definition of $\Lambda_{(k)s}$ in (6.15) implies that $(\Lambda_{(k)s})_{\min} = (\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2)_{\min}$ and therefore completes the proof of (6.11). □

Proof (6.12). The proof for (6.12) is given by first replacing this equation of matrix functions by an equation of scalar functions with arguments ϕ^2 and w. Without loss of generality we omit the subscripts k and s in order to simplify notation and reformulate (6.12) as

$$(0,\{1 - (1 - \phi^2 w^{-1})^2\})_{\min} = (0,\{1 - (1 - \phi_{\max}^2 w^{-1})^2\})_{\min}, \qquad (6.16)$$

with $0 \leq \phi^2 \leq \phi^2_{\max}$ and $0 \leq w \leq \phi^2_{\max}$,

where ϕ^2 denotes the possible values of the diagonal elements of Φ^2_k, with $\phi^2_{\max}$ equal to the maximum value of ϕ^2,

and w denotes the possible values of the corresponding balancing factors $\hat{w}_{(k)s}$.

The boundaries for w can easily be verified in (3.9). We evaluate in detail the functions $\{1 - (1 - \phi^2 w^{-1})^2\}$ and $\{1 - (1 - \phi^2_{\max} w^{-1})^2\}$ in (6.16) for the cases $\phi^2 \leq w$ and $\phi^2 > w$ and obtain

$$0 \leq \{1 - (1 - \phi^2 w^{-1})^2\} \leq 1 \qquad \text{for } 0 \leq \phi^2 \leq w \leq \phi^2_{\max}$$

$$\text{and} \quad \{1 - (1 - \phi^2_{\max} w^{-1})^2\} \leq \{1 - (1 - \phi^2 w^{-1})^2\} < 1 \qquad \text{for } 0 \leq w < \phi^2 \leq \phi^2_{\max}. \quad (6.17)$$

The first term does not contain a function with $\phi^2_{\max}$, because for $0 \leq \phi^2 \leq w \leq \phi^2_{\max}$ we have only one possible value for w, when $\phi^2 = \phi^2_{\max}$, namely $w = \phi^2_{\max}$. This implies that the function $\{1 - (1 - \phi^2_{\max} w^{-1})^2\}$ can only be equal to 1 in the case $\phi^2 \leq w$. The lower bounds in (6.17) imply that (6.16) and therefore (6.12) is always valid. □

Summarising the previous exposé we can now replace Step 6 by

Step 6. Compute $c^i_s = -(0, \{1 - (1 - \phi^2_{1k}(w^i_{(k)s})^{-1})^2\})_{\min}$. $\forall s$

Another computational short cut for the simultaneous SCA algorithm in section 6.1.4 concerns the size of the $\mathbb{P}_k$ and $\mathbb{B}_s$ matrices, especially when $n \gg \sum_k m_k$. The $\mathbb{P}_k$ matrices are $(n \times p_k)$ and therefore all the $\mathbb{B}_s$ matrices are $(n \times n)$. Manipulations with these matrices make much greater demands on computer time if n is large compared to some fixed total number of variables $\sum_k m_k$. A simple remedy for this phenomenon is offered by adding or changing the following steps of the SCA algorithm in section 6.1.4

Step 0. Compute SVD $\mathbb{H} = \mathbb{P}\Phi\mathbb{Q}'$

Step 1. Compute SVD $(\mathbb{P}'\mathbb{H}_k) = \mathbb{P}_k\Phi_k\mathbb{Q}_k'$ $\forall k$

Step 12. Compute $\mathbb{X}^{i+2} = \mathbb{P}\mathbb{X}^{i+1}$ $\forall k$

For the SVD $\mathbb{H}=(\mathbb{H}_1,..,\mathbb{H}_k,..,\mathbb{H}_K)=\mathbb{P}\Phi\mathbb{Q}'$ we have the orthonormal singular vector matrices $\mathbb{P}$ $(n\times P)$ and $\mathbb{Q}$ $(\Sigma_k m_k \times P)$ corresponding to non-zero singular values in diagonal matrix Φ $(P\times P)$, with $P \leq \Sigma_k m_k$. The only purpose of Step 0 is to find a description of the space of $\mathbb{H}$ by an orthonormal basis $\mathbb{P}$ with a number of columns P equal to the rank of $\mathbb{H}$. Therefore any other (faster) technique for finding such a space would be acceptable. Take for instance the complete orthogonal factorization of $\mathbb{H}$ (see e.g. Gill, Murray & Wright, 1981, page 39). As a result of Steps 0 and 1 the size of the $\mathbb{P}_k$ matrices in the iteration steps now is reduced to $(P\times p_k)$ and the size of the $\mathbb{B}_s$ matrices to $(P\times P)$. In Step 12 we represent the solution in the original n-dimensional orthonormal basis instead of the auxiliary P-dimensional orthonormal basis.

6.1.6 Successive SCA solution

In successive SCA, the fit function formulated in (3.5) and (3.7) is maximized in successive steps for each dimension s. In other words for $s = 1,\ldots,p$ we maximize

$$\text{SCAsu:} \quad \text{Fit}(\mathbf{x}_s, w_{(k)s}) = \sum_{k=1}^{K} \mathbf{x}_s'\mathbb{A}_s\mathbb{P}_k\{\mathbb{I} - (\mathbb{I} - \Phi_k^2 w_{(k)s}^{-1})^2\}\mathbb{P}_k'\mathbb{A}_s\mathbf{x}_s, \tag{6.18}$$

with $\mathbb{A}_s = \mathbb{I}$ for $s = 1$

$\mathbb{A}_s = (\mathbb{I} - \mathbb{X}_{s-1}\mathbb{X}_{s-1}')$ for $s = 2,\ldots,p$, with $\mathbb{X}_{s-1} = (\mathbf{x}_1,\ldots,\mathbf{x}_{s-1})$.

This fit function is to be preferred to the simultaneous SCA fit function if we are interested in the highest possible fit for the first dimension and not in the highest possible fit for all p dimensions simultaneously. An alternative for (6.18) would be deflation, i.e. taking the antiprojection on the previous dimensions for each set $\mathbb{H}_{(k)s}$

$\mathbb{H}_{(k)s} = \mathbb{H}_k$ for $s = 1$ $\forall k$

$\mathbb{H}_{(k)s} = (\mathbb{I}-\mathbf{x}_{s-1}\mathbf{x}_{s-1}')\mathbb{H}_{(k)s-1}.$ for $s = 2,\ldots,p$ $\forall k$

and computing new eigenvectors $\mathbb{P}_{(k)s}$ and eigenvalues $\Phi_{(k)s}^2$ for each successive dimension. In principle this will lead to other solutions than maximization of (6.18). The solution for the deflation method can be found by computing p times a one dimensional simultaneous solution for each successive group of K matrices $\mathbb{H}_{(k)s}$.

6.1.7 The algorithm for successive SCA

The p -dimensional solution of $\mathbb{X}$ for successive SCA is essentially obtained by computing p times a one dimensional simultaneous solution. The only difference with this one dimensional simultaneous solution is that we have to add the antiprojection matrices $\mathbb{A}_s$ as defined in (6.18). The resulting algorithm for successive SCA is as follows.

```
Initialization:
    Step 1.   Compute SVD $\mathbb{H}_k = \mathbb{P}_{(k)1}\Phi_k\mathbb{Q}_k'$          $\forall k$
    Step 2.   Compute EigenVD $\sum_{k=1}^{K} \mathbb{P}_{(k)1}\{\mathbb{I} - (\mathbb{I} - \Phi_k^2/\phi_{1k}^2)^2\}\mathbb{P}_{(k)1}' = \mathbb{K}\Lambda\mathbb{K}'$
    Step 3.   Set $s = 1$ and $\mathbb{X}^0 = {}_n\mathbb{K}_p = (\mathbf{k}_1,\ldots,\mathbf{k}_s,\ldots,\mathbf{k}_p)$
Iterations:
    Step 4.   Compute $w_{(k)s}^i = \dfrac{\mathbf{x}_s^{i\,\prime}\mathbb{P}_{(k)s}\Phi_k^4\mathbb{P}_{(k)s}'\mathbf{x}_s^i}{\mathbf{x}_s^{i\,\prime}\mathbb{P}_{(k)s}\Phi_k^2\mathbb{P}_{(k)s}'\mathbf{x}_s^i}$          $\forall k$
    Step 5.   Compute $\underline{\mathbb{B}}_s^i = \sum_{k=1}^{K} \mathbb{P}_{(k)s}\{\mathbb{I} - (\mathbb{I} - \Phi_k^2(w_{(k)s}^i)^{-1})^2\}\mathbb{P}_{(k)s}'$
    Step 6.   Compute $c_s^i = -\lambda_{\min}(\underline{\mathbb{B}}_s^i)$
    Step 7.   Compute $\mathbb{B}_s^i = c_s^i\mathbb{I} + \underline{\mathbb{B}}_s^i$
    Step 8.   Compute $\mathbf{b}^i = \mathbb{B}_s^i\mathbf{x}_s^i$
    Step 9.   Compute $\mathbf{x}_s^{i+1} = \mathbf{b}^i(\mathbf{b}^{i\,\prime}\mathbf{b}^i)^{-1/2}$
    Step 10.  Evaluate $\mathrm{SCAsu}(\mathbf{x}_s^{i+1}) = \mathbf{x}_s^{i+1\,\prime}\mathbb{B}_s^i\mathbf{x}_s^{i+1} - c_s^i$
              If $\mathrm{SCAsu}(\mathbf{x}_s^{i+1}) - \mathrm{SCAsu}(\mathbf{x}_s^i) > \varepsilon$, for some small value $\varepsilon$,
              then go to Step 4.
    Step 11.  Compute $s = s + 1$
```

If $s > p$ then stop.

Step 12. Compute $\mathbb{P}_{(k)s} = (\mathbb{I} - \mathbb{X}_{s\text{-}1}\mathbb{X}_{s\text{-}1}')\mathbb{P}_{(k)1}$, $\forall k$
with $\mathbb{X}_{s\text{-}1} = (\mathbb{x}_1,\ldots,\mathbb{x}_{s\text{-}1})$

Go to Step 4.

Step 8 and 9 for updating $\mathbb{x}_s$ can be repeated in inner iterations many times as in the simultaneous algorithm of section 6.1.4.

6.1.8 Computational short cuts revised

The computational short cuts described in section 6.1.5 can also be applied to the successive SCA algorithm. We only need to make a minor adaptation in verifying Theorem 6.2, because the meaning of $\underline{\mathbb{B}}_{(k)s}$ changes from

$$\underline{\mathbb{B}}_{(k)s} = \mathbb{P}_k\{\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2\}\mathbb{P}_k' \text{ in section 6.1.5} \qquad \forall k,s$$

into $$\mathbb{B}_{(k)s} = \mathbb{P}_{(k)s}\{\mathbb{I} - (\mathbb{I} - \Phi_k^2(w_{(k)s}^{i})^{-1})^2\}\mathbb{P}_{(k)s}' \text{ in this section.} \qquad \forall k,s$$

The proof of (6.10) remains valid if we substitute $\mathbb{B}_{(k)s}$ for $\underline{\mathbb{B}}_{(k)s}$. For (6.11) this is not so obvious, because the equations in (6.15) no longer hold. Therefore it remains to prove that

$$\lambda_{\min}(\mathbb{B}_{(k)s}) \geq (\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2)_{\min}. \qquad \forall k,s \quad (6.19)$$

Proof. From Step 12 in section 6.1.7 we derive the relation between $\mathbb{B}_{(k)s}$ and $\underline{\mathbb{B}}_{(k)s}$:

$$\mathbb{B}_{(k)s} = \mathbb{A}_s\underline{\mathbb{B}}_{(k)s}\mathbb{A}_s, \qquad \forall k,s \quad (6.20)$$

with $\mathbb{A}_s = \mathbb{I}$ for $s = 1$
$\mathbb{A}_s = (\mathbb{I} - \mathbb{X}_{s\text{-}1}\mathbb{X}_{s\text{-}1}')$ for $s = 2,\ldots,p$, with $\mathbb{X}_{s\text{-}1} = (\mathbb{x}_1,\ldots,\mathbb{x}_{s\text{-}1})$.

Using the definition in (6.14) we have $\lambda_{\min}(\underline{\mathbb{B}}_{(k)s}) = \mathbb{y}_{(k)s}'\underline{\mathbb{B}}_{(k)s}\mathbb{y}_{(k)s}$ and $\lambda_{\min}(\mathbb{B}_{(k)s}) = \mathbb{y}_{(k)s}'\mathbb{B}_{(k)s}\mathbb{y}_{(k)s}$, where the vectors $\mathbb{y}_{(k)s}$ and $\mathbb{y}_{(k)s}$ denote the eigenvectors corresponding to the smallest eigenvalues of respectively $\underline{\mathbb{B}}_{(k)s}$ and $\mathbb{B}_{(k)s}$, $\forall k,s$.

When $(\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2)_{\min} = 0$, we know from (6.11) that $\lambda_{\min}(\underline{\mathbb{B}}_{(k)s}) = \mathbb{y}_{(k)s}'\underline{\mathbb{B}}_{(k)s}\mathbb{y}_{(k)s} \geq 0$. Because $\underline{\mathbb{B}}_{(k)s}$ is positive semi-definite, we have $\mathbb{z}'\underline{\mathbb{B}}_{(k)s}\mathbb{z} \geq 0$ for any other vector $\mathbb{z}$ and therefore $\mathbb{y}_{(k)s}'\mathbb{A}_s\underline{\mathbb{B}}_{(k)s}\mathbb{A}_s\mathbb{y}_{(k)s} = \lambda_{\min}(\mathbb{B}_{(k)s}) \geq 0$.

When $(\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2)_{\min} < 0$ and $\lambda_{\min}(\mathbb{B}_{(k)s}) \geq 0$, (6.19) is obviously true.

