Canonical Variate Analysis and Related Methods with Longitudinal Data

Michael Beaghen

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Eric P. Smith, Chair
Jesse C. Arnold
Robert V. Foutz
Donald R. Jensen
Keying Ye

November 13, 1997
Blacksburg, Virginia

Keywords: Redundancy Analysis, Procrustes Rotation, Common Principal Components

Canonical Variate Analysis with Longitudinal Data

Michael Beaghen

(ABSTRACT)

Canonical variate analysis (CVA) is a widely used method for analyzing group structure in multivariate data. It is mathematically equivalent to a one-way multivariate analysis of variance and often goes by the name of canonical discriminant analysis. Change over time is a central feature of many phenomena of interest to researchers. This dissertation extends CVA to longitudinal data. It develops models whose purpose is to determine what is changing and what is not changing in the group structure. Three approaches are taken: a maximum likelihood approach, a least squares approach, and a covariance structure analysis approach. All methods have in common that they hypothesize canonical variates which are stable over time.

The maximum likelihood approach models the positions of the group means in the subspace of the canonical variates. It also requires modeling the structure of the within-groups covariance matrix, which is assumed to be constant or proportional over time. In addition to hypothesizing stable variates over time, one can also hypothesize canonical variates that change over time. Hypothesis tests and confidence intervals are developed.

The least squares methods are exploratory. They are based on three-mode PCA methods such as the Tucker2 and parallel factor analysis. Graphical methods are developed to display the relationships between the variables over time.

Stable variates over time imply a particular structure for the between-groups covariance matrix. This structure is modeled using covariance structure analysis, which is available in the SAS package Proc Calis.

Methods related to CVA are also discussed. First, the least squares methods are extended to canonical correlation analysis, redundancy analysis, Procrustes rotation and correspondence analysis with longitudinal data. These least squares methods lend themselves equally well to data from multiple datasets. Lastly, a least squares method for the common principal components model is developed.

Dedicated to my parents.

ACKNOWLEDGEMENTS

Foremost I thank my advisor for his superb guidance. I also thank my committee members for their efforts and the Statistics faculty for their fine teaching and support.

TABLE OF CONTENTS

CHAPTER EIGHT

CANONICAL VARIATE ANALYSIS OVER TIME 81

CHAPTER NINE

SCALING THE VARIABLES 119

LIST OF FIGURES

LIST OF TABLES

Notation

There are several conventions that are adhered to in this dissertation. A column vector is bold and small lettered. Matrices are bold and large lettered. Scalars are small lettered and not bold. A constant is italicized, an index is not.

CHAPTER ONE

INTRODUCTION

Change over time is a central feature of many phenomena of interest to researchers. Analysis of data measured over time is well developed for univariate analysis (Hand & Crowder 1989, Diggle, Liang & Zeger 1994), but not as well developed for multivariate analysis. In this dissertation I extend certain multivariate methods to longitudinal data so that one can investigate the change over time in the underlying structures in the data.

The main interest of the dissertation is extending canonical variate analysis (CVA) to longitudinal data. However, other related multivariate methods which I extend to longitudinal data are canonical correlation analysis (CCA), redundancy analysis (RA) and Procrustes rotation (PR). These are kindred methods which relate two sets of variables, which I shall designate X-variables and the Y-variables. In the case of CVA the X-variables are group indicators, while for CCA, RA and PR the variables may be either continuous or categorical. Of these four methods, I give particular attention to canonical variate analysis because of its obvious usefulness and because it is the most mathematically tractable when extended to data taken over time.

In addition to change over time I investigate change that occurs over different datasets or groups. Although not a primary interest of this dissertation, the problem of modeling canonical correlation analysis, canonical variate analysis, redundancy analysis and Procrustes rotation with data from multiple datasets is closely related to the problem of modeling these analyses with longitudinal data. I also devote a chapter to common principal components analysis (Flury 1988), which models principal components with data from multiple datasets and shares conceptual similarities to the other models I discuss.

Several multivariate models for data over time have already been developed. Related to some of the approaches to be taken in this dissertation, Kiers (1991) gives an overview of using three-mode principal components to model principal components over time. Three-mode principal components decomposes three-mode data, that is, data which is in the form of a three-way array. Swaminathan (1984) has developed models for factor analysis with longitudinal data. Modeling the relationship between two sets of variables over time, however, has not been well developed. Regression can be viewed as a model relating two sets of variables, where one set consists of the response. Regression over time has been modeled by different methods. Ware (1985) describes random effects models, autoregressive models and multivariate models. Liang and Zeger (1986) discuss using generalized linear models for longitudinal regression. Akin to factor analysis and multiple regression is LISREL (LInear Structural RELations, Jöreskog 1989), which relates variables in a structural model. Jöreskog (1979) describes how LISREL can be approached with longitudinal data, although the results are limited.

In addition to multivariate models for longitudinal data, a few multivariate methods for data from multiple datasets or groups have also been developed. A method for performing principal components analysis on data from multiple datasets is common principal components (see Chapter Four). Closely related to one method developed in this dissertation is a method for performing canonical variate analysis on data from multiple datasets (Campbell and Tomenson 1983).

What is shared by the models mentioned above is that they hypothesize common or stable variates across the multiple occasions or datasets. I shall take the same approach to modeling change. The common component or variate shall be the *leitmotiv* of this dissertation. The nature of change shall be investigated by asking the question: What remains stable over time, and what changes? In particular, I ask if it is useful to model variates as constant or stable over time. If the common variate approach is deemed useful, then the nature of the change is indicated by the strength of the relationship, i.e., the orrelation or covariance, between the pairs of variates at each occasion.

Having outlined the problem and the nature of the attempted solution, I will now lay out the organization of the dissertation. Chapter Two provides the reader with the background material necessary to frame the problem and approach the solutions attempted in the rest of the dissertation. Chapter Two first discusses canonical correlation analysis, followed by canonical variate analysis, which is a special case of canonical correlation analysis. Then it discusses redundancy analysis and Procrustes rotation, two methods closely related to canonical correlation analysis. Next three-mode methods are discussed. Lastly, Campbell and Tomenson's model for canonical variate analysis for multiple datasets is discussed.