When $(\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2)_{\min} < 0$ and $\lambda_{\min}(\mathbb{B}_{(k)s}) < 0$, we also have $\lambda_{\min}(\underline{\mathbb{B}}_{(k)s}) < 0$, because in this case $\lambda_{\min}(\underline{\mathbb{B}}_{(k)s}) = (\mathbb{I} - (\mathbb{I} - \Phi_k^2\hat{w}_{(k)s}^{-1})^2)_{\min}$. In other words we must now verify $\lambda_{\min}(\mathbb{B}_{(k)s}) \geq \lambda_{\min}(\underline{\mathbb{B}}_{(k)s})$. In general we have for any projector matrix $\mathbb{A}_s = \mathbb{A}_s\mathbb{A}_s$ and for any unit normalized vector $\mathbb{z}$, $0 \leq \mathbb{z}'\mathbb{A}_s\mathbb{z} \leq 1$. Multiplying $\mathbb{y}_{(k)s}'\mathbb{A}_s\mathbb{y}_{(k)s} \leq 1$ with the negative value $\mathbb{y}_{(k)s}'\mathbb{B}_{(k)s}\mathbb{y}_{(k)s}$ we obtain

$$(\mathbb{y}_{(k)s}'\mathbb{B}_{(k)s}\mathbb{y}_{(k)s})(\mathbb{y}_{(k)s}'\mathbb{A}_s\mathbb{y}_{(k)s}) \;\geq\; \mathbb{y}_{(k)s}'\mathbb{B}_{(k)s}\mathbb{y}_{(k)s} \qquad \forall k,s$$

and therefore

$$\mathbb{y}_{(k)s}'\mathbb{B}_{(k)s}\mathbb{y}_{(k)s} \;\geq\; \frac{\mathbb{y}_{(k)s}'\mathbb{B}_{(k)s}\mathbb{y}_{(k)s}}{\mathbb{y}_{(k)s}'\mathbb{A}_s\mathbb{y}_{(k)s}}. \qquad \forall k,s \qquad (6.21)$$

Furthermore for vector $\mathbb{A}_s\mathbb{y}_{(k)s}(\mathbb{y}_{(k)s}'\mathbb{A}_s\mathbb{y}_{(k)s})^{-1/2}$ we have,

$$\mathbb{y}_{(k)s}'\underline{\mathbb{B}}_{(k)s}\mathbb{y}_{(k)s} \;\leq\; \frac{\mathbb{y}_{(k)s}'\mathbb{A}_s\underline{\mathbb{B}}_{(k)s}\mathbb{A}_s\mathbb{y}_{(k)s}}{\mathbb{y}_{(k)s}'\mathbb{A}_s\mathbb{y}_{(k)s}} = \frac{\mathbb{y}_{(k)s}'\mathbb{B}_{(k)s}\mathbb{y}_{(k)s}}{\mathbb{y}_{(k)s}'\mathbb{A}_s\mathbb{y}_{(k)s}}. \qquad \forall k,s \qquad (6.22)$$

Combining (6.21) and (6.22) into $\mathbb{y}_{(k)s}\mathbb{B}_{(k)s}\mathbb{y}_{(k)s} \geq \mathbb{y}_{(k)s}'\underline{\mathbb{B}}_{(k)s}\mathbb{y}_{(k)s}$ we have completed our proof for (6.19). □

6.2 Computation of the NRDA method

Before we develop in section 6.2.1 and 6.2.2 an algorithm for the maximization of the NRDA fit function (4.37) we first reformulate this function. To simplify notation we will write $\mathbb{F}$ for $\mathbb{F}(\mathbb{H})$. Instead of maximizing in (4.37) the variance of the columns of $\mathbb{P}'\mathbb{F}_{\mathbb{B}}$ accounted for by $\mathbb{V}$, we can also maximize the variance of the rows of $\mathbb{P}'\mathbb{F}_{\mathbb{B}}$ accounted for by the orthonormal basis $\mathbb{C}$. In other words we can maximize the fit function tr $\mathbb{C}'\mathbb{F}_{\mathbb{B}}'\mathbb{P}\mathbb{P}'\mathbb{F}_{\mathbb{B}}\mathbb{C}$ instead of tr $\mathbb{V}'\mathbb{P}'\mathbb{F}_{\mathbb{B}}\mathbb{F}_{\mathbb{B}}'\mathbb{P}\mathbb{V}$. The optimal discriminant space $\mathbb{P}\mathbb{V}$ of (4.37) for fixed $\mathbb{F}$ is found by the eigenvalue decomposition of $\mathbb{P}'\mathbb{F}_{\mathbb{B}}\mathbb{F}_{\mathbb{B}}'\mathbb{P}$, which produces eigenvectors $\mathbb{V}$. $\mathbb{V}$ is also defined by the equivalent eigenvalue decomposition of $\mathbb{F}_{\mathbb{B}}'\mathbb{P}\mathbb{P}'\mathbb{F}_{\mathbb{B}}=\mathbb{F}_{\mathbb{B}}'\mathbb{P}\mathbb{P}'\mathbb{P}\mathbb{P}'\mathbb{F}_{\mathbb{B}}$. The discriminant space is derived from the matrix $\mathbb{C}$ with the first p eigenvectors by unit normalizing $\mathbb{P}\mathbb{P}'\mathbb{F}_{\mathbb{B}}\mathbb{C}$. Next we decompose the projection of the mirror variates $\mathbb{F}_{\mathbb{B}}\mathbb{C}$ back on to space $\mathbb{P}$ in two parts. In figure 6.1 the Pythagorean decomposition is illustrated for one variate $\mathbb{F}\mathbb{c}$, mirror variate $\mathbb{F}_{\mathbb{B}}\mathbb{c}$ and reflected variate $\mathbb{P}\mathbb{P}'\mathbb{F}_{\mathbb{B}}\mathbb{c}=\mathbb{F}\mathbb{n}$.

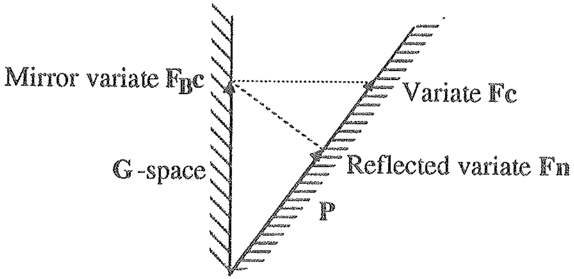

Mirror variate $\mathbb{F}_{\mathbb{B}}\mathbb{c}$
Variate $\mathbb{F}\mathbb{c}$
G-space
Reflected variate $\mathbb{F}\mathbb{n}$
$\mathbb{P}$

Figure 6.1 *Pythagorean decomposition of reflected variate* $\mathbb{F}\mathbb{n}$.

The small triangle in figure 6.1 shows the Pythagorean decomposition of $\mathbb{F}\mathbb{n}$ into $\mathbb{F}_{\mathbb{B}}\mathbb{C}-(\mathbb{F}_{\mathbb{B}}\mathbb{C}-\mathbb{F}\mathbb{N})$, with $\mathbb{F}\mathbb{n}$ orthogonal to $(\mathbb{F}_{\mathbb{B}}\mathbb{C}-\mathbb{F}\mathbb{N})$. The resulting fit function is

$$\begin{aligned}
\text{NRDA:}\quad \text{Fit}(\mathbb{C},\mathbb{N}) &= \text{tr } \mathbb{C}'\mathbb{F}_{\mathbb{B}}'\mathbb{P}\mathbb{P}'\mathbb{F}_{\mathbb{B}}\mathbb{C} \\
&= \text{tr } \mathbb{C}'\mathbb{F}_{\mathbb{B}}'\mathbb{F}_{\mathbb{B}}\mathbb{C} - (\mathbb{F}_{\mathbb{B}}\mathbb{C} - \mathbb{F}\mathbb{N})'(\mathbb{F}_{\mathbb{B}}\mathbb{C} - \mathbb{F}\mathbb{N}) \\
&= \text{tr } 2\mathbb{N}'\mathbb{F}'\mathbb{G}\mathbb{F}\mathbb{C} - \mathbb{N}'\mathbb{F}'\mathbb{F}\,\mathbb{N} \\
&= \text{tr } 2\mathbb{N}'\mathbb{B}\mathbb{C} - \mathbb{N}'\mathbb{T}\mathbb{N}, \qquad (6.23)
\end{aligned}$$

with $\mathbb{C}'\mathbb{C} = \mathbb{I}$,

where $\mathbb{F}$ is shorthand for $\mathbb{F}(\mathbb{H})$
and denotes the nonlinear transformed values of $\mathbb{H}$,
$\mathbb{C}$ denotes the variable weights of the transformed variables $\mathbb{F}$,
$\mathbb{P}$ denotes the orthonormal basis of $\mathbb{F}$, with $\mathbb{F}=\mathbb{P}\Phi\mathbb{Q}'$,
$\mathbb{F}_{\mathbb{B}}$ denotes $\mathbb{G}\mathbb{F}$, which are the between-variables of $\mathbb{F}$,
$\mathbb{F}\mathbb{N}$ denotes the non normalized discriminant space,
$\mathbb{B}$ denotes the between group variance-covariance matrix, $\mathbb{F}_{\mathbb{B}}'\mathbb{F}_{\mathbb{B}}$,
$\mathbb{T}$ denotes the total variance-covariance matrix, $\mathbb{F}'\mathbb{F}$.

6.2.1 Maximization of reformulated NRDA fit function

The maximization of NRDA (6.23) proceeds in two main alternating steps. First the parameters $\mathbb{C}$ and $\mathbb{N}$ are estimated with $\mathbb{F}$ fixed and in the next main step $\mathbb{F}$ is estimated with $\mathbb{C}$ and $\mathbb{N}$ fixed. The two steps are repeated until convergence is reached.

Estimation of $\mathbb{C}$ and $\mathbb{N}$ with $\mathbb{F}$ fixed.

We compute the singular value decomposition $\mathbb{P}'\mathbb{F}_{\mathbb{B}}=\mathbb{K}\Lambda\mathbb{L}'$, with non-zero Λ in descending order and singular vectors $\mathbb{K}'\mathbb{K}=\mathbb{L}'\mathbb{L}=\mathbb{I}$. For a p dimensional solution we have $\mathbb{C}=\mathbb{L}_p$, where $\mathbb{L}_p$ are the first p singular vectors, and $\mathbb{N}=\mathbb{Q}\Phi^{-1}\mathbb{K}_p\Lambda_p$, with $\mathbb{F}=\mathbb{P}\Phi\mathbb{Q}'$.

The corresponding discriminant space $\mathbb{P}\mathbb{V}=\mathbb{F}\mathbb{A}$ in (4.37) for fixed $\mathbb{F}$ is given by $\mathbb{P}\mathbb{V}=\mathbb{P}\mathbb{K}_p$. This implies for the reflected variates $\mathbb{P}\mathbb{P}'\mathbb{F}_{\mathbb{B}}\mathbb{C}=\mathbb{F}\mathbb{N}$, that they are a rescaling of the discriminant space $\mathbb{P}\mathbb{P}'\mathbb{F}_{\mathbb{B}}\mathbb{C}=\mathbb{P}\mathbb{K}_p\Lambda_p=\mathbb{P}\mathbb{V}\Lambda_p=\mathbb{F}\mathbb{A}\Lambda_p$. For the discriminant weights $\mathbb{A}$ we have $\mathbb{A}=\mathbb{N}\Lambda^{-1}$.

Estimation of $\mathbb{F}$ with $\mathbb{C}$ and $\mathbb{N}$ fixed.

The estimation of $\mathbb{F}$ proceeds variable-wise. Successively all variables of $\mathbb{F}$ are updated. If the variable is numerical or multiple nominal the estimation for variable k is skipped (see section 4.5.3). We maximize NRDA (6.23) with all parameters fixed except one variable $\mathbb{f}_k$ of $\mathbb{F}$. The remaining K–1 variables are fixed and gathered in matrix $\mathbb{F}_{-k}$. With $\mathbb{F}_{-k}$, $\mathbb{C}$ and $\mathbb{N}$ fixed we rewrite (6.23) in two steps into a simpler form with respect to variable $\mathbb{f}_k$.

$$\begin{aligned}\text{NRDA:}\quad \mathrm{Fit}(\mathbb{f}_k) &= \mathrm{tr}\ 2\mathbb{N}_{-k}'\mathbb{F}_{-k}'\mathbb{G}\mathbb{F}_{-k}\mathbb{C}_{-k}\\ &\quad + 2(\mathbb{f}_k'\mathbb{G}\mathbb{f}_k\mathbb{n}_k'\mathbb{c}_k + \mathbb{f}_k'\mathbb{G}\mathbb{F}_{-k}\mathbb{C}_{-k}\mathbb{n}_k + \mathbb{f}_k'\mathbb{G}\mathbb{F}_{-k}\mathbb{N}_{-k}\mathbb{c}_k)\\ &\quad - \mathrm{tr}\ \mathbb{N}_{-k}'\mathbb{F}_{-k}'\mathbb{F}_{-k}\mathbb{N}_{-k} - \mathbb{f}_k'\mathbb{f}_k\mathbb{n}_k'\mathbb{n}_k - 2\mathbb{f}_k'\mathbb{F}_{-k}\mathbb{N}_{-k}\mathbb{n}_k, \qquad (6.24)\end{aligned}$$

with $\quad \mathbb{f}_k'\mathbb{f}_k = 1.$

where $\quad \mathbb{F}_{-k}$ denotes matrix $\mathbb{F}$ with column k deleted,

$\mathbb{c}_k$ denotes a column vector with row k of matrix $\mathbb{C}$,

$\mathbf{C}_{-k}$ denotes matrix $\mathbf{C}$ with row k deleted,

$\mathbf{n}_k$ denotes a column vector with row k of matrix $\mathbf{N}$,

$\mathbf{N}_{-k}$ denotes matrix $\mathbf{N}$ with row k deleted.

We decompose the term $\mathbf{f}_k'\mathbf{G}\mathbf{f}_k=\mathbf{f}_k'\mathbf{G}\mathbf{G}\mathbf{f}_k$ into two parts by projecting variable $\mathbf{f}_k$ on to the group space $\mathbf{G}\mathbf{D}^{-1/2}$ as we did for the projection of the mirror variates $\mathbf{F}_{\mathbf{B}}\mathbf{C}$ on to space $\mathbf{P}$.

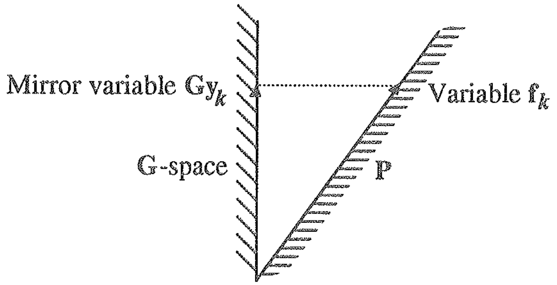


Figure 6.2 *Pythagorean decomposition of mirror variable* $\mathbf{G}\mathbf{y}_k$.

In figure 6.2 the Pythagorean decomposition $\mathbf{f}_k'\mathbf{G}\mathbf{f}_k=\mathbf{f}_k'\mathbf{f}_k-(\mathbf{f}_k-\mathbf{G}\mathbf{y}_k)'(\mathbf{f}_k-\mathbf{G}\mathbf{y}_k)$ is illustrated for variable $\mathbf{f}_k$ and mirror variable $\mathbf{G}\mathbf{f}_k=\mathbf{G}\mathbf{y}_k$. By adding parameter $\mathbf{y}_k$ we rewrite (6.24) into

$$\text{NRDA:}\quad \text{Fit}(\mathbf{f}_k,\mathbf{y}_k) = 2\mathbf{f}_k'\{\mathbf{G}\mathbf{F}_{-k}(\mathbf{C}_{-k}\mathbf{n}_k + \mathbf{N}_{-k}\mathbf{c}_k) - \mathbf{F}_{-k}\mathbf{N}_{-k}\mathbf{n}_k\}$$
$$+ 2(2\mathbf{f}_k'\mathbf{G}\mathbf{y}_k - \mathbf{y}_k'\mathbf{D}\mathbf{y}_k)\mathbf{n}_k'\mathbf{c}_k$$
$$+ \text{tr } \mathbf{N}_{-k}'\mathbf{F}_{-k}'(2\mathbf{G}\mathbf{F}_{-k}\mathbf{C}_{-k} - \mathbf{F}_{-k}\mathbf{N}_{-k}) - \mathbf{n}_k'\mathbf{n}_k, \qquad (6.25)$$

with $\mathbf{f}_k'\mathbf{f}_k = 1$,

where $\mathbf{y}_k$ denotes weights for the orthogonal group indicator matrix $\mathbf{G}$.

For $\mathbf{f}_k$ fixed maximizing NRDA($\mathbf{f}_k$,$\mathbf{y}_k$) comes down to maximizing $2\mathbf{f}_k'\mathbf{G}\mathbf{y}_k-\mathbf{y}_k'\mathbf{D}\mathbf{y}_k$. This optimization problem is equivalent to minimizing $(\mathbf{f}_k-\mathbf{G}\mathbf{y}_k)'(\mathbf{f}_k-\mathbf{G}\mathbf{y}_k)$, with $\mathbf{f}_k'\mathbf{f}_k=1$, which reaches a minimum for $\mathbf{y}_k=\mathbf{D}^{-1}\mathbf{G}'\mathbf{f}_k$.

For $\mathbf{y}_k$ fixed maximizing NRDA($\mathbf{f}_k$,$\mathbf{y}_k$) is equivalent to minimizing the residual variance $\mathbf{e}_k'\mathbf{e}_k$, where $\mathbf{e}_k=\mathbf{f}_k-(\mathbf{G}\mathbf{F}_{-k}(\mathbf{C}_{-k}\mathbf{n}_k+\mathbf{N}_{-k}\mathbf{c}_k)+2\mathbf{G}\mathbf{y}_k\mathbf{n}_k'\mathbf{c}_k-\mathbf{F}_{-k}\mathbf{N}_{-k}\mathbf{n}_k)$ and $\mathbf{f}_k$ gives the appropriate nonlinear transformation of variable $\mathbf{h}_k$. (See Gifi, 1990, page

529 and Kruskal &Carroll, 1969.) In section 4.5.3 various transformations are mentioned.

6.2.2 *The algorithm for NRDA*

Summarising in this section the preceding elaborations we define an algorithm for the maximization of the NRDA fit function (4.37).

Initialization:

Step 1. Expand $\mathbf{H}=(\mathbf{h}_1,\ldots,\mathbf{h}_k,\ldots,\mathbf{h}_K)$
into $\mathbf{H}=(\mathbf{H}_1,\ldots,\mathbf{H}_k,\ldots,\mathbf{H}_K)$, see (4.36),
with $\mathbf{H}_k = \mathbf{h}_k$ for single variables,
and $\mathbf{H}_k = \mathbf{J}\mathbf{G}_k\mathbf{D}_k^{-1/2}$ for multiple nominal variables.

Step 2. Set $\mathbf{F}=\mathbf{H}$, $i=1$, and $\Lambda_p^{i-1}=0$.

Iterations:

Step 3. Compute SVD $\mathbf{F}=\mathbf{P}\Phi\mathbf{Q}'$.

Step 4. Compute SVD $\mathbf{P}'\mathbf{G}\mathbf{F}=\mathbf{K}\Lambda\mathbf{L}'$.

Step 5. Compute $\mathbf{C}=\mathbf{L}_p$, where $\mathbf{L}_p$ are the first p singular vectors.

Step 6. Compute $\mathbf{N}=\mathbf{Q}\Phi^{-1}\mathbf{K}_p\Lambda_p^i$.

Step 7. If $(\mathrm{tr}\Lambda_p^i - \mathrm{tr}\Lambda_p^{i-1}) > \varepsilon$, for some small ε, then stop.

Step 8. Minimize $\mathbf{e}_k'\mathbf{e}_k$ for single non-numerical vars, $\forall k$ successive,
with $\mathbf{e}_k=\mathbf{f}_k^i-\{\mathbf{G}(\mathbf{F}_{-k}(\mathbf{C}_{-k}\mathbf{n}_k + \mathbf{N}_{-k}\mathbf{c}_k)+2\mathbf{f}_k^{i-1}\mathbf{n}_k'\mathbf{c}_k)-\mathbf{F}_{-k}\mathbf{N}_{-k}\mathbf{n}_k\}$,
where $\mathbf{f}_k$ gives the appropriate unit normalized nonlinear transformation of variable $\mathbf{h}_k$ and $\mathbf{F}_{-k}$ is updated before all successive steps.

Step 9. Set, $i=i+1$ and go to Step 3.

The parameter $\mathbf{y}_k$ derived from the maximization of (6.25) is incorporated in Step 8 in the 'old' nonlinear transformation $\mathbf{f}_k^{i-1}$ of variable $\mathbf{h}_k$.

Chapter 7

EXAMPLES

We present analyses of real-life data using three methods developed in the preceding chapters. For a psychometric application of Set Component Analysis (chapter 3) we compare the SCA solution of the Miller-Nicely data with the corresponding INDSCAL solution. Reflected Discriminant Analysis from chapter 4 is applied on mass spectrometric barley tissue profiles and compared with results for PC-DA. The barley tissue profiles are also analysed with Nonlinear Reflected Discriminant Analysis.

Introduction

Although a wide range of methods have been presented in the preceding chapters we give only a modest number of real-life applications. Several considerations lead to this approach. Many methods which have been discussed are well-known, and although there was up to now no overall criterion for the PLS methods, all these methods have already been applied for many years. For the 'new' methods like SCA and RDA we already gave a fairly diverse impression of their properties by simulation studies. The number of new methods that can be generated with DC or LDC (chapter 5) is so large that a separate future treatment of corresponding applications is justified. For the moment we confine ourselves to present in the next three main sections real-life examples for respectively SCA, RDA and NRDA. The analyses in this chapter were performed by programming all involved methods in APL (*A Programming Language*).

7.1 SCA and INDSCAL on psychometric Miller-Nicely data

In Soli & Arabie (1979) the utility of phonetic features versus acoustic properties for describing perceptual relations among speech sounds was evaluated with a multidimensional scaling analysis of the consonant confusions data of Miller & Nicely (1955). A general review of the many analyses this classic dataset has supported is given by Shepard (1987). In section 7.1.1 we introduce the experimental data. In section 7.1.2 we present some details on transformation, normalization and symmetrization applied by Soli & Arabie. The computed INDSCAL solution is briefly discussed. In section 7.1.3 the SCA solution is presented and compared to INDSCAL results.

7.1.1 Experimental data

The data from Miller & Nicely's experiment consist of full 16×16 matrices $\mathbb{C}_k$ of identification confusions between 16 consonant phonemes obtained in K=17 different listening conditions. Four subjects listened while a fifth subject served as a speaker, reading lists of consonant-vowel syllables formed by pairing the consonants /p, t, k, f, θ, s, ʃ, b, d, g, v, ∂, z, ζ, m, n/ with the vowel /a/. (The phonemes /θ/, /ʃ/, /∂/, and /ζ/ are respectively pronounced as in *th*in, *sh*awl, *th*at, *Zh*ivago and the vowel /a/ as in f*a*ther.) The subjects rotated as speakers and listeners within each experimental condition k. The listeners recorded the consonant they had heard after each syllable was spoken. The consonants are classified by phoneme features in five groups shown in table 7.1.

Table 7.1 *Phoneme features of 16 consonants.*

	Stops	Fricatives	Nasals
Voiceless	/p, t, k/	/f, θ, s, ʃ/	
Voiced	/b, d, g/	/v, ∂, z, ζ/	/m, n/

The 17 experimental listening conditions are summarized in table 7.2 and may be classified under three general headings. First were the noise-masking conditions, in which only the signal-to-noise (S/N) ratio changed. The S/N ratio was manipulated by varying the amplitude of random noise which had been low-pass filtered at 6500 Hz. Second were the low-pass conditions, in which a constant S/N ratio of 12 dB was maintained while the speech was low-pass filtered at the cutoff frequencies given in table 7.2. The final conditions were high-pass, in which the same constant S/N ratio of 12 dB was again maintained while the speech channel was high-pass filtered at the cutoff frequencies also given in table 7.2. For each condition we computed the efficient rank $I_k m_k$ of the INDSCAL scalar product matrix according to (2.17). The efficient rank gives an indication of the efficiency of information transfer by estimating the number of reliable dimensions. By adding more noise or by narrowing the filter bandwidth the efficient rank gradually goes down. Only the efficient rank of N6 is remarkably high. The mean efficient rank over all conditions is 5.6 and the minimum is 4. Therefore the computation of a four dimensional solution is advisable, if one wants to find as many reliable dimensions as possible which are common to all conditions.

Table 7.2 *Listening conditions.*

Condition heading	Label	Efficient Rank $I_k m_k$	Speech-to-noise ratio (dB)	Bandwidth (Hz)
Noise masking	N1L1	7.8	12	200-6500
	N2	6.0	6	200-6500
	N3	5.5	0	200-6500
	N4	4.7	–6	200-6500
	N5	4.1	–12	200-6500
	N6	7.2	–18	200-6500
Low-pass filtering	L2H1	6.2	12	200-5000
	L3	5.9	12	200-2500
	L4	5.5	12	200-1200
	L5	5.4	12	200-600
	L6	5.3	12	200-400
	L7	4.0	12	200-300
High-pass filtering	H2	5.9	12	1000-5000
	H3	5.4	12	2000-5000
	H4	6.0	12	2500-5000
	H5	5.6	12	3000-5000
	H6	4.1	12	4500-5000

7.1.2 *INDSCAL analysis*

Soli & Arabie (1979) employed the INDSCAL method and program with the original data $\mathbb{C}_k$ log transformed to enhance consistency with the linear INDSCAL model. The actual normalization and symmetrization applied by Soli & Arabie were recovered in steps. With the Appendix of Arabie & Soli (1982) and Shepard (1972) we reconstructed the formula for deriving from the original confusion data $\mathbb{C}_k$ the INDSCAL input similarity matrices $\mathbb{V}_k$.

$$(\mathbb{V}_k)_{ij} = {}^{10}\log\,\Big(\frac{(\mathbb{C}_k)_{ij} + (\mathbb{C}_k)_{ji}}{(\mathbb{C}_k)_{ii} + (\mathbb{C}_k)_{jj}} + 0.001\Big) \qquad (7.1)$$

where $(\mathbb{C}_k)_{ij}$, denote the elements of the confusion matrix $\mathbb{C}_k$,
$(\mathbb{V}_k)_{ij}$ denote the elements of the INDSCAL input similarity matrices $\mathbb{V}_k$.

The (7.1) log transformation of the Miller-Nicely data was verified in Appendix A of Arabie, Carroll & DeSarbo (1987), where some of the log transformed values are listed. The scalar products matrices $\mathbb{S}_k$ were derived from the similarity matrices $\mathbb{V}_k$ by the following formula.

$$\mathbb{S}_k = -1/2\mathbb{J}(c_k(\mathbb{1}\mathbb{1}'-\mathbb{I}) - \mathbb{V}_k)^2\mathbb{J} \tag{7.2}$$

where $\mathbb{J}$ denotes the centring matrix $(\mathbb{I} - \mathbb{1}(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}')$,
c_k denotes the maximum of $(-\mathbb{V}_k)_{ij} - (-\mathbb{V}_k)_{il} - (-\mathbb{V}_k)_{lj}$, $\forall i,j,l$.

> The additive constant c_k gives an estimate of the smallest constant approximating satisfaction of the triangle inequality $d_{ij} \leq d_{il} + d_{lj}$, $\forall i,j,l$, with $d_{ij} = c_k + (-\mathbb{V}_k)_{ij}$. The additive constant method applied in INDSCAL is described in Torgerson (1958, pp. 276-277). The resulting scalar product matrices $\mathbb{S}_k$ have large positive eigenvalues. The small eigenvalues are distributed about zero and are assumed to be 'error' dimensions.

The derived scalar products matrices (7.2) were analysed matrix conditional, which is the default option in the INDSCAL program. Therefore these matrices were multiplied by a normalizing constant required to set the sum of squares for each matrix $\mathbb{S}_k$ equal to unity. So $\mathbb{S}_k = \mathbb{S}_k(\mathrm{tr}\mathbb{S}_k\mathbb{S}_k)^{-1/2}$, $\forall k$. The INDSCAL dimensions and weights for the unit normalized matrices $\mathbb{S}_k$ were computed according to the INDSCAL procedure of Carroll & Chang (1970). Our results are nearly equal to the results presented in Soli & Arabie (1979). On the basis of interpretability and only slight increments in the INDSCAL fit for dimension five and six they decided to choose the four dimensional solution as most appropriate for describing the perceptual relationships between the 16 consonants. This is in agreement with the 'four dimension' advice in section 7.1.1 based on the efficient rank. We shall refer to the INDSCAL solution of Soli & Arabie as the *original* solution. Arabie, Carroll & DeSarbo (1987) also report slight differences in their 1987 reanalysis compared to the original solution computed in 1976 and suggest this is probably due to the use of different hardware. Their reanalysed proportion of variance accounted for was 0.6907 for the four dimensional INDSCAL solution and equal to our computed total proportion of VAF. The original value was 0.6922.

> The total proportion of VAF for scalar product matrices S_k according to Arabie, Carroll & DeSarbo (1987) is computed by
>
> $$VAF_{\text{prop}} = 1 - \frac{\sum\limits_{k=1}^{K} \mathrm{tr}(XW_kX'-S_k)(XW_kX'-S_k)}{\sum\limits_{k=1}^{K} \mathrm{tr}S_kS_k}. \qquad (7.3)$$
>
> The proportion of VAF for each matrix S_k is computed by restricting the summation in the numerator and denominator to apply only to data from each single source.

Nevertheless they presented the original figures in order to maintain consistency with earlier published accounts. We present our figures in table 7.3.A and 7.6.A, mainly because the proportions of VAF by each dimension are not at all like the values originally presented. Due to these differences dimension 3 and 4 are interchanged.