While the material presented in Chapter Two is strictly a review, the chapters which follow present mostly new material. Background material appears in Chapters Four, Six, Seven and Nine. These chapters are self-contained; they have both new material and the related background.

In Chapter Three I show the partitioning of sums of squares and prove the nestedness of the solutions for the PARAFAC (orth.) model. These results are important for the developments of Chapters Four and Five.

Chapter Four compares the maximum likelihood and least squares approaches to common principal components (CPC). It starts with background on Flury's (1984) maximum likelihood approach, then outlines how CPC can be approached by three-mode PCA. This chapter is the one least organically connected to the rest of the dissertation. Besides interest in the CPC model for its own sake, it relates to the main thesis in two ways. First, it is a clear exposition of a common variate model. Second, it shows how maximum likelihood and least squares methods complement each other. In this sense it suggests possible developments of common variate models for canonical variate analysis, canonical correlation analysis and redundancy analysis over time which could be approached by both maximum likelihood and least squares.

Chapters Five and Six present least squares methods. Chapter Five models canonical variate analysis, canonical correlation analysis, redundancy analysis and Procrustes rotation over multiple occasions and over multiple datasets with three-mode principal components. In Chapter Six graphical techniques are developed to be used in conjunction with the least squares methods of Chapter Five. I do not develop hypothesis tests for the least squares methods. Hence the least squares methods are not as powerful as the maximum likelihood method. However, they are more flexible. They can be applied to data where the X-variables are continuous. They can be applied to data where Y-variables are categorical, such as correspondence analysis. Furthermore, they are well suited for exploratory analysis.

In Chapter Seven canonical variate analysis over time is approached by the analysis of covariance structures or COSAN (COvariance Structure ANalysis, McDonald 1978), which is related to LISREL modeling.

In Chapter Eight I develop a model for canonical variates over time which is estimated by maximum likelihood. Like Campbell and Tomenson's approach (see Section **2.5**), it models group means. This is the only maximum likelihood method that I develop fully in the dissertation. I develop hypothesis tests based on the likelihood ratio principal and confidence intervals based on the inverse of the information matrix.

In Chapter Nine I discuss the important issue of the scaling of the variables. Chapter Ten concludes the dissertation and suggests further research.

CHAPTER TWO

BACKGROUND

In this dissertation I develop methods for certain multivariate applications when one has three-mode data; that is, data taken over multiple occasions or datasets. In this chapter I present background material that is central to the development of these methods. Section **2.1** details the singular value decomposition (SVD). The SVD is important to this dissertation for several reasons. First, it unifies the methods for relating two sets of variables, as they can be put into the framework of a SVD. Second, the three-mode PCA models are generalizations of the SVD. Lastly, the SVD is the basis for the biplot, a fundamental graphical technique. Section **2.2** describes canonical correlation analysis, canonical variate analysis, redundancy analysis and Procrustes rotation. These are kindred methods for relating two sets of variables which I will generalize to longitudinal and multiple group data. Section **2.3** delves into three-mode principal components, which is one of the approaches I take to generalizing the aforementioned multivariate methods. An advantage of three-mode PCA is that it lends itself readily to graphical displays. Lastly, Section **2.4** discusses the Campbell and Tomenson model for canonical variate analysis for data from multiple datasets. To the best of my knowledge it is the only model that

accomplishes something similar to that which the models in this dissertation will accomplish. In particular, the model I develop in Chapter Eight can be viewed as an extension of Campbell and Tomenson's model.

2.1 THE SINGULAR VALUE DECOMPOSITION

An $m \times n$ matrix $\mathbf{X}$ can be decomposed as follows (Kshirsagar 1972):

$$\mathbf{X} = \mathbf{\Pi \Sigma \Omega}'$$

where $\mathbf{\Pi}$ is an $m \times m$ orthogonal matrix, $\mathbf{\Omega}$ is an $n \times n$ orthogonal matrix, and $\mathbf{\Sigma}$ is an $m \times n$ matrix with elements σ_{ij}, where the singular values are $\sigma_{jj} = \sigma_j$, and $\sigma_{ij} = 0$ for $i \neq j$. Such a decomposition is called the singular value decomposition (SVD). An equivalent form of the SVD (assume $m \geq n$) specifies that $\mathbf{\Pi}$ be an $m \times n$ matrix with orthonormal columns and $\mathbf{\Sigma}$ an $n \times n$ diagonal matrix with $\sigma_{jj} = \sigma_j$. Which form is being referred to will be clear from the context.

The SVD has the optimality property (Eckart and Young 1936) that it yields the best low rank approximation to a matrix in a least squares sense. If $\mathbf{X}$ has rank r and one wants a rank s approximation to $\mathbf{X}$, $s < r$, then the optimal approximation is $\hat{\mathbf{X}} = \mathbf{\Pi}_s \mathbf{\Sigma}_s \mathbf{\Omega}_s'$, where $\mathbf{\Sigma}_s$ is an $s \times s$ diagonal matrix whose diagonal elements are the s largest singular values of $\mathbf{X}$, and $\mathbf{\Pi}_s$ and $\mathbf{\Omega}_s$ contain the columns of $\mathbf{\Pi}$ and $\mathbf{\Omega}$ corresponding to those s singular values. It is this optimality property that allows the SVD to be used as the basis for biplots. For a discussion of some other uses of the SVD in statistics see Good (1969).

If $\mathbf{X}$ is a symmetric matrix then the singular value decomposition is equivalent to a spectral decomposition. $\mathbf{\Pi}_s = \mathbf{\Omega}_s =$ the matrix of eigenvectors of $\mathbf{X}$, while $\mathbf{\Sigma}_s$ is a diagonal matrix of eigenvalues of $\mathbf{X}$.

2.2 CANONICAL CORRELATION AND RELATED MODELS

Canonical correlation analysis (CCA) is the most widely known of the multivariate methods that relate two sets of variables. CCA relates p X-variables to q Y-variables. The assignment of variables into the X-set and the Y-set is always performed before the analysis and is based on the nature and purpose of the study. For example, a medical researcher may want to relate lifestyle variables, X-variables, such as daily caloric intake and exercise, with cardiovascular variables, Y-variables, such as cholesterol level and blood pressure. Canonical correlation analysis (CCA), redundancy analysis (RA) and Procrustes rotation (PR) all require this a priori division of the variables.