Table 7.3 *INDSCAL and SCA weights and proportion of variance accounted for*

	A:*Our four dim. INDSCAL solution*					B:*Four dim. SCA solution*				
Filter Label	4.1	4.2	4.3	4.4	VAF_{prop}	4.1	4.2	4.3	4.4	VAF_{prop}
N1L1	0.36	0.41	0.26	0.43	0.59	0.48	0.37	0.32	0.29	0.55
N2	0.46	0.54	0.25	0.39	0.75	0.64	0.36	0.37	0.34	0.79
N3	0.51	0.56	0.20	0.37	0.81	0.71	0.35	0.35	0.26	0.82
N4	0.60	0.54	0.17	0.19	0.77	0.78	0.29	0.32	0.11	0.81
N5	0.73	0.43	0.19	0.20	0.84	0.76	0.32	0.37	0.07	0.83
N6	0.49	0.41	0.10	0.15	0.47	0.51	0.18	0.42	0.14	0.49
L2H1	0.40	0.55	0.27	0.40	0.74	0.55	0.39	0.45	0.29	0.74
L3	0.45	0.53	0.24	0.42	0.77	0.59	0.39	0.44	0.28	0.77
L4	0.54	0.52	0.10	0.33	0.72	0.66	0.25	0.42	0.25	0.74
L5	0.52	0.59	0.13	0.28	0.76	0.69	0.27	0.41	0.25	0.78
L6	0.69	0.41	0.16	0.22	0.77	0.68	0.25	0.38	0.20	0.71
L7	0.65	0.52	0.05	0.08	0.76	0.76	0.10	0.35	0.14	0.73
H2	0.32	0.37	0.38	0.45	0.63	0.48	0.51	0.17	0.30	0.61
H3	0.37	0.15	0.55	0.29	0.57	0.34	0.55	0.18	0.09	0.46
H4	0.25	0.21	0.56	0.29	0.54	0.28	0.59	0.16	0.25	0.51
H5	0.19	0.10	0.69	0.25	0.62	0.16	0.61	0.16	0.24	0.48
H6	0.06	0.08	0.77	0.12	0.63	0.06	0.59	0.07	0.28	0.43
VAF_{prop}	0.26	0.22	0.15	0.10	0.69	0.33	0.16	0.11	0.06	0.66

We computed the VAF_{prop} dimension-wise by substituting $x_s w_{(k)s} x_s$ instead of XW_kX' in (7.3) for each dimension s separately, where $w_{(k)s}$ gives the appropriate diagonal value of W_k. The four VAF_{prop} by dimension do not sum to the total of 0.69, because the dimensions are not orthogonal. To obtain the sequence of the original solution the four dimensions labeled by their VAF_{prop} should be permutated to 0.26 0.22 0.10 and 0.15. In the original published table Soli & Arabie gave the values 0.33 0.13 0.16 and 0.07. These VAF_{prop} values are rather misleadingly derived from the total VAF_{prop} of lower dimensional solutions. In table 7.4.A we give the original VAF_{prop} with increments up to a 6 dimensional solution. In table 7.4.B we give our corresponding reanalysed results.

Table 7.4 *INDSCAL fit for different dimensionalities.*

	A:*Original INDSCAL solutions*		B:*Our INDSCAL solutions*	
Dimensionality	VAF_{prop}	Increment	VAF_{prop}	Increment
1	0.33		0.34	
2	0.46	0.13	0.50	0.16
3	0.62	0.16	0.62	0.12
4	0.69	0.07	0.69	0.07
5	0.73	0.04	0.74	0.05
6	0.77	0.04	0.77	0.03

The most striking difference is the fit of the two dimensional INDSCAL solutions. In the original analysis the iteration process is stopped too early compared to our reanalysed results. Apart from this minor practical error the matching of the values of table 7.4.A to the corresponding dimensions of the original four dimensional solution is theoretically dubious. We illustrate this in table 7.5 where correlations are given between the dimensions of our 1, 2 (dim. 2.1 and 2.2) and 3 (dim. 3.1, 3.2 and 3.3) dimensional INDSCAL solution with our 4 dimensional INDSCAL and SCA solution. Redundant zeros are left out in this table.

Table 7.5 *Correlations with four dimensional INDSCAL and SCA solution.*

INDSCAL	A:*Our four dim. INDSCAL solution*				B:*Four dim. SCA solution*			
Dimensions	4.1	4.2	4.3	4.4	4.1	4.2	4.3	4.4
1	-0.84	0.74	-0.03	-0.07	1	-0.03	-0.01	0.04
2.1	-0.84	0.75	-0.01	-0.09	1	-0.01	0	0.04
2.2	0.18	0.11	0.91	-0.59	-0.02	0.98	0.03	0.03
3.1	0.99	-0.32	0.06	0.13	-0.86	0.08	0.49	-0.02
3.2	-0.21	1	0.04	0.04	0.71	0.07	0.7	0.05
3.3	0.11	0.06	0.92	-0.61	0	0.98	-0.07	0.02

First we observe in table 7.5.A that our four dimensional solution gives the same sequence of dimension 3 and 4, if we use the correlations with the three dimensional solution as a criterion to classify the fourth dimension. In the original Soli & Arabie solution our third dimension is classified as the fourth dimension. Secondly the correlations with the one and two dimensional solution in table 7.5 show that the unique orientation of the INDSCAL dimensions is not consistent over comparable solutions, because it is dependent on the dimensionality of the solution. Therefore the *VAF* values presented in table 7.4 should not be substituted in table 7.3.

Table 7.6 *INDSCAL and SCA consonant dimensions.*

	A:*Our four dim. INDSCAL solution*				B:*Four dim. SCA solution*			
Consonants	4.1	4.2	4.3	4.4	4.1	4.2	4.3	4.4
/p/	0.275	-0.304	-0.151	-0.089	-0.357	-0.089	-0.044	-0.255
/t/	0.317	-0.289	-0.057	-0.236	-0.362	0.078	-0.001	-0.414
/k/	0.283	-0.329	-0.097	-0.198	-0.363	0.017	-0.113	-0.346
/f/	0.255	-0.097	-0.110	0.359	-0.230	-0.270	0.171	0.129
/θ/	0.248	-0.050	-0.185	0.267	-0.207	-0.178	0.189	0.180
/s/	0.245	0.021	0.206	0.098	-0.174	0.238	0.193	0.442
/ʃ/	0.195	-0.089	0.713	-0.001	-0.200	0.507	-0.017	0.468
/b/	-0.107	0.180	-0.202	0.395	0.135	-0.376	0.196	0.134
/d/	-0.098	0.265	-0.159	-0.386	0.245	0.126	0.075	-0.124
/g/	-0.131	0.307	-0.119	-0.276	0.275	0.104	0.131	-0.108
/v/	-0.160	0.232	-0.132	0.355	0.198	-0.303	0.163	0.064
/∂/	-0.137	0.269	-0.116	0.153	0.250	-0.173	0.158	-0.073
/z/	-0.123	0.342	0.068	-0.165	0.292	0.134	0.159	-0.150
/ζ/	-0.180	0.199	0.490	-0.335	0.281	0.464	-0.045	-0.237
/m/	-0.432	-0.338	-0.062	0.120	0.090	-0.192	-0.572	0.194
/n/	-0.449	-0.321	-0.087	-0.061	0.128	-0.089	-0.645	0.096

The INDSCAL dimensions in table 7.6.A are displayed graphically in figure 7.1 and the corresponding weights from table 7.3 in figure 7.2. Dimension 3 and 4 are presented in the same orientation as the original publication in order to facilitate a visual comparison.

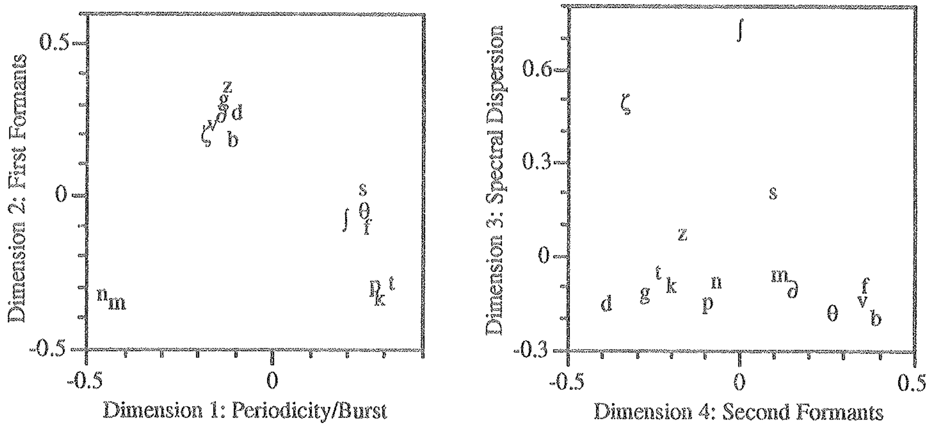


Figure 7.1 *INDSCAL dimensions mapping consonants of Miller-Nicely data.*

We give a summary extracted from Arabie, Carroll & DeSarbo (1987) of the interpretation of the four dimensions. For readers not familiar with phonetics it is useful to know that the vocal tract resonates at overtone frequencies. These resonant frequencies are known as the *formants.*

"The first dimension of the object space appears to specify the temporal relationship between onset of periodic formant resonance and the initiation of broadly dispersed acoustic energy. An attempt to capture this generality led Soli & Arabie to select the abbreviated label 'periodicity/burst order'. In choosing a label for the second dimension, the perceptual weights for this dimension in all listening conditions were also examined (see Table 7.3 and figure 7.2). The pattern of weights implied that the second dimension specified spectral changes in the lower portion of the speech spectrum that are excited by relatively large amounts of acoustic energy, corresponding to 'first formant transitions', which becomes the label for the second dimension.

The fourth dimension of figure 7.1 seems to specify the shape of voiced second formant transitions in the syllables, and resembles the dimension in Wish's analysis labeled 'second formant transitions'. That label has been retained in the current analysis. The arrangement of the phonemes on the third dimension corresponds quite well to the amount of spectrally dispersed acoustic energy located below 5 kHz in the speech spectrum. Because of this correspondence, the dimension has been given the label 'spectral dispersion'.

Perhaps the most succinct summary of this object space is to note that acoustic properties rather than phonetic features gave the most interpretable account of the dimensions. This conclusion differs from previous analyses and runs counter to traditional theorizing by some phoneticians."

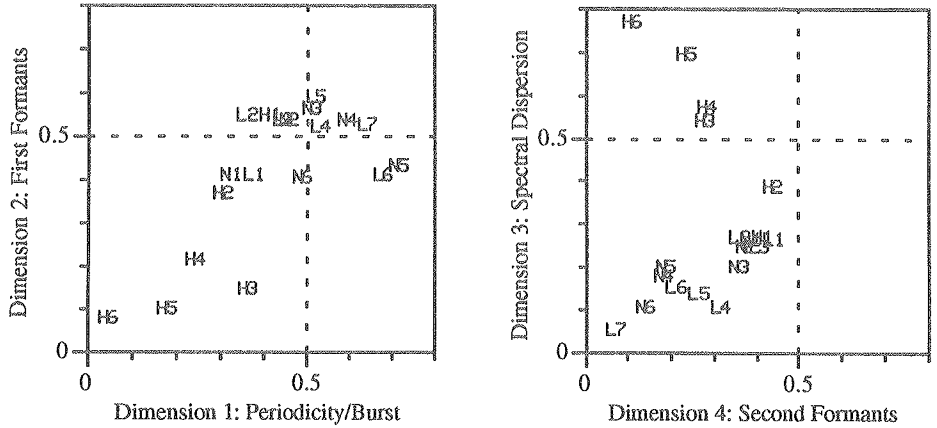


Figure 7.2 *INDSCAL weights mapping filter conditions of Miller-Nicely data.*

The INDSCAL weights in figure 7.2 reveal a simple structure as generally predicted in chapter 3. Some filter conditions load high and others very low on the respective dimensions, but there is no gradual transition of the weights from one dimension to another for instance in the form of a quarter circle. According to many introductions in INDSCAL theory this differential weighting is an attractive feature of the model. In chapter 3 we also argued that INDSCAL solutions will be dominated by sets with a low efficient rank, if the sets are normalized to the same total sum of squares. The correlation between the efficient ranks in table 7.2 and the VAF_{prop}'s for the listening conditions in table 7.3.A is -0.53, which is really far from zero and confirms the theoretical results.

7.1.3 SCA analysis

The consonant dimensions of the SCA solution are listed in table 7.6.B and graphically displayed in figure 7.3. The weights for the SCA dimensions are computed according to the INDSCAL procedure of Carroll & Chang (1970) and referred to as SCA weights. The SCA weights are listed in table 7.3.B and graphically displayed in figure 7.5. An impression of the matching between the SCA and INDSCAL consonant *space* is given by the canonical correlations between these two sets with four variables. The canonical correlations 0.999, 0.997, 0.996 and 0.771 indicate an almost complete overlap in three dimensions. Table 7.5 shows that the orientation of the SCA dimensions differs from the four dimensional INDSCAL solution. The orientation of the first two SCA dimensions is more or less equal to the two dimensional INDSCAL solution and therefore clearly different from the orientation of the first two dimensions of the four dimensional INDSCAL solution. The interpretation of the SCA consonant dimensions changes considerably due to the different orientation within the INDSCAL space and some change outside this space.

The first SCA dimension in figure 7.3 separates the voiceless (unvoiced) consonants from the voiced consonants as can be verified in table 7.1 and is labeled 'voicing'. Four of the five phonetic groups in table 7.1 are separated by the first SCA dimension. Only the voiced stops and the voiced fricatives are not distinguished.

For the interpretation of the second dimension we provide some additional information. The most widely used set of symbols for phonetic transcription is that of the International Phonetic Association (IPA).

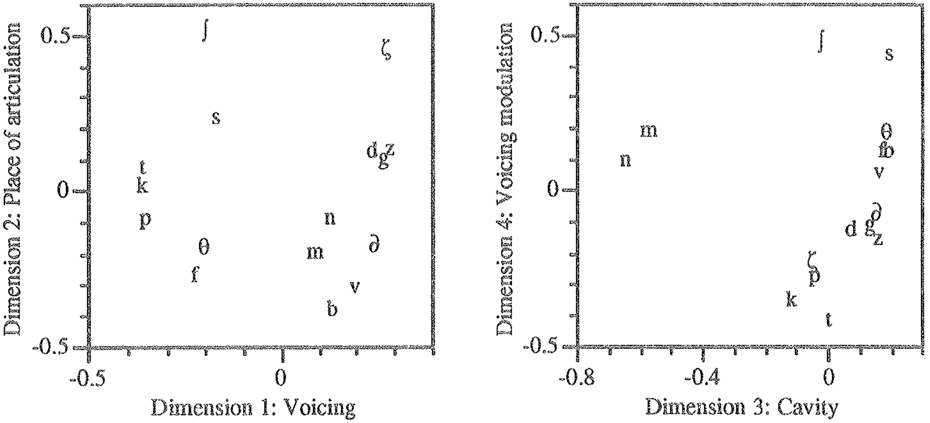


Figure 7.3 *SCA dimensions mapping consonants of Miller-Nicely data.*

From the IPA chart given in the Encyclopaedia Britannica (15th edition, 1984) we extracted the classification of the 16 involved consonant phonemes. The extracted chart is given in table 7.7.