As a preliminary, define the covariance matrix $\mathbf{S}$:

$$\mathbf{S} = \left(\frac{1}{n-1}\right)\begin{bmatrix} \mathbf{Y}'\mathbf{Y} & \mathbf{Y}'\mathbf{X} \\ \mathbf{X}'\mathbf{Y} & \mathbf{X}'\mathbf{X} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{YY} & \mathbf{S}_{YX} \\ \mathbf{S}_{XY} & \mathbf{S}_{XX} \end{bmatrix},$$

where $\mathbf{X}$ is an $n \times p$ matrix of measurements of p variables on n units, and $\mathbf{Y}$ is an $n \times q$ matrix of measurements of q variables on n units. I shall assume throughout the discussion that $\mathbf{X}$ and $\mathbf{Y}$ are centered by variable. If the variables are also scaled to unit variance $\mathbf{S}$ is a correlation matrix.

2.2.1 Canonical Correlation

Hotelling (1935) proposed canonical correlation analysis as a model to relate two sets of variables measured on the same units. He derived linear combinations of the X-variables and linear combinations the Y-variables that were maximally correlated, subject to the constraints that each derived variate was uncorrelated with the other variates in its set and that each variate had a variance of one. Denote the vector of X-variables by **x**, and the vector of Y-variables by **y**. CCA finds linear compounds of **x** and **y**, a_i and b_i:

$$a_i = \mathbf{w}_i'\mathbf{x}, \qquad b_i = \mathbf{v}_i'\mathbf{y},$$

choosing $\mathbf{w}_i$ and $\mathbf{v}_i$ to maximize the correlation between a_i and b_i, subject to the constraints:

$$\mathbf{w}_i'\mathbf{S}_{XX}\mathbf{w}_j = 0, \qquad \mathbf{v}_i'\mathbf{S}_{YY}\mathbf{v}_j = 0 \qquad \forall\, i \neq j$$

and

$$\mathbf{w}_i'\mathbf{S}_{XX}\mathbf{w}_i = 1, \qquad \mathbf{v}_i'\mathbf{S}_{YY}\mathbf{v}_i = 1.$$

The canonical coefficients, $\mathbf{w}_i$ and $\mathbf{v}_i$, can be obtained by a spectral analysis. The $\mathbf{w}_i$ are eigenvectors of the matrix $\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}$ and the $\mathbf{v}_i$ are eigenvectors of $\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}$. The eigenvalues of these two matrices are equal, and they are the canonical correlations between the pairs of canonical variates, $\mathbf{w}_i$ and $\mathbf{v}_i$.

Let **W** represent the matrix whose columns consist of $\mathbf{w}_i$ and **V** the matrix whose columns consist of $\mathbf{v}_i$. For a positive definite matrix **M** define the unique symmetric positive definite square root matrix $\mathbf{M}^{1/2}$ such that $\mathbf{M} = \mathbf{M}^{1/2}\mathbf{M}^{1/2}$. Clearly, $\mathbf{M}^{1/2} = \mathbf{L}\Lambda^{1/2}\mathbf{L}'$, where $\mathbf{M} = \mathbf{L}\Lambda\mathbf{L}'$ is the spectral decomposition of **M**. Then it is also possible to get the canonical coefficients from a singular value decomposition of the matrix $\mathbf{S}_{XX}^{-1/2}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1/2}$ (Gittins 1985) as follows:

$$\mathbf{S}_{XX}^{-1/2}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1/2} = \mathbf{W}^*\mathbf{E}\mathbf{V}^{*\prime},$$

where $\mathbf{W}^* = \mathbf{S}_{XX}^{1/2}\mathbf{W}$, $\mathbf{V}^* = \mathbf{S}_{YY}^{1/2}\mathbf{V}$, **E** is a diagonal matrix with the canonical correlations as its elements.

Canonical correlation has the appealing property of biorthogonality. Biorthogonality is the property that each canonical variate in the X-domain is uncorrelated with the canonical variates in the Y-domain except the corresponding Y-variate. This is equivalent to requiring that **W** and **V** diagonalize $\mathbf{S}_{XY}$, that is;

$$\mathbf{W}'\mathbf{S}_{XY}\mathbf{V} = \mathbf{E}.$$

where **E** is a diagonal matrix. Biorthogonality implies that the relationship between the X-variables and Y-variables can be partitioned by the pairs of canonical variates, enhancing the interpretability of the analysis.

As an aid to interpreting the canonical variates, researchers often examine structure coefficients (Meredith 1964). The structure coefficients for the X-variables are the correlations between the X-variables and the canonical variates for the X-domain. The matrix of these terms is $\mathbf{S}_{XX}\mathbf{W}$. Similarly, the structure coefficients for the Y-variables are the correlations between

the Y-variables and the canonical variates of the Y-domain, and the matrix of these terms is $\mathbf{S}_{YY}\mathbf{V}$.

2.2.2 Canonical Variate Analysis

Canonical variate analysis (CVA) is a widely used method for analyzing group structure in multivariate data. It is mathematically equivalent to a one-way multivariate analysis of variance (MANOVA) and often goes by the name canonical discriminant analysis. CVA can be interpreted as a special case of canonical correlation analysis where one set of variables consists of group indicators (Gittins 1985). This formulation of CVA as a canonical correlation analysis will be exploited in Chapters Five and Six. A geometrical formulation of CVA given by Campbell and Atchley (1981) will be exploited in Chapter Eight.