Table 7.7 *Phoneme features of 16 consonants.*

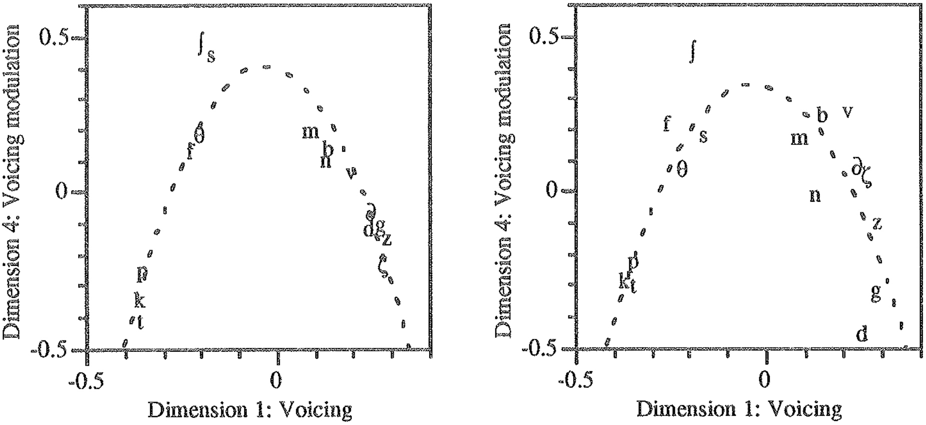
Place of articulation	Voiceless stops	Voiceless fricatives	Nasals	Voiced stops	Voiced fricatives
Velar	/k/			/g/	
Palato alveolar		/ʃ/			/ʒ/
Alveolar	/t/	/s/	/n/	/d/	/z/
Dental		/θ/			/∂/
Labio-dental		/f/			/v/
Bilabial	/p/		/m/	/b/	

We split the IPA 'dental and alveolar' group in a 'dental' and an 'alveolar' group according to supplemental information in the Encyclopaedia Britannica. The place of articulation is ordered from back (velar) to front (bilabial). The columns are ordered in such a way to facilitate a comparison with the first two SCA dimensions in figure 7.3. The second dimension separates 'back' consonants from 'front' consonants and is labeled 'place of articulation'. Above -0.89 we find velar, palato alveolar and alveolar consonants and below -0.89 we find dental, labio-dental and bilabial consonants. The place of articulation of all fricatives (Table 7.7) is perfectly ordered by the second

SCA dimension. The same holds perfectly for the nasals, but not perfectly for the stops. The velar articulation points are located slightly too much to the front. Apparently the velar and alveolar stops are not discriminated with this place of articulation dimension.

The third SCA dimension separates nasal consonants from oral consonants and is labeled 'cavity'. We could also have used the label 'nasality', but this label interferes too much with the phonetic meaning of nasality and the corresponding ordering of consonants.

The fourth SCA dimension (SCA_4) can be quite well predicted with a quadratic function of the first dimension and is labeled by 'voicing modulation'. The correlation between SCA_4 and $(-6.392 \times SCA_1^2 - 0.407 \times SCA_1 + 0.400)$ is 0.908. We plotted the first and fourth dimension in figure 7.4.A to show the functional relation visually. The interpretation of the 'voicing modulation' dimension is not univocal. It can be an artefact of analyzing nonlinear data with a linear technique or it can be that listeners are apparently able to discriminate neutral voicing from extreme voicing. Anyway the voiceless fricatives are separated extra from the other consonants by this functional relation. It is interesting to notice that the four dimensional INDSCAL solution also contained this functional relation, but less pronounced. To show the relation we predicted the SCA_1 and the SCA_4 from the INDSCAL space with multiple linear regression. The multiple correlations were respectively 0.999 and 0.785.



A *SCA consonants with curve fitting.* **B** *INDSCAL-SCA approximation.*

Figure 7.4 *Dimension 1 and 4 mapping consonants of Miller-Nicely data.*

We obtained the consonant dimensions IND_{SCA1} and IND_{SCA4} and plotted these INDSCAL approximations of the first and the fourth SCA dimension in figure 7.4.B. The correlation between IND_{SCA4} and ($-5.524\times\mathrm{IND}^2_{SCA1}-0.352\times\mathrm{IND}_{SCA1}+0.346$) is 0.761. We remark that optimization of the orientation of the first and the fourth dimension with respect to a quadratic relation might improve the correlation for both SCA and IND_{SCA}. Secondly the prediction of the fourth dimension by the first dimension can be improved by applying other simple nonlinear functions in figure 7.4. For instance a separate linear regression for the voiced and the voiceless consonants results in correlations 0.962 and 0.754 between true and predicted fourth dimension for respectively SCA and IND_{SCA}.

Summarizing, the SCA consonant dimensions can satisfactorily be interpreted with phonetic features, contrary to the interpretation of the original INDSCAL dimensions with acoustic properties by Soli & Arabie.

The SCA weights in figure 7.5 show a nice gradual transition for the first two dimensions in the form of a quarter circle and offer an almost ideal example of differential weighting of dimensions. Generally high-pass filtering conditions result in better than average discrimination of place of articulation and worse than average discrimination of voicing. This tendency is reversed for low-pass filtering and noise masking conditions. The same transition not compared to average discrimination but by measuring relative discrimination can be observed for voicing modulation and cavity. In table 7.8 the weight ratio's for dimensions 1/2 and 3/4 are computed.

Table 7.8 *Weight ratio's for dimension 1 divided by 2 and 3 divided by 4.*

Filter Label	1/2	3/4	Filter Label	1/2	3/4	Filter Label	1/2	3/4
N1L1	1.28	1.10	L2H1	1.41	1.56	H2	0.94	0.59
N2	1.78	1.07	L3	1.50	1.57	H3	0.62	2.02
N3	2.06	1.34	L4	2.64	1.71	H4	0.47	0.62
N4	2.72	2.87	L5	2.58	1.67	H5	0.26	0.67
N5	2.38	5.36	L6	2.68	1.86	H6	0.11	0.26
N6	2.78	2.95	L7	7.69	2.49			

All noise masking and low-pass filtering conditions discriminate better on dimension 1 and 3 compared to respectively dimension 2 and 4 than the high-pass filtering conditions except for H3(3/4). Low-pass filtering and noise masking tend to increase gradually the discrimination of dimension 1 and 3 compared to respectively dimension 3 and 4. For high-pass filtering this tendency is reversed.

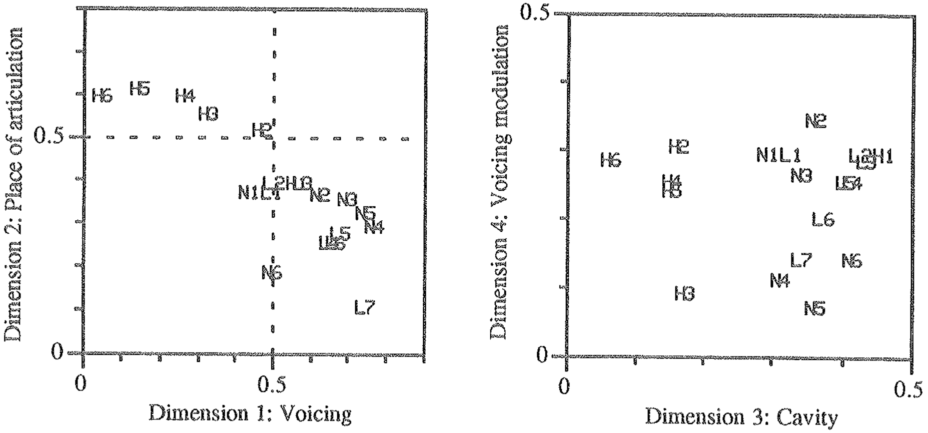


Figure 7.5 *SCA weights mapping filter conditions of Miller-Nicely data.*

In summary the analysis of the Miller-Nicely data with INDSCAL and SCA seems to confirm the theoretical expectations.

- The orientation of the SCA dimensions appears simpler to interpret than the INDSCAL orientation. Although this could be expected, because SCA eliminates as much as possible the unique components in the listening conditions, it remains to be seen in future if this property is repeated for other real-life examples.
- The configuration of the SCA weights shows a gradual transition from one dimension to another and approximates more to the concept of differential weighting of dimensions. The INDSCAL weights are more grouped in bundles as is usual for simple structure configurations.
- The SCA solution is not dominated by sets with low efficient rank. The correlation between the VAF_{prop} and the efficient rank for all listening conditions is -0.26. For the INDSCAL solution this correlation is -0.53, which implies relatively high loadings for listening conditions with low efficient rank.

7.2 RDA and PC-DA on mass spectrometric barley tissue profiles

In Tas, Angelino, La Vos & van der Greef (1991) barley tissue profiles are analysed with PC-DA, which is frequently applied in chemometrics to pyrolysis profiles.

Pyrolysis stands for the thermal degradation of usual complex (bio)chemical systems like micro-organisms, cells, cell walls, food, soil, plant materials, fossil deposits, body fluids and tissues.
In section 7.2.1 we introduce the experimental data. In section 7.2.2 the PC-DA solution is presented. Next the barley tissue profiles are analyzed with RDA in section 7.2.3. We also provide Leaving-One-Out (L-O-O) error rates (4.31) to compare group prediction of RDA with PC-DA.

7.2.1 *Experimental data*

Six tissue elements (husk, aleurone, endosperm, scutellum, radicle and coleoptil) were prepared from the barley variety Triumph. Samples were obtained at the beginning of the malting process from starting material (day 0), after four days of germination (day 4) and after six days of germination and subsequent kilning (day 7). Py-DCI/MS (pyrolysis-direct chemical ionization/mass spectrometry) was performed with the number of MS measurements on each tissue sample listed in table 7.9.

Table 7.9 *Number of MS measurements on barley tissue samples.*

Seed particle	Native barley	4th day of germination	After kilning (7th day)
Husk	3	0	3
Aleurone	6*	3	3
Endosperm	3	3	3
Scutellum	3	3	3
Radicle	3	3	3
Coleoptil	3	6*	6*

n=3 for each cell, cells marked with an asterisk (*) are sampled twice: n=6

The 60 measured spectra were normalized to total ion current to correct for differences in sample size. The resulting patterns were reduced to subsets of 235 variables, the highest Fisher weights being the selection criterion. Finally the 235 variables were transformed to unit normalized variables in deviations from their mean. We refer to the rows of the resulting 60×235 matrix as the *barley tissue profiles.*

The Between-to-Total ratio *BT* (4.27) of the barley tissue profiles for the six tissue groups is 0.26. It should be noticed that *BT* was more or less maximized in the preprocessing step, where variables with the highest Fisher weights were selected (Fisher, 1936). We computed the efficient rank *Im* of the barley tissue profiles according to (2.17). The efficient rank is *Im*=17.4, which is close to the total number of 17 measured cells in table 7.9.

7.2.2 Principal Component - Discriminant Analysis

To explore the differences in MS pattern between the six tissue elements the 60 barley tissue profiles are partitioned into six groups. Differences in sampling time are neglected. For the final PC-DA solution the 60×235 matrix was first reduced with PCA to rank 9. Next the DA solution was computed in a second step. We refer to the resulting solution as the PC9-DA solution, because of the rank 9 reduction in the first step. The 60 objects in the PC9-DA discriminant space are plotted in figure 7.6, where the objects are labeled by the first letter of their group tissue name.

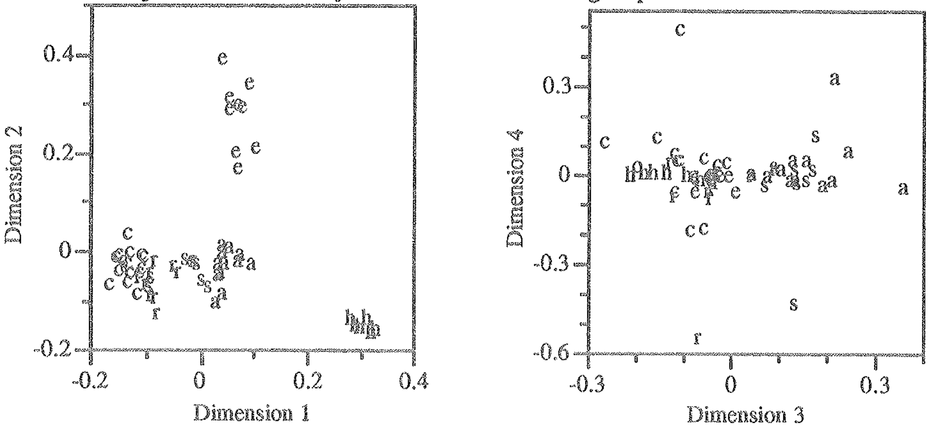


Figure 7.6 *PC9-DA discriminant space mapping 60 barley tissue profiles.*

The loadings are given in figure 7.7. Only loadings outside a circle with radius 0.5 are displayed. We will not elaborate on the interpretation of the mass numbers. Our first two PC9-DA dimensions rotated 45 degrees are consistent with the PC-DA results of Tas, Angelino, La Vos & van der Greef (1991). They considered the replicates as groups (20 groups, see table 7.9) and reduced the spectra to subsets of 57 variables.

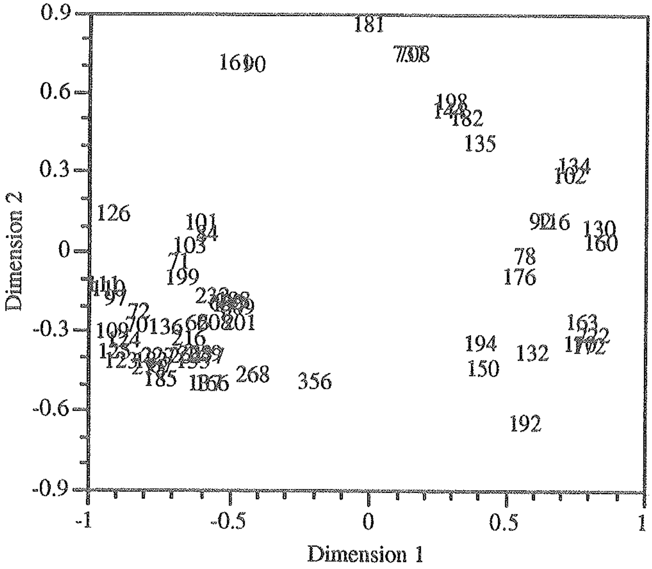


Figure 7.7 *PC9-DA discriminant space mapping barley loadings.*

7.2.3 *Reflected Discriminant Analysis*

The discriminant space of the RDA solution is graphically displayed in figure 7.8.