To start, however, it is useful to review the traditional formulation of CVA. Krzanowski (1988, page 291) summarizes that the objective of CVA is to, "provide a low-dimensional representation of the data that highlights as accurately as possible the true differences existing between the *m* subsets of points in the full configuration". One finds a weighted sum of the variables whose between-groups variation is maximized with respect to its within-groups variation. That is, find the $p \times 1$ vector $\mathbf{v}_1$ maximizing $\frac{\mathbf{v}_1'\mathbf{B}\mathbf{v}_1}{\mathbf{v}_1'\mathbf{C}\mathbf{v}_1}$, subject to $\mathbf{v}_1'\mathbf{C}\mathbf{v}_1 = 1$, where $\mathbf{B} = \sum_{g=1}^{m} n_g(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})'$, $\mathbf{C} = \sum_{g=1}^{m}\sum_{i=1}^{n_g}(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)'$, $\bar{\mathbf{x}}_g$ is vector of sample means for the g^{th} group and $\bar{\mathbf{x}}$ is the vector means for the entire dataset. $\mathbf{C}$ is referred to as the within-groups covariance matrix and $\mathbf{B}$ as the between-groups covariance matrix. Find further $\mathbf{v}_i$, $i = 2,\ldots,r$, which maximize $\frac{\mathbf{v}_i`\mathbf{B}\mathbf{v}_i}{\mathbf{v}_i`\mathbf{C}\mathbf{v}_i}$ subject to $\mathbf{V}'\mathbf{C}\mathbf{V} = \mathbf{I}$, where $\mathbf{v}_i$ is the i^{th} canonical variate, and $\mathbf{V}$ is a $p \times r$ matrix whose i^{th} column is $\mathbf{v}_i$. Note that $r \le \min(m, p)$ and that $\mathbf{B} + \mathbf{C} = \mathbf{S}_{XX}$.

The matrix of canonical variates is obtained by finding the eigenvectors of $\mathbf{C}^{-1}\mathbf{B}$. However, it is easier to take the eigenvectors of $\mathbf{C}^{-½}\mathbf{B}\mathbf{C}^{-½}$. Denote the matrix of eigenvectors of $\mathbf{C}^{-½}\mathbf{B}\mathbf{C}^{-½}$ by $\mathbf{U}$. Then $\mathbf{V} = \mathbf{C}^{-½}\mathbf{U}$. There are two hypothesis tests of interest. The first is the test that the vectors of group means are equal. The second is the test of dimensionality; i.e., how many canonical variates are statistically significant. For details on these tests see Kshirsagar (1972).

To put CVA in the framework of canonical correlation analysis, create m - 1 binary variables, $x_1,\ldots,x_{m-1}$. If a subject belongs to the s^{th} group, then $x_s = 1$, otherwise $x_s = 0$. Now one can obtain canonical variates as described in Section **2.2.1**.

Campbell and Atchley (1981) give a geometrical formulation of CVA. They model the group means to lie in the subspace defined by the canonical variates, in particular, by $\mathbf{\Sigma V}$, as seen below:

$$\boldsymbol{\mu}_g = \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}\mathbf{V}\mathbf{e}_g, \tag{2.1}$$

for $g = 1,\ldots,m$, where m is the number of groups, p is number of variables, $\boldsymbol{\mu}_g$ is a $p \times 1$ vector of means for the g^{th} group, $\boldsymbol{\mu}_0$ is a $p \times 1$ vector of overall means, $\boldsymbol{\Sigma}$ is a $p \times p$ within-groups covariance matrix, and $\mathbf{e}_g$ is an $f \times 1$ vector of scores of group means on each canonical variate, i.e., $\mathbf{e}_g = \mathbf{X}'\mathbf{V}$. When one assumes multivariate normality and estimates by maximum likelihood, the solution to **V** is the same as given in Section **2.2.1**.

Campbell and Atchley argue that one can view CVA as a principal components analysis performed on the group means in the space obtained by transforming the variables by the Mahalanobis transformation; that is, $\mathbf{x}^* = \mathbf{S}_{XX}^{-½}\mathbf{x}$. In this space Euclidean distance equals Mahalanobis distance, where the Mahalanobis distance between two group means $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ is defined as $(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$. Further, in this space $\mathbf{S}_{X^*X^*} = \mathbf{I}$. The principal components of the group means in this transformed space correspond to the canonical variates. To illustrate these points, consider **Figure 2.1**, which shows a scatterplot of the data from two groups for two variables. Compare **Figure 2.1** to **Figure 2.2**, which shows a scatterplot for the same data in the space of the transformed variables.

To see that this view of CVA leads to (2.1) consider that the positions of the group means in the transformed space is $\boldsymbol{\Sigma}^{-½}\boldsymbol{\mu}_g$, for $g = 1,\ldots,m$. A principal components analysis of the positions of these group means leads to the spectral decomposition of

$$\boldsymbol{\Sigma}^{-½}(\boldsymbol{\mu}_g - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_g - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-½} = \tfrac{n}{g-1}\boldsymbol{\Sigma}^{-½}\mathbf{B}\boldsymbol{\Sigma}^{-½}.$$

The matrix principal components is **U**, where $\mathbf{U} = \boldsymbol{\Sigma}^{½}\mathbf{V}$. Now reverse the transformation by multiplying **U** by $\boldsymbol{\Sigma}^{½}$, and one has the group means in the space defined by $\boldsymbol{\Sigma}\mathbf{V}$.

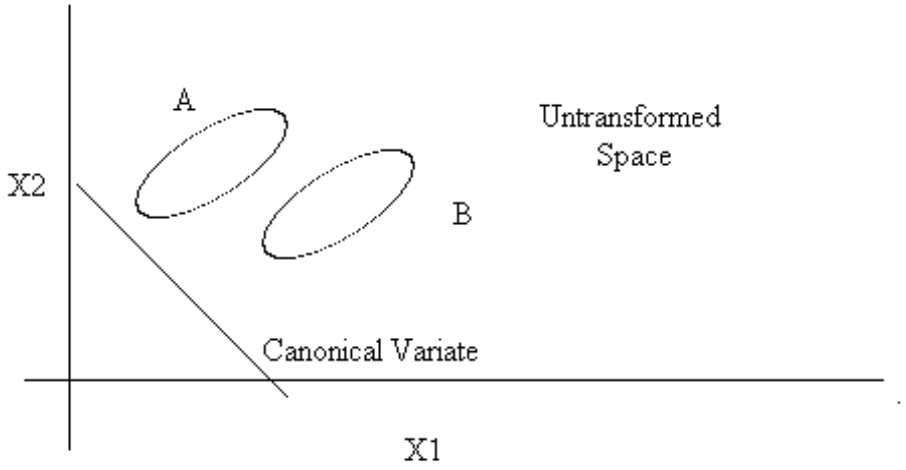

Figure 2.1 Scatterplots in the Untransformed Space

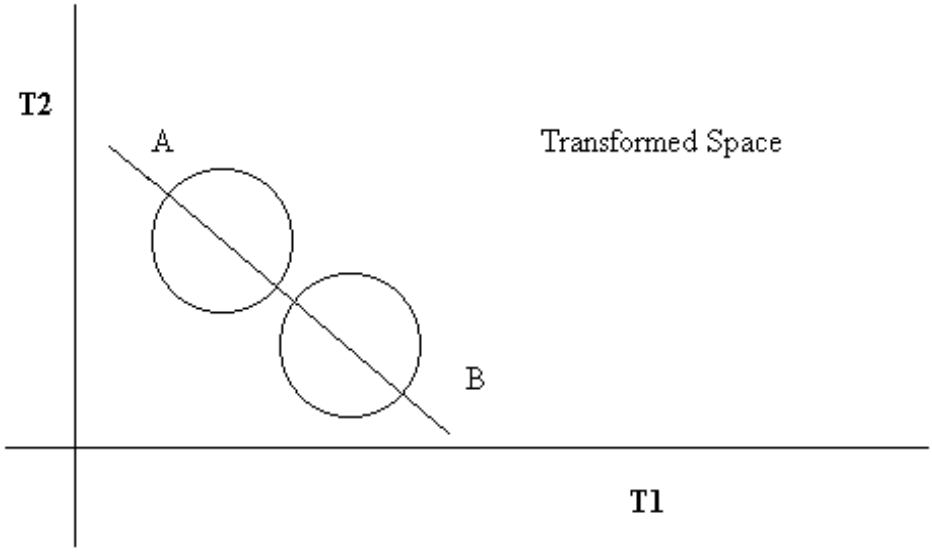

Figure 2.2 Scatterplots in the Transformed Space

2.2.3 Correspondence Analysis

Another method which can be viewed as a special case of CCA is correspondence analysis. Correspondence analysis can be interpreted as a canonical correlation analysis on data where both sets of variables are group indicators; that is, the data are in the form of a two-way

contingency table. Correspondence analysis is an alternative to loglinear models that lends itself well to graphical displays (Greenacre 1984). Some of the methods developed for CCA will be applicable to correspondence analysis, providing a method for generalizing correspondence analysis to longitudinal or multiple group data.

2.2.4 Redundancy Analysis

CCA sometimes finds variates that are correlated but not of practical interest to the researcher because they explain little variance. Redundancy analysis (RA) was devised by Van den Wollenberg (1978) as an alternative to CCA that avoids this problem. RA derives uncorrelated compounds of the X-variables, called redundancy variates, which maximize the variance explained of the Y-variables. Rao (1964) had earlier proposed the same method.

The weights for the redundancy variates, $\mathbf{w}_i$, are determined by the following eigenvalue equation:

$$(\mathbf{S}_{XY}\mathbf{S}_{YX} - \lambda_i \mathbf{S}_{XX})\mathbf{w}_i = 0,$$

where λ_i is the variance explained by the i^{th} variate.

In Van den Wollenberg's original development of RA, only redundancy variates for the X-variables are found. Johansson (1981) extended redundancy analysis by deriving variates in the Y-set that correspond to the redundancy variates in the X-set. Linear combinations of the Y-variables, $\mathbf{v}_i'\mathbf{y}$, are extracted such that the absolute value of $\mathbf{w}_i'\mathbf{S}_{XY}\mathbf{v}_i$ is maximized subject to the constraints $\mathbf{v}_i'\mathbf{v}_i = 1$ and $\mathbf{v}_i'\mathbf{v}_j = 0$ for $i \neq j$. The resulting solution is:

$$\mathbf{v}_i = \lambda_i^{-½}\mathbf{S}_{YX}\mathbf{w}_i$$

where $\lambda_i = \mathbf{w}_i'\mathbf{S}_{XY}\mathbf{S}_{XY}'\mathbf{w}_i$ (Tyler 1982). An alternate solution is to perform a singular value decomposition on $\mathbf{S}_{XX}^{-½}\mathbf{S}_{XY}$,

$$\mathbf{S}_{XX}^{-½}\mathbf{S}_{XY} = \mathbf{W}^*\mathbf{E}\mathbf{V}' \tag{2.2}$$

where $\mathbf{W}^* = \mathbf{S}_{XX}^{½}\mathbf{W}$, $\mathbf{V}$ is the matrix of variates for the Y-variables and $\mathbf{E}$ is the diagonal matrix whose elements are the square roots of λ_i.

RA has the property of biorthogonality, which was defined in Section **2.2.1** for CCA. However, Tyler (1982) showed that RA has an even stronger property. The pairs of redundancy variates additively partition the total variation of the Y-variables that is explained by the X-variables. In other words, all of the variance explained by a redundancy variate $\mathbf{w}_i'\mathbf{x}$ is associated with one vector in the Y-set, $\mathbf{v}_i'\mathbf{y}$. This property also holds for CCA. However, the redundancy variates partition the variance in an optimal way, as they successively maximize the variance explained in the Y-variables.

Redundancy analysis (RA) has analogues to canonical variate analysis. When the X-variables are dummy variables indicating group membership, as in canonical variate analysis, RA yields a procedure similar to canonical variate analysis. However, it differs in that RA determines pairs of variates that maximize the between-group variance, whereas CCA determines pairs of variates that maximize the ratio of the between-group variance to the within-group variance. One can also look at RA with this kind of data as a principal components analysis on

the group means in the untransformed space, with the principal components corresponding to the redundancy variates.

2.2.5 Procrustes Rotation

Procrustes analysis (Gower 1975) is an analysis where two sets of variables measured on the same units are translated, dilated and rotated such that the point configurations are as similar as possible in a least squares sense. The interest in this section is in the rotation part of the Procrustes analysis, which shall be referred to simply as Procrustes rotation (PR). PR is closely related to CCA and RA. In later chapters it will be seen to be more suitable than CCA or RA for particular kinds of data when extended to a three-mode model.