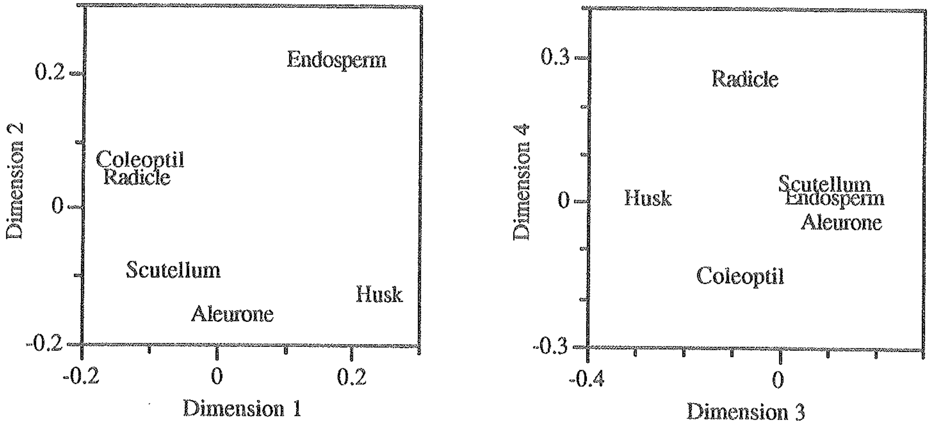


Figure 7.8 *RDA discriminant space mapping 60 barley tissue profiles.*

The within-group variance has completely vanished and therefore all objects within one group are positioned exactly on one group point. The first RDA dimension is very similar to the first PC9-DA dimension. The loadings for the first two dimensions are given in figure 7.9. Only loadings outside a circle with radius 0.5 are displayed.

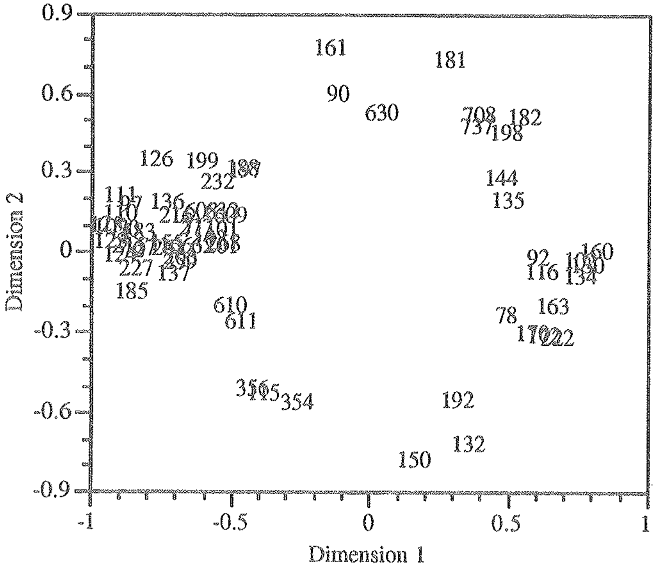


Figure 7.9 *RDA discriminant space mapping barley loadings.*

Although the RDA results simplify remarkably compared to PC-DA, this does not imply that group prediction is also improved. Therefore we have computed L-O-O error rates (4.31) for assessing prediction in several dimensions for both PC-DA and RDA. The L-O-O error rate is a measure of misclassification of group prediction. In table 7.10 we present five criteria for comparing PC-DA, RDA and NRDA solutions. NRDA results will be discussed in section 7.3. The first column in table 7.10 gives the number of dimensions of the discriminant space. The second column gives the squared canonical correlation $\rho^2_{\mathrm{CVA}} = \mathbf{v}_s'\mathbb{P}'\mathbb{P}_{\mathbb{B}}\mathbb{P}_{\mathbb{B}}'\mathbb{P}\mathbf{v}_s$ between each dimension s of the discriminant space $\mathbb{P}\mathbb{V}$ and the corresponding projection on the group space. The squared correlations ρ^2_{CVA} are the diagonal values of $\mathbb{V}'\mathbb{P}'\mathbb{P}_{\mathbb{B}}\mathbb{P}_{\mathbb{B}}'\mathbb{P}\mathbb{V}$ in $\mathrm{CVA}_{\mathrm{BT}}$ (4.13). The other four criteria are the proportion of variance accounted for, VAF_{prop}, the proportion of reflected variance accounted for, $RVAF_{\mathrm{prop}}$, the L-O-O error rate for 6 groups, *LOO6* and the L-O-O error rate for 5 groups computed with the 6 group solutions, *LOO5*, where radicle and coleoptil are merged to one group. It is interesting to realize that the Between-to-Total ratio *BT* (4.27) gives an upper bound for $RVAF_{\mathrm{prop}}$ with $RVAF_{\mathrm{prop}} \leq BT$. The upper bound can for instance be reached in the complete rank case as has been formulated in section 4.3.7. Because the barley tissue profiles are a complete rank case, the upper bound of 0.26 is reached with five dimensions.

Table 7.10 *A comparison of PC9-DA, RDA and NRDA solutions.*

Number of dimensions		ρ^2_{CVA}	VAF_{prop}	$RVAF_{\mathrm{prop}}$	*LOO6* error rate	*LOO5* error rate
1	PC9-DA	0.96	0.14	0.13	0.37	0.28
1	RDA	1.00	0.15	0.15	0.38	0.22
1	NRDA	1.00	0.21	0.21	0.38	0.33
2	PC9-DA	0.92	0.21	0.19	0.22	0.13
2	RDA	1.00	0.20	0.20	0.25	0.12
2	NRDA	1.00	0.30	0.30	0.25	0.20
3	PC9-DA	0.78	0.26	0.22	0.23	0.13
3	RDA	1.00	0.24	0.24	0.22	0.05
3	NRDA	1.00	0.37	0.37	0.08	0.02
4	PC9-DA	0.11	0.32	0.22	0.23	0.12
4	RDA	1.00	0.25	0.25	0.15	0.07
4	NRDA	1.00	0.40	0.40	0.12	0.07
5	PC9-DA	0.03	0.37	0.22	0.23	0.12
5	RDA	1.00	0.26	0.26	0.13	0.05
5	NRDA	1.00	0.42	0.42	0.12	0.03

The squared correlations ρ^2_{CVA} for each dimension show that the fourth and fifth PC9-DA dimension have almost no discriminating power between groups, whereas the RDA solution discriminates perfectly and completely nullifies within-group variance. The lowest error rate of *LOO6* is 0.09 higher for PC9-DA than for RDA. The difference is 0.07 for *LOO5*. These values are consistent with the results for low error levels of the simulation study in chapter 4 (see table 4.6). Prediction with PC9-DA is not substantially improved by using more than two discriminant dimensions. With RDA all dimensions are exploited to separate specific groups. It is remarkable that for each extra dimension the prediction is improved, especially if we consider the very small proportion of *VAF* (0.01) for dimension 4 and 5. For instance the improvement of 0.07 for *LOO6* in RDA prediction from three to four dimensions is mainly caused by the separation of radicle and coleoptil (see figure 7.8, dimension 4). This separation is not achieved by PC-DA (see figure 7.6).

In summary the analysis of the barley tissue profiles with PC-DA and RDA seems to confirm the theoretical expectations.

- Prediction with RDA compared to PC-DA is consistent with the results for low error levels of the simulation study in chapter 4 (see table 4.6).
- The RDA results simplify compared to PC-DA due to filtering out of within information.

An interesting property of RDA revealed by the analysis of the barley tissue profiles is that RDA is able to improve prediction with relatively small proportions of *VAF* up to the last dimension inclusive.

7.3 NRDA on barley tissue profiles

In this section we investigate the properties of Nonlinear Reflected Discriminant Analysis (NRDA) applied on the barley data described in section 7.2.1 and analysed in the previous section with PC-DA and RDA. We selected isotone transformations for the variables of the barley tissue profiles. With isotone transformations of variables not only the order of object values is preserved, but also the increase or decrease has to remain consistent. The Between-to-Total ratio *BT* (4.27) of the isotone transformed barley tissue profiles is 0.42, which is much higher than the non transformed value of 0.26.

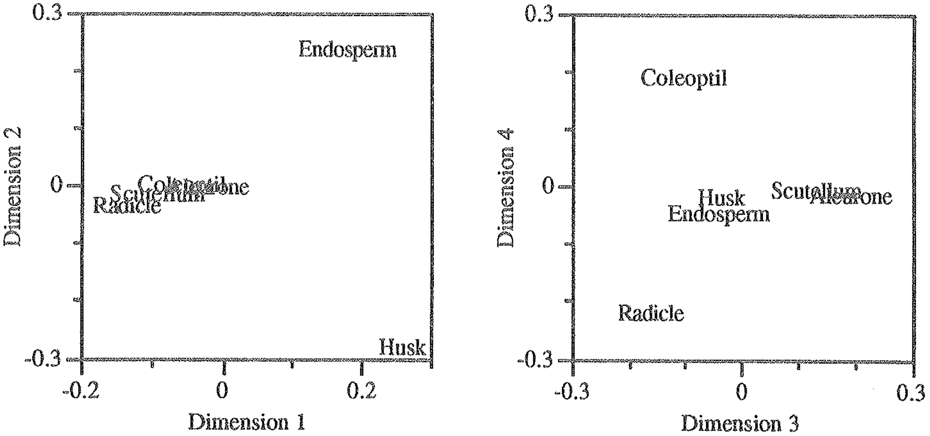


Figure 7.10 *NRDA discriminant space mapping 60 barley tissue profiles.*

Consequently the VAF_{prop} in table 7.10 reaches for NRDA a higher maximum than for RDA. The efficient span does not change notably from Im=17.4 to Im=16.9. The discriminant space of the NRDA solution is graphically displayed in figure 7.10. The solution is similar to the RDA solution in figure 7.8. A salient difference is the domination of the first two NRDA dimensions by husk and endosperm, whereas these two tissues disappear in the last three dimensions. In the RDA solution husk and endosperm contribute substantially to the third dimension. The loadings of the isotone transformed variables on to the first two NRDA dimensions are given in figure 7.11. Only loadings outside a circle with radius 0.5 are displayed. The cluster with loadings -0.673 0.728 contains mass numbers 72, 97, 101, 103, 110, 111 and 126. The complementary cluster with loadings 0.673 -0.728 contains 130, 163, 170, 172, 192, 194 and 222. The squared multiple correlation of all the isotone transformed variables of these two clusters with the first two NRDA dimensions is 0.98.

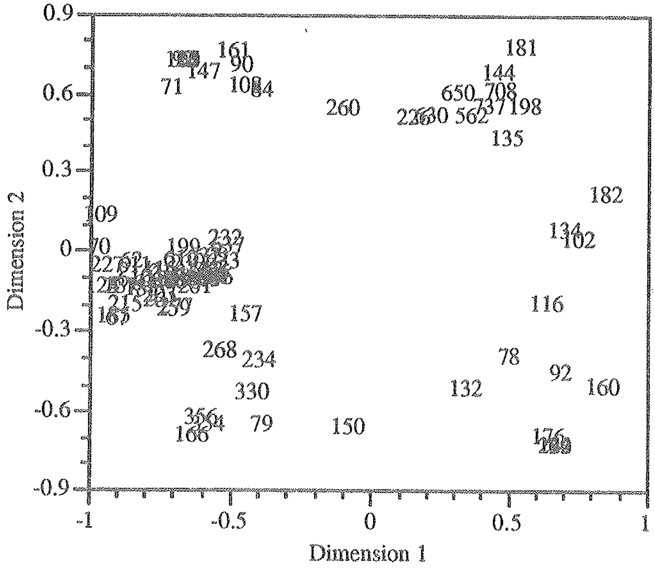


Figure 7.11 *NRDA discriminant space mapping loadings of transformed vars.*

For mass number 110 and 181 we will show how the variables are transformed. In figure 7.12 we display the isotone transformations of the objects for mass number 110. The objects are labeled by the first letter of their group tissue name. The 'non transformed values' give the not isotone transformed objects of the unit normalized variable in deviation from the mean as defined at the end of section 7.2.1 and the 'transformed values' give the corresponding optimal isotone transformations computed with NRDA.

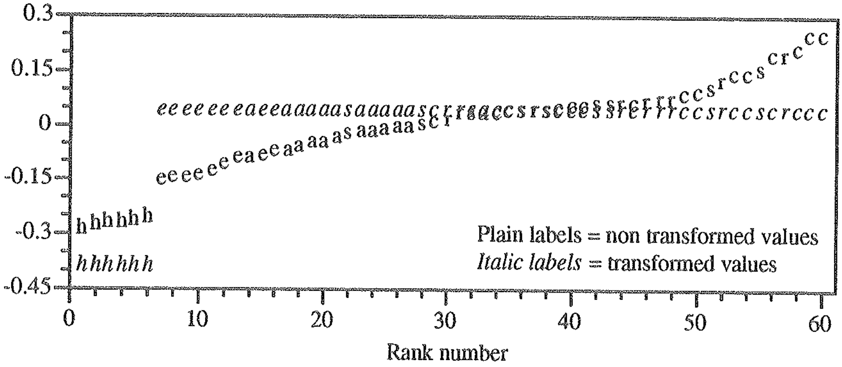


Figure 7.12 *NRDA transformations for mass number 110.*

The small gap between husk and all other tissues is widened by the NRDA transformation. The low intensities for husk become even lower by the isotonic transformation. The non transformed values of the other variables (72, 97, 101, 103, 111 and 126) in the cluster with loadings -0.673 0.728 are somewhat different, but the isotone transformed values are exactly the same as for number 110. The complementary cluster with mass numbers 130, 163, 170, 172, 192, 194 and 222 has exactly the same isotone transformation with the sign reversed and the 6 husk measurements are perfectly discriminated towards the positive side with prominent intensities. In figure 7.13 we illustrate the isotonic transformation for mass number 181. Endosperm is clearly separated from the other tissues by the isotone transformations. Only one transformed value of endosperm is intermediate, because this endosperm value was originally lower than a coleoptil value. The examples above show how the optimal transformations exaggerate differences in order between groups. This makes it easy to find some clear boundaries between groups of tissue with respect to profiles of mass intensities.

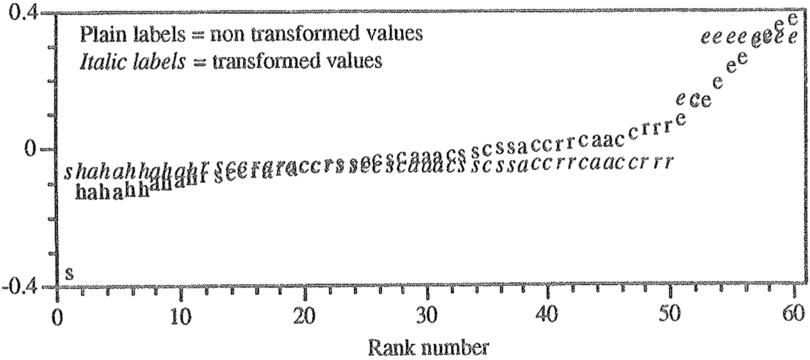


Figure 7.13 *NRDA transformations for mass number 181.*

We investigated NRDA prediction by computing L-O-O error rates (4.31) for the NRDA solutions. The NRDA L-O-O error rate is calculated by omitting one object from the raw data prior to NRDA. The transformed value of the omitted objects are computed by *neighbour quantification*. For neighbour *c* quantification *c* objects with the closest value to the value.of the omitted object in the raw data are selected for each variable separately. The mean of the corresponding isotone transformed values of these closest values is assigned to the omitted object as the transformed value of the omitted object for this variable. The substituted transformed object is projected into the NRDA discriminant space and classified to the closest group mean. This is repeated for all objects in the raw data, and the L-O-O error rate is given by the fraction of objects that are misclassified. We emphasize that by this procedure the isotone transformation of the remaining objects is independent of the the omitted object. We apply neighbour 9 quantification on the barley tissue profiles with *c* equal to the mean number of group objects minus one. The resulting NRDA L-O-O error rates are listed in table 7.10. Prediction is better with NRDA than with RDA, 0.05 for *LOO6* and 0.03 for *LOO5*. The minimum NRDA error rate for *LOO6* and *LOO5* is reached with three dimensions. Here the most striking improvement of 0.14 compared to RDA is scored for *LOO6*.