A Procrustes analysis usually starts with a translation of the data. But that is obviated in this discussion by the centering of both sets of variables. Since PR can be performed independently of the dilation, as will be shown shortly, I start with the rotation. If **X** and **Y** are mean centered, Procrustes rotation finds the orthogonal matrix **Q** such that m is minimized, where: $m = \sum_i \sum_j \left(x_{ij} - y^*_{ij} \right)^2$ and $\mathbf{Y}^* = \left[y^*_{ij} \right] = \mathbf{YQ}$. **Q** can be shown to be: $\mathbf{Q} = \mathbf{VW}'$, where one derives **W** and **V** by performing a SVD on $\mathbf{S}_{XY}$, so that

$$\mathbf{S}_{XY} = \mathbf{WEV}'.$$

Now one sees that if one performs a dilation on either **x** or **y** by multiplying by a scalar c, one obtains the same **W** and **V** as c is factored into the matrix of singular values, **E**.

Like CCA and RA, Procrustes rotation can be viewed as a method which finds pairs of variates that relate two sets of variables. These pairs of variates are orthogonal and maximize the covariance. To see this, note that the SVD of $\mathbf{S}_{XY}$ is equivalent to successively finding pairs of $\mathbf{w}_i$ and $\mathbf{v}_i$ such that $\mathbf{w}_i' \mathbf{S}_{XY} \mathbf{v}_i$ is maximized; but $\mathbf{w}_i' \mathbf{S}_{XY} \mathbf{v}_i$ is just the covariance of $\mathbf{w}_i' \mathbf{x}$ and $\mathbf{v}_i' \mathbf{y}$. Thus **E** is a diagonal matrix with the covariances of $\mathbf{w}_i' \mathbf{x}$ and $\mathbf{v}_i' \mathbf{y}$ in the i^{th} diagonal position. For the rest of the dissertation when Procrustes rotation is referred to it shall be in the sense of finding orthogonal variate pairs such that the covariance of the pairs is maximized.

At this point it is worth emphasizing the unity of form of CCA, RA and PR. All are derived from a SVD of a transformation of $\mathbf{S}_{XY}$. In CCA both the X-variables and Y-variables are transformed by their respective Mahalanobis transformations and $\mathbf{S}_{XX}^{-1/2} \mathbf{S}_{XY} \mathbf{S}_{YY}^{-1/2}$ is decomposed. In RA only the X-variables are transformed and $\mathbf{S}_{XX}^{-1/2} \mathbf{S}_{XY}$ is decomposed. In PR neither the X-variables nor the Y-variables are transformed before decomposition. This unity of form will be exploited in the generalizations of CCA, RA and PR to three-mode data in Chapter Four.

2.3 THREE-MODE PRINCIPAL COMPONENT ANALYSIS

Three-mode PCA is a method which I will use to generalize CCA, RA and PR to three-mode data. The discussion on three-mode PCA in this section is derived from Kroonenberg (1983). A mode is an index for the data. Traditional PCA has two modes, typically the subject and variable modes. Each measurement is indexed by subject and variable (the units shall be

referred to as subjects, as is the convention in three-mode PCA). A third mode is a third index for the data, such as conditions or occasions. An example of a three-mode observation would be to measure a subject on a variable on a given occasion (or under a given condition).

2.3.1 The Tucker3 Model

Tucker (1966) generalized the SVD to three-mode data with his Tucker3 and Tucker2 models. The more general of the two models is the Tucker3, which decomposes a three-way array into a set of orthonormal vectors (components) for each mode and a three-dimensional core matrix ("box") of values relating these components. The components are a weighted sum of subjects, variables or occasions, and are interpreted the same way as are principal components, that is, as summaries of variation. The core box is analogous to the matrix of eigenvalues in a SVD, although the core box is generally not diagonal. Let the modes be indexed by i, j, k. Typically, x_{ijk} could stand for the value of variable j on condition (or time) k for subject i; hence $\underline{\mathbf{X}} = [x_{ijk}]$. Let n indicate the number of measurements in the i-mode, i.e. the number of subjects; let m indicate the number of measurements in the j-mode, i.e. the number of variables; and let l indicate the number measurements in the k-mode, i.e. the number of conditions or occasions. Further, let s indicate the number of components for the i-mode, t the number of components for the j-mode, and u the number of components for the k-mode. Let **G** denote the $n \times s$ matrix of components for the i-mode, **H** the $m \times t$ matrix of components for the j-mode, **E** the $l \times u$ matrix of components for the k-mode, and $\underline{\mathbf{C}}$ the $s \times t \times u$ core box. Without loss of generality the matrices **G**, **H** and **E** are specified to be columnwise orthonormal to identify the solution.

The Tucker3 model is expressed in terms of a single observation as follows:

$$x_{ijk} = \sum_{p=1}^{s} \sum_{q=1}^{t} \sum_{r=1}^{u} g_{ip} h_{jq} e_{kr} c_{pqr} .$$

To express the three-mode decomposition in matrix form it is necessary to reformulate $\underline{\mathbf{X}}$ and $\underline{\mathbf{C}}$. Let **X** be the $n \times lm$ matrix formed by laying out the l $n \times m$ subject $\times$ variable data matrices side by side. Let **C** be the $s \times ut$ matrix formed by laying out the u $s \times t$ core matrices side by side. Then

$$\mathbf{X} = \mathbf{GC}(\mathbf{H}' \otimes \mathbf{E}'),$$

where $\otimes$ is the Kronecker product (see Appendix One).

2.3.2 Special Cases of Three-Mode Principal Components

A special case of the Tucker3 model is the Tucker2 model (Kroonenberg 1983), which restricts **E** to be the identity matrix. One can specify the Tucker2 model in terms of matrices as follows:

$$\mathbf{X}_k = \mathbf{GC}_k\mathbf{H}', \qquad k = 1,\ldots,l.$$

If one makes the further restriction that the $\mathbf{C}_k$ matrices be diagonal, one has the PARAFAC (PARAllel FACtor analysis, Harshman 1970) or Candecomp (CANonical DECOMPosition, Carroll and Chang 1970) model, henceforth referred to simply as PARAFAC. The PARAFAC model does not assume orthonormal columns for the components matrices **G** and **H**, and, in

contrast to the Tucker2 and Tucker3 models, to require such results in a loss in generality. However, the PARAFAC model with the restriction of orthonormality on the two sets of components will play an important role in future chapters. It shall be henceforth referred to as the PARAFAC (orth.) model.

Three-mode PCA models were envisaged for the situation where one wants to perform a PCA or factor analysis on data that was not only multivariate but had measurements taken over multiple occasions or conditions. However, they can also model data from multiple datasets or groups. One way to accomplish this is to model the crossproduct matrices, typically the covariance matrices, for each dataset (Kiers 1991). This is the approach taken in Chapter Three on common principal components. Modeling covariances implies symmetry between the **G** and **H** components. If the PARAFAC model is restrained so that $\mathbf{G} = \mathbf{H}$ one has the INDSCAL (INDividual SCALing, Carroll and Chang 1970) model.

2.4 THE CAMPBELL AND TOMENSON MODEL

Campbell and Tomenson (1983) hypothesize common canonical variates over multiple datasets. Their model is both a competitor of some of the models to be presented later and a starting point for them. In Campbell and Tomenson's formulation the common canonical variates build a reduced space in which the group means for each dataset are located. They present a hierarchy of models from most general to most specific. Briefly outlined, the most general model is that the canonical variates differ in each dataset. The next most general model is that the canonical variates are common in each dataset, but the location of the group means on them differ. The most specific model is that the variates are the same and the group means have common coordinates on them. A more detailed statement of these models follows.

The model in general form is:

$$\boldsymbol{\mu}_{gk} = \boldsymbol{\mu}_{0k} + \boldsymbol{\Sigma}\mathbf{V}_k\mathbf{e}_{gk}$$

where p indicates the number of variables, $\boldsymbol{\mu}_{gk}$ is the $p \times 1$ vector of means for the g^{th} group in the k^{th} dataset, $\boldsymbol{\mu}_{0k}$ is the $p \times 1$ vector of overall group means for the k^{th} dataset, $\boldsymbol{\Sigma}$ is the $p \times p$ within-group covariance matrix assumed common over group and dataset, $\mathbf{V}_k$ is the $p \times r$ matrix whose r columns are the canonical variates for dataset t, and $\mathbf{e}_{gk}$ is an $r \times 1$ vector specifying the coordinates on the r canonical variates for the g^{th} group mean in the k^{th} dataset. The three models described above are:

1. $\mathbf{V}_i \neq \mathbf{V}_j$ and $\mathbf{e}_{gi} \neq \mathbf{e}_{gj}$ for $i \neq j$; the canonical variates differ over dataset.
2. $\mathbf{V}_i = \mathbf{V}_j$ but $\mathbf{e}_{gi} \neq \mathbf{e}_{gj}$ for $i \neq j$; the canonical variates are common to each dataset, but the coordinates of the group means on the variates differ.
3. $\mathbf{V}_i = \mathbf{V}_j$ and $\mathbf{e}_{gi} = \mathbf{e}_{gj}$, but $\boldsymbol{\mu}_{0i} \neq \boldsymbol{\mu}_{0j}$ for $i \neq j$; the canonical variates are common to each dataset, the coordinates of the group means on the variates are the same, but the overall center of the means is different.

Campbell and Tomenson assume that the data are normally distributed, then derive maximum likelihood estimates of the parameters and hypothesis tests based on likelihood ratios.

2.5 SUMMARY

This chapter provides the starting point for the rest of the dissertation. It presents those multivariate methods I want to extend to longitudinal data, which are CCA, CVA, RA and PR. It presents two models that I will build on to achieve this: three-mode PCA, which is a flexible least squares method which is suitable for graphical displays such as joint plots: and Campbell and Tomenson's model, which is based on maximum likelihood methods.

CHAPTER THREE

PRELIMINARY RESULTS FOR PARAFAC WITH ORTHOGONALITY CONSTRAINTS

3.1 INTRODUCTION

In this chapter I discuss some preliminary results pertaining to the PARAFAC model with orthogonality constraints, which will be henceforth referred to as PARAFAC (orth.). These results shall be important for both Chapter Four, which is on Common Principal Components, and Chapter Six, which is on using three-mode methods for CCA, CVA, RA and PR over time. In Section **3.2** I show certain properties of the PARAFAC (orth.) model related to the sums of squares of fit and error. These properties will allow me in Chapter Six to show that the CVA, CCA, RA and PR/time models are optimal in maximizing sums of squared correlations, variance explained, or sums of squared covariances, depending on the method. In Section **3.3** I prove that the PARAFAC (orth.) solutions are nested.

3.2 OPTIMALITY PROPERTIES OF THE PARAFAC MODEL WITH ORTHOGONALITY CONSTRAINTS

Kroonenberg (1983) shows that the least squares solutions to the Tucker2 and Tucker3 models have certain useful properties with respect to the sums of squares. He shows that the total sums of squares can be partitioned into sums of squares fit and sums of squares residuals. He also shows that the sums of the squared elements of the core matrices equal the sums of squares fit. In particular, the square of the i^{th}, j^{th}, k^{th} core element is the fit contributed by the combination of the i^{th} element of the first (subject) mode, the j^{th} element of the second (variable) mode and the k^{th} element of the third (occasion) mode. An important consequence is that minimizing the sums of squares lack of fit is equivalent to maximizing the sums of squares fit, and thus equivalent to maximizing the sums of squares of the elements of the core matrix. In this section I show that these properties hold true for the PARAFAC (orth.) model. Of particular importance to future developments is **Proposition 3.2**.