More research is needed on the optimal neighbour *c* quantification. For neighbour 1 quantification we obtain a smallest value of 0.15 for *LOO6* and of 0.07 for *LOO5*.

Prediction is only slightly worse than RDA prediction, which is an indication for the robustness of the NRDA procedure. Application of neighbour $c=g-1$ quantification on the barley tissue profiles with c equal to the number of group objects minus one, might improve NRDA prediction even further.

In summary the analysis of the barley tissue profiles with RDA and NRDA seems to give promising results.

- Maximum prediction is better with NRDA than with RDA,
- NRDA has a more efficient predictive capacity with a smaller number of discriminating dimensions.
- The optimal transformations exaggerate differences in order between groups, which make it easy to find some clear boundaries between groups of tissue with respect to profiles of mass intensities.

Chapter 8

CONCLUSIONS

In this monograph the integration of multiset MVA methods has been achieved in several ways. In chapter 2 multiset MVA methods are described in a comprehensive filter system of methods by filtering the eigenvalues of the sets. Hybrid MVA methods are placed in this system by combining different types of filter or by defining compound filters. In chapter 3 and 4 adjusted methods are formulated with corresponding filters. The integration approach of directed correlations in chapter 5 defines a wider scope of methods than the filter system. The equivalence of algorithms produced by Wold's basic PLS method of Soft modelling (Wold, 1982) and algorithms produced by the maximization of specific LDC path models illustrates the extended range of methods. Many related PLS algorithms can be derived from a corresponding fit function by specifying an appropriate LDC path model. The elaboration of the filter system in chapter 2 and directed correlations in 5 is illustrated with a selection of most characteristic methods. An exhaustive treatment of all possible methods is not pursued. For instance the relation of LDC with some three mode PLS algorithms still has to be studied. In the next sections we evaluate some results in more detail and outline future prospects.

8.1 Efficient rank

Embedded in the filter theory of chapter 2 we elaborated on ideas about the efficiency of information transfer by defining the information span with a corresponding measure for efficient rank. The efficient rank seems to give a reasonable estimate of the number of stable dimensions. In chapter 7 the efficient rank of the Miller-Nicely data was in agreement with the number of interpretable dimensions mentioned in previous publications about these data. Furthermore, the efficient rank of barley tissue profiles was consistent with the measurement design. A comparative study with other real-life data and other rank measures will be necessary for extensively assessing the properties of efficient rank.

8.2 Adjusted methods

In chapter 3 on Set Correlation with Set Variance Constraints we formulated the adjusted method of Set Component Analysis (SCA) and in chapter 4 on Set Variance with Set Correlation Constraints the Reflected Variance methods. In SCA the emphasis was primarily on maximizing the sum of squared correlations between set variates and secondly on improving variance accounted for. In Reflected Variance methods the emphasis was primarily on maximizing variance accounted for and secondly on improving squared canonical correlations. Theoretically and practically SCA was compared with INDSCAL. Reflected Discriminant Analysis (RDA) was compared with two other forms of discriminant analysis. The results indicate that the secondary improving constraint dominates the properties of the adjusted methods. More specifically SCA provides even a more adequate estimate of the true common dimensions of the INDSCAL model than the INDSCAL procedure of Carroll & Chang (1970). Other theoretical properties of SCA and INDSCAL are confirmed in chapter 7. For instance the SCA dimension weights of the Miller-Nicely data approximate more to the concept of differential weighting of dimensions than the INDSCAL weights. It is very convenient that the SCA solution is also simpler to interpret, but only further investigations of other data can give conclusive results. Chapter 4 shows how the rank reducing step in PC-DA can capitalize on the wrong information and how DA can capitalize on spurious regions. RDA does not have these drawbacks. A simulation study and a real-life analysis of barley tissue profiles confirm the better predictive capacities of RDA. Nonlinear extension of RDA provides new possibilities in group analysis. Datasets which could not be analysed with Nonlinear Discriminant Analysis (Gifi, 1990) can now be analysed with NRDA. The benefits of nonlinear transformations are illustrated with an example in chapter 7. Isotone transformations of barley tissue profiles computed with NRDA tend to emphasize boundaries between groups by exaggerating differences in order. By compensating the increase of freedom with neighbour quantification, group prediction is even improved compared to RDA.

8.3 Set Variance with Set Variance Constraints

Instead of integrating Set Correlation with Set Variance Constraints (chapter 3) or Set Variance with Set Correlation Constraints (chapter 4) we can also integrate Set Variance with Set Variance Constraints or Set Correlation with Set Correlation Constraints. We did not elaborate on these combinations in the respective chapters, but at this point like to confine ourselves to giving one example of such a method. An attractive fit function for a two sets Set Variance with Set Variance Constraints adjusted method would be to maximize tr $\mathbb{Z}_1'\mathbb{S}_1^{1/2}\mathbb{S}_2^{1/2}\mathbb{Z}_2$, referred to as Double Variance Analysis (DVA). DVA is attractive because it summarizes the fit functions of two complementary adjusted methods, tr $\mathbb{Z}_1'\mathbb{S}_1^{1/2}\mathbb{S}_2\mathbb{S}_1^{1/2}\mathbb{Z}_1$ and tr $\mathbb{Z}_2'\mathbb{S}_2^{1/2}\mathbb{S}_1\mathbb{S}_2^{1/2}\mathbb{Z}_2$ in one function with the same optimal solutions. The p dimensional DVA solution for $\mathbb{Z}_1$ and $\mathbb{Z}_2$ is equal to the first p singular vectors of $\mathbb{S}_1^{1/2}\mathbb{S}_2^{1/2}$ with the corresponding singular values in descending order. DVA therefore shows dual features comparable with Principal Component Analysis. The PCA solution has matching principal components and loadings for one set of variables, whereas the DVA solution has matching principal components for two sets.

8.4 Future prospects

The promising results for the nonlinear extension (Gifi, 1990) of RDA indicate that it would be interesting to investigate nonlinear extensions for SCA and LDC (including PLS) as well. Common scale transformations as developed by Van der Lans (1992) can add useful features to the nonlinear extensions by restricting the degrees of freedom. The fitting of reflected variance methods with a corresponding LDC path model has lead to a PLS2 variant with theoretically better predictive capacities than the usual PLS2 method for essential multivariate problems. Practical testing of theory is needed. A useful generalization of the INDRES model is expected by substituting the approximation of $\mathbb{S}_k$ with $\mathbf{X}\mathbf{W}_k\mathbf{X}'$ by the approximation of $\mathbb{H}_k$ with $\mathbf{X}\mathbf{W}_k\mathbf{Y}'$, where the number of variables for each set $\mathbb{H}_k$ must be the same. The multiset decomposition with this model might provide a more adequate decomposition of K sets than the CANDECOMP procedure of Carroll & Chang (1970), because fitting the INDRES model with SCA in chapter 3 improved the estimation of true common dimensions of the INDSCAL model compared to the INDSCAL procedure of Carroll & Chang (1970). Other methods for fitting the INDRES model might even further improve results. Finally we mention that Reflected Component Analysis can be adapted for performing a cluster analysis. For this purpose we have to assume that the mirror matrix U defines some unknown group space. The ideal group classification for some fixed number of groups is given by the global RCA maximum for unknown group space and unknown reflected discriminant space.

References

Anderson, T.W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics, 22*, 327-351.

Arabie, P., & Soli, S.D. (1982). The interface between the types of regression and methods of collecting proximity data. In R.G. Golledge & J.N. Rayner (Ed.), *Proximity and preference. Problems in the multidimensional analysis of large data sets* (pp. 90-115). Minneapolis:University of Minnesota Press.

Arabie, P., Carroll, J.D., & DeSarbo, W.S. (1987). Three-way scaling and clustering. *Sage University Paper series on Quantitative Applications in the Social Sciences* (Series no. 07-065). Beverly Hills: Sage Pubns.

Bennett, J.H. (1974). *Collected papers of R.A. Fisher.* University of Adelaide, South Australia: Coudrey Offset Press.

Bijleveld, C.J.H. (1989). *Exploratory linear dynamic systems analysis.* Leiden: DSWO Press, Leiden University.

Bloxom, B. (1968). *Individual differences in multidimensional scaling.* (ETS RM 68-45.) Princeton, New Jersey: Educational Testing Service.

Bookstein, F.L. (1982). The geometric meaning of soft modelling, with some generalizations. In K.G. Jöreskog & H. Wold (Ed.), *Systems under indirect observation. Part II* (pp. 55-74). Amsterdam: North Holland.

Campbell, N.A. (1980). Shrunken estimators in discriminant and canonical variate analysis. *Applied Statistics, 29*, 5-14.

Carroll, J.D. (1968). A generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th annual convention of the American Psychological Association, 3*, 227-228.

Carroll, J.D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*, 283-319.

de Jong, S., & Kiers, H.A.L. (1992). Principal covariates regression. Part I. Theory. *Chemometrics and Intelligent Laboratory Systems, 14*, 155-164.

de Leeuw, J., & Bijleveld, C.J.H. (1987). *Fitting reduced rank regression models by alternating least squares.* (Report RR-87-05.) Leiden: University of Leiden, Department of Data Theory.

de Leeuw, J., & Heiser, W.J. (1980). Multidimensional scaling with restrictions on the configuration. In P.R. Krishnaiah (Ed.), *Multivariate analysis V* (pp. 501-522). Amsterdam: North-Holland.

de Leeuw, J., & Pruzansky, S. (1978). A new computational method to fit the weighted euclidian distance model. *Psychometrika, 43*, 479-490.

Fisher, R.A. (1936). The use of multiple measurements on taxonomic problems. *Annals of Eugenics, 7*, 179-188.

Fornell, C., & Bookstein F.L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research, 19*, 440-452.

Fortier, J.J. (1966). Simultaneous linear prediction. *Psychometrika, 31*, 369-381.

Friedman, J.H., & Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, C-23*, 881-890.

Geladi, P., & Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta, 185*, 1-17.

Geladi, P., Martens, H., Martens, M., Kalvenes, S., & Esbensen, K. (1988). Multivariate comparison of laboratory measurements. In P. Tørboll (Ed.), *Symposium i anvendt statistik, København, 25-27 januar* (pp. 15-30). Danmarks edb-center for forskning og uddannelse.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Cichester: Wiley.

Gill, P.E., Murray, W., & Wright, M.H. (1981). *Practical optimization.* London: Academic Press, Inc.

Gittins, R. (1985). *Canonical analysis, a review with applications in ecology.* Berlin: Springer-Verlag.

Golub, G.H., & van Loan, C.F. (1990). *Matrix computations.* Baltimore: John Hopkins University Press.

Gower, J.C. (1992). The geometry of matrices. In S. Schach & G. Trenkler (Ed.), *Data analysis and statistical inference: Festschrift in honour of Prof. Dr. Friedhelm Eicker* (pp. 547-566). Köln: Verlag Josef Eul.

Green, B.F. (1969). Best linear composites with a specified structure. *Psychometrika, 34*, 301-318.

Hall, P. (1927). The distribution of means for samples of size N drawn from a population in which the variate takes between 0 and 1, all such values being equally probable. *Biometrika, 19*, 240-245.

Hauser, R.M., & Goldberger, A.S. (1971). The treatment of unobservable variables in path analysis. In H.L. Costner (Ed.), *Sociological Methodology* (pp. 81-117). San Francisco: Jossey-Bass.

Hoerl, A.E., & Kennard, W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics, 12*, 55-67.

Hoogerbrugge, R., Willig, S.J., & Kistemaker, P.G. (1983). Discriminant analysis by double stage principal component analsysis. *Analytical Chemistry, 55*, 1711-1712.

Horan, C.B. (1969). Multidimensional scaling: combining observations when individuals have different perceptual structures. *Psychometrika, 34*, 139-165.

Horst, P. (1961). Relations amoung m sets of measures. *Psychometrika, 26*, 129-149.

Huber, P.J. (1985). Projection pursuit. *The Annals of Statistics, 13*, 435-475.

Kiers, H.A.L. (1989). *Three-way methods for the analysis of qualitative and quantitative two-way data.* Leiden: DSWO Press, Leiden University.

Kroonenberg, P.M. (1983). *Three-mode principal component analysis: theory and applications.* Leiden: DSWO Press, Leiden University.

Kruskal, J.B., & Carroll, J.D. (1969). Geometrical models and badness-of-fit functions. In P.R. Krishnaiah (Ed.), *Multivariate Analysis, Vol. II* (pp. 639-671). New York: Academic Press.

L'Hermier des Plantes, H. (1976). *Structuration des tableaux à trois indices de la statistiques.* Thèse de 3ème cycle, Université Montpelier II.

Lohmöller, J.-B. (1989). *Latent variable path modelling with partial least squares.* Heidelberg: Physica-Verlag.

Lorber, A., Wangen, L.E., & Kowalski, B.R. (1987). A theoretical foundation for the PLS algorithm. *Journal of Chemometrics, 1*, 19-31.

MacCallum, R.C., & Cornelius, E.T. (1977). A Monte Carlo inverstigation of recovery of structure by ALSCAL. *Psychometrika, 42*, 401-428.

Manne, R. (1987). Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems, 2*, 187-197.

Martens, M., & Martens, H. (1986). Partial least squares regression. In J. Piggott (Ed.), *Statistical procedures in food research* (pp. 293-359). London: Elsevier Applied Science.

Maxwell, A.E. (1977). *Multivariate Analysis in Behavioral Research.* London: Chapman and Hall.

Meulman, J.J. (1986). *A distance approach to nonlinear multivariate analysis.* Leiden: DSWO Press, Leiden University.

Miller, G.A., & Nicely, P.E (1955). An analysis of perceptual confusions amoung some English consonants. *J. Acoust. Soc. Am., 27*, 338-352.

Nierop, A.F.M. (1989). *Generalized set component analysis: A toolbox for multiple sets analysis.* (Report RR-89-04.) Leiden: University of Leiden, Department of Data Theory.

Nierop, A.F.M. (1993). The INDRES model: an INDSCAL model with residuals orthogonal to INDSCAL dimensions. In R. Steyer, K.F. Wender & K.F. Widaman (Ed.), *Proceedings of the 7th European Meeting of the Psychometric Society* (pp. 366-370). Stuttgart and New York: Gustav Fisher Verlag.

Nierop, A.F.M. (1991). Reflected Variables: their use in discriminant analysis. *Third Conference of the International Federation of Classification Societies, Edinburgh, Scotland, August 6-9, 1991.*

Noonan, R., & Wold, H. (1982). PLS path modelling with indirectly observed variables: a comparison of alternative estimates for the latent variable. In K.G. Jöreskog & H. Wold (Ed.), *Systems under indirect observation. Part II* (pp. 75-94). Amsterdam: North Holland.

Ramsay, J.O. (1988). Monotone regression splines in action. *Statistical Science, 3*, 425-461.

Shepard, R.N. (1972). Psychological representation of speech sounds. In E.E. David, Jr. & P.B. Denes (Ed.), *Human communications: a unified view* (pp. 67-113). New York: MacGraw Hill.

Shepard, R.N. (1987). George Miller's data and the birth of methods for representing cognitive structures. In W. Hirst (Ed.), *Giving birth to cognitive science: A festschrift for George A. Miller* (). New York: Cambridge University Press.

Soli, S.D., & Arabie, P. (1979). Auditory versus phonetic accounts of observed confusions between consonant phonemes. *J. Acoust. Soc. Am., 66*, 46-59.

Stone, M., & Brooks, R.J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society B, 52*, 237-269.

Tas, A.C., Angelino, S.A.G.F, La Vos, G.F., & van der Greef, J. (1991). Data analysis of pyrolysis-mass spectrometric profiles generated with soft ionization detection: application to barley tissue profiles. *Journal of Analytical and Applied Pyrolysis, 20*, 73-85.

ten Berge, J.M.F. (1983). A generalization of Kristof's theorem on the trace of certain matrix products. *Psychometrika, 48*, 519-523.

ten Berge, J.M.F. (1986). A general solution for the Maxbet problem. In J. de Leeuw, W. J. Heiser, J. Meulman & F. Critchley (Ed.), *Multidimensional data analysis* (pp. 81-87). Leiden: DSWO Press, Leiden University.

ten Berge, J.M.F. (1988). Generalized approaches to the Maxbet problem and the Maxdiff problem, with applications to canonical correlations. *Psychometrika, 53*, 487-494.

Torgerson, W.S. (1958). *Theory and methods of scaling.* New York: Wiley.

Tucker, L.R. (1951). *A method for synthesis of factor analytic studies.* (Personnel Research Section Report No. 984.) Washington, D.C.: Department of the Army.

Tucker, L.R. (1958). An inter-battery method of factor analysis. *Psychometrika, 23*, 111-136.

van de Geer, J.P. (1984). Linear relations among K sets of variables. *Psychometrika, 49*, 79-94.

van de Geer, J.P. (1986). *Introduction to linear multivariate data analysis.* Leiden: DSWO Press, Leiden University.

van den Wollenberg, A.L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika, 42*, 207-219.

van der Burg, E. (1988). *Nonlinear canonical correlation and some related techniques.* Leiden: DSWO Press, Leiden University.

van der Lans, I.A. (1992). *Nonlinear multivariate analysis for multiattribute preference data.* Leiden: DSWO Press, Leiden University.

van der Leeden, R. (1990). *Reduced rank regression with structured residuals.* Leiden: DSWO Press, Leiden University.

van Rijckevorsel, J.L.A. (1987). *The application of fuzzy coding and horseshoes in multiple correspondence analysis.* Leiden: DSWO Press, Leiden University.

Wilkinson, J.H. (1965). *The algebraic eigenvalue problem.* Oxford: University Press.

Wold, H. (1982). Soft modelling: the basic design and some extensions. In K.G. Jöreskog & H. Wold (Ed.), *Systems under indirect observation. Part II* (pp. 1-54). Amsterdam: North Holland.

Yendle, P.W., & Macfie, H.J.H. (1989). Discriminant principal components analysis. *Journal of Chemometrics, 3*, 589-600.

Young, F.W. (1970). Nonmetric multidimensional scaling: recovery of metric information. *Psychometrika, 35*, 455-473.

Summary

In chapter 1 on *Integration of divergent aims in Multidimensional Analysis* the predictive value of multiset multivariate methods is related to the optimal integration of two criteria: stability and exactness. Stability of prediction is linked to *set variance* and exactness of prediction is linked to *set correlation*. Based on strategies to combine stable with exact prediction, we introduce a classification of hybrid and adjusted multivariate methods. Some considerations on mathematical tools and presentation are added. An outline of the structure of this monograph is provided.

In chapter 2 on *A Filter View on Multiset Models* we illustrate that many Multivariate Analysis (MVA) methods are build up with set variance and set correlation constituents. Our first aim is to show a variety of construction methods and not an exhaustive inventory of methods. Two new methods are proposed, based on *potential variance accounted for* and *information span*. The last three main sections show how set variance and set correlation can be integrated with competitive subfunctions and therefore illustrate the concept of hybrid methods.

In chapter 3 on *Set Correlation with Set Variance Constraints* we describe *Set Component Analysis* from several points of view. (1) The method integrates a set correlation and a set variance part by maximizing the sum of squared set correlations and adjusting the set variates with set variance constraints. (2) SCA is identical to Multiset CCA with proportionality restrictions on the variable weights. (3) By defining a free quadratic filter, SCA is related with the filter theory formulated in chapter 2. We conclude this chapter by indicating relations with other methods and presenting a simulation study of INDSCAL compared with SCA. The relation between INDSCAL and SCA is established by proposing and fitting a new model, the *INDRES model*.

In chapter 4 on *Set Variance with Set Correlation Constraints or Reflected Variance* we introduce *Reflected Variance methods* also from several points of view. (1) The Reflected Variance methods integrate a set variance and a set correlation part by maximizing the variance accounted for by set variates and adjusting the set variates with set correlation constraints. (2) The Reflected Variance methods project variables from one set on to another set, project these variables back and then compute principal components of the *reflected variables*. (3) By defining reflecting filters, Reflected Variance methods are related with the filter theory formulated in chapter 2. The

principle of reflected variables is elaborated by defining *Reflected Component Analysis* (RCA) and *Reflected Discriminant Analysis* (RDA). It will be shown theoretically how and under which conditions RDA can improve group prediction compared to Discriminant Analysis (DA) and Principal Component - Discriminant Analysis (PC-DA). In a simulation study theoretical results are confirmed. Some multiset and nonlinear extensions are proposed.

In chapter 5 on *Directed Correlations and Partial Least Squares* a new multiplicative hybrid method is formulated that maximizes the product of two complementary fit functions, a local and a global MVA function. The local function gives a multiset alternative for maximizing variance accounted for. The global function maximizes correlations as formulated in chapter 3. These adjusted correlations are called *directed correlations* and are embedded in a multiset path analysis framework utilizing *primary* and *secondary* predictions. The product function that globally maximizes directed correlations and locally increases set variance as much as possible is called Lifted Directed Correlations (LDC). LDC is able to describe many existing MVA methods, hybrid and adjusted methods. It gives one fit function for cyclic hybrid methods like the basic and extended Partial Least Squares (PLS) method of path modelling, Consensus PLS and PLS Hierarchical Components.

In chapter 6 on *Algorithms* we present two algorithms for non eigenvalue-eigenvector problems. First a simultaneous and successive monotone convergent algorithm for Set Component Analysis (chapter 3) is developed, where an interesting general algorithmic subproblem is to maximize the variance of different matrices accounted for by corresponding orthogonal latent variables. Secondly we elaborate a monotone convergent algorithm for Nonlinear Reflected Discriminant Analysis (chapter 4).

In chapter 7 on *Examples* we present analyses of real-life data using three methods developed in the preceding chapters. For a psychometric application of Set Component Analysis (chapter 3) we compare the SCA solution of the Miller-Nicely data with the corresponding INDSCAL solution. Reflected Discriminant Analysis from chapter 4 is applied on mass spectrometric barley tissue profiles and compared with results for PC-DA. The barley tissue profiles are also analysed with Nonlinear Reflected Discriminant Analysis.

Finally we draw our conclusions in chapter 8.

Author Index

Curriculum Vitae

Dré Nierop werd geboren op 29 maart 1954 te Leiden. Van 1966 tot 1972 bezocht hij het Bonaventura College in Leiden en voltooide daar zijn Gymnasium ß opleiding. Vervolgens ging hij biologie studeren aan de Rijksuniversiteit te Leiden. In 1976 behaalde hij het Kandidaatsexamen en in 1981 het Doctoraal examen met hoofdvak morfologie en bijvakken datatheorie (I.s.m. het Rijksmuseum van Natuurlijke Historie te Leiden), milieukunde en onderwijskunde. Van 1980 tot 1981 was hij studentassistent voor halve dagen bij de Vakgroep Datatheorie, Fakulteit Sociale Wetenschappen, Rijksuniversiteit Leiden. In 1983 was hij aangesteld als assistent onderzoeker bij het Max Planck Instituut te Nijmegen. In het kader van een taalpsychologisch onderzoek naar de relatie tussen denken en taalproduktie analyseerde hij het gebruik van stopwoorden. Van 1984 tot 1986 was hij aangesteld als onderzoeksmedewerker en toegevoegd onderzoeker op een ZWO (nu NWO) project bij de Vakgroep Ontwikkelingspsychologie, Fakulteit Sociale Wetenschappen, Rijksuniversiteit Leiden. In het kader van een gedragsonderzoek analyseerde hij het ontwikkelingsverloop in de interactie patronen bij verzorgers met prikkelbare en niet-prikkelbare babies. Van 1985 tot 1987 was hij als onderzoeker betrokken bij een project voor de ontwikkeling van Taalscreeningsinstrumenten voor Drie- tot Zesjarigen bij de Vakgroep Ontwikkelingspsychologie, Rijksuniversiteit Leiden, gesubsidieerd door het Praeventiefonds te 's-Gravenhage. Tijdens deze periode was hij op een gelijksoortig project werkzaam voor het Nederlands Instituut voor het Dove en Slechthorende Kind te Amsterdam. Van 1986 tot 1987 begeleidde hij als onderzoeker voor het Nederlands Astmafonds te Leusden een project over het ziekteverloop bij CARA patienten in samenwerking met het St. Antonius Ziekenhuis te Nieuwegein, de DSWO en de Vakgroep Datatheorie, Rijksuniversiteit Leiden. Van 1987 tot 1992 was hij als assistent in opleiding aangesteld bij de Vakgroep Datatheorie, Fakulteit Sociale Wetenschappen, Rijksuniversiteit Leiden. Vanaf 1992 is hij door het Centrum voor Bio-Farmaceutische Wetenschappen, Rijksuniversiteit Leiden aangesteld als onderzoeker op een innovatieproject patroonherkenning bij TNO - Voeding, Afdeling Structuuropheldering en Instrumentele Analyse te Zeist.

Errata

- The upper part of page 35 must be

$$\mathbf{v}_c = (\mathbf{I} + \nu_c\Phi_c^{-2}\underline{\underline{)}}^{-1}\mathbf{P}_c'\mathbf{x}. \qquad \forall c \quad (2.34)$$

After insertion of $\mathbf{Q}_c\Phi_c^{-1}\mathbf{v}_c$ for the MRR weights $\mathbf{t}_c$ in (2.32) with $\mathbf{v}_c$ according to (2.34) we maximize

$$\mathrm{MRR}_{p=1}\text{: } \underline{\underline{\mathrm{Fit}}}(\mathbf{x}) = \sum_{c=1}^{K} \mathbf{x}'\mathbf{P}_c(\mathbf{I} + \nu_c\Phi_c^{-2})^{-1}\mathbf{P}_c'\mathbf{x}, \qquad (2.35)$$

with $\mathbf{x}'\mathbf{x}=1$ and $\nu_c \geq 0\ \forall c$.

After maximization of (2.35) the optimal MRR variates are

$$\mathbf{H}_c\mathbf{t}_c = \mathbf{P}_c\mathbf{v}_c = \mathbf{P}_c(\mathbf{I} + \nu_c\Phi_c^{-2}\underline{\underline{)}}^{-1}\mathbf{P}_c'\mathbf{x}. \qquad \forall c \quad (2.36)$$

- Some traces must be added:

On page 27 in formula 2.25

$\mathbf{V}_1'\mathbf{P}_1'\mathbf{P}_2\mathbf{P}_2'\mathbf{P}_1\mathbf{V}_1$ *and* $\mathbf{V}_2'\mathbf{P}_2'\mathbf{P}_1\mathbf{P}_1'\mathbf{P}_2\mathbf{V}_2$ *become*
$\underline{\underline{\mathrm{tr}}}\ \mathbf{V}_1'\mathbf{P}_1'\mathbf{P}_2\mathbf{P}_2'\mathbf{P}_1\mathbf{V}_1$ *and* $\underline{\underline{\mathrm{tr}}}\ \mathbf{V}_2'\mathbf{P}_2'\mathbf{P}_1\mathbf{P}_1'\mathbf{P}_2\mathbf{V}_2$.

On page 55 in line 9

$\mathbf{X}'\mathbf{E}_k\mathbf{E}_k\mathbf{X}$ and $\mathbf{X}'\mathbf{M}_k\mathbf{M}_k\mathbf{X}$ *must be* $\underline{\underline{\mathrm{tr}}}\ \mathbf{X}'\mathbf{E}_k\mathbf{E}_k\mathbf{X}$ and $\underline{\underline{\mathrm{tr}}}\ \mathbf{X}'\mathbf{M}_k\mathbf{M}_k\mathbf{X}$.

On page 65 in formula ^{1}MCCA

$\mathbf{V}_1'\mathbf{P}_1'\mathbf{P}_2\mathbf{P}_2'\mathbf{P}_1\mathbf{V}_1$ *becomes* $\underline{\underline{\mathrm{tr}}}\ \mathbf{V}_1'\mathbf{P}_1'\mathbf{P}_2\mathbf{P}_2'\mathbf{P}_1\mathbf{V}_1$ *and three lines down*
$\mathbf{V}'\mathbf{P}'\mathbf{U}\mathbf{U}'\mathbf{P}\mathbf{V}=\mathbf{X}'\mathbf{P}\mathbf{P}'\mathbf{U}\mathbf{U}'\mathbf{P}\mathbf{P}'\mathbf{X}$ *must be* $\underline{\underline{\mathrm{tr}}}\ \mathbf{V}'\mathbf{P}'\mathbf{U}\mathbf{U}'\mathbf{P}\mathbf{V}=\underline{\underline{\mathrm{tr}}}\ \mathbf{X}'\mathbf{P}\mathbf{P}'\mathbf{U}\mathbf{U}'\mathbf{P}\mathbf{P}'\mathbf{X}$.

- On page 112 after formula (5.19) delete 'usually' and add 'Only'

> The right-hand and left-hand eigenvectors of A are ~~usually~~ not orthogonal (Wilkinson, 1965). Only if A is symmetric we have U'U=I, and $\mathbf{U}^{-1}$=U'.