I start with some necessary definitions. Define $\underline{\mathbf{C}}$ to be an $m \times n \times p$ three-way array, $\mathbf{C}_k = \underline{\mathbf{C}}[,,k]$, $k = 1,\ldots,p$, to be the p $m \times n$ slices of $\underline{\mathbf{C}}$, $\mathbf{F}$ to be an $m \times m$ orthogonal matrix, $\mathbf{G}$ to be an $n \times n$ orthogonal matrix, and $\mathbf{D}_k$ to be the $m \times n$ matrix $\mathbf{D}_k = \mathbf{F}'\mathbf{C}_k\mathbf{G}$. Further define $\mathbf{D}_k[i,j]$ to be the element in the i^{th} row and j^{th} column of $\mathbf{D}_k$, $\mathbf{f}_i$ to be the i^{th} column of $\mathbf{F}$, and $\mathbf{g}_j$ to be the j^{th} column of $\mathbf{G}$.

I proceed with two identities. First, one has the following decomposition of $\mathbf{C}_k$:

$$\mathbf{C}_k = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{g}_i\mathbf{f}_j'\mathbf{D}_k[i,j].$$

Second, the total sums of squares of $\mathbf{D}_k$ equals the total sums of squares of $\mathbf{C}_k$:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{C}_k^2[i,j] = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{D}_k^2[i,j], \tag{3.1}$$

since

$$\text{trace}(\mathbf{C}_k'\mathbf{C}_k) = \text{trace}(\mathbf{G}'\mathbf{C}_k'\mathbf{F}\mathbf{F}'\mathbf{C}_k\mathbf{G}) = \text{trace}(\mathbf{D}_k'\mathbf{D}_k).$$

Define a rank-one approximation to $\mathbf{C}_k$ as $\hat{\mathbf{C}}_k = \mathbf{f}_i\mathbf{g}_j'h_{ij}$. Note that by a theorem by Penrose (1955) for given normal $\mathbf{f}_i$ (i.e., $\|\mathbf{f}_i\|^2 = 1$) and $\mathbf{g}_j$, the optimal h_{ij} is $h_{ij} = \mathbf{f}_i'\mathbf{C}_k\mathbf{g}_j = \mathbf{D}_k[i,j]$. Let T be a non-zero set of combinations of i,j, $1 \le i \le m$, and $1 \le j \le n$. Then the sum of the rank-one approximations to $\mathbf{C}_k$ defined by $\mathbf{f}_i$ and $\mathbf{g}_j$, $(i,j) \in T$, and $h_{ij} = \mathbf{D}_k[i,j]$, is itself an approximation to $\mathbf{C}_k$. Define this as $\hat{\mathbf{C}}_k^T = \sum_{(i,j)\in T} \mathbf{f}_i\mathbf{g}_j'\mathbf{D}_k[i,j]$.

Proposition 3.1. For the modeling of $\mathbf{C}_k$ by $\hat{\mathbf{C}}_k^T$, the sums of squares total is additively partitioned into the sums of squares fit and sums of squares lack of fit.

Proof: The sums of squares fit is

$$\sum_{(i,j)\in T}(\hat{\mathbf{C}}_k^T[i,j])^2 = \text{trace}(\hat{\mathbf{C}}_k^{T\prime}\hat{\mathbf{C}}_k^T) = \text{trace}\left(\sum_{(i,j)\in T}\mathbf{f}_i\mathbf{g}_j'\mathbf{D}_k[i,j]\right)'\left(\sum_{(i,j)\in T}\mathbf{f}_i\mathbf{g}_j'\mathbf{D}_k[i,j]\right)$$

$$= \sum_{(i,j)\in T}\mathbf{D}_k^2[i,j]\text{trace}\left(\mathbf{f}_i\mathbf{g}_j'\right)'\mathbf{f}_i\mathbf{g}_j'$$

$$= \sum_{(i,j)\in T}\mathbf{D}_k^2[i,j]. \qquad (3.2)$$

The sums of squares lack of fit is:

$$= \text{trace}(\mathbf{C}_k - \sum_{(i,j)\in T}\mathbf{f}_i\mathbf{g}_j'\mathbf{D}_k[i,j])'(\mathbf{C}_k - \sum_{(i,j)\in T}\mathbf{f}_i\mathbf{g}_j'\mathbf{D}_k[i,j])$$

$$= \text{trace}(\mathbf{C}_k'\mathbf{C}_k) + \sum_{(i,j)\in T}\text{trace}(\mathbf{f}_i\mathbf{g}_j'\mathbf{g}_j\mathbf{f}_i'\mathbf{D}_k^2[i,j]) - \sum_{(i,j)\in T}2\text{trace}(\mathbf{C}_k'\mathbf{f}_i\mathbf{g}_j'\mathbf{D}_k[i,j])$$

$$= \text{trace}(\mathbf{C}_k'\mathbf{C}_k) + \sum_{(i,j)\in T}\mathbf{D}_k^2[i,j] - \sum_{(i,j)\in T}2\mathbf{D}_k[i,j]\text{trace}(\mathbf{G}\mathbf{D}_k\mathbf{F}'\mathbf{f}_i\mathbf{g}_j')$$

$$= \text{trace}(\mathbf{C}_k'\mathbf{C}_k) + \sum_{(i,j)\in T}\mathbf{D}_k^2[i,j] - \sum_{(i,j)\in T}2\mathbf{D}_k[i,j]\text{trace}(\mathbf{g}_j'\mathbf{G}\mathbf{D}_k'[i,])$$

$$= \text{trace}(\mathbf{C}_k'\mathbf{C}_k) - \sum_{(i,j)\in T}\mathbf{D}_k^2[i,j].$$

Clearly, the sums of squares fit and sums of squares lack of fit equal the sums of squares total. Ž

For a given $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{C}_k$, $k = 1,\ldots,p$, one can consider the class of $\hat{\mathbf{C}}_k^T$, $(i,j)\in T$, as a class of models. I shall refer to these as orthogonal models. Thus by **Proposition 3.1** for an orthogonal model the sums of squares total can be partitioned into a sums of squares fit and a sums of squares error.

With **Proposition 3.1** in place I move to **Proposition 3.2**. Denote the rank-*r* PARAFAC (orth.) solution to $\mathbf{C}_k$ as $\mathbf{F}^*$, $\mathbf{G}^*$ and $\mathbf{H}^*$, where $\mathbf{F}^*$ is an $m\times r$ columnwise orthonormal matrix, $\mathbf{G}^*$ is an $n\times r$ columnwise orthonormal matrix and $\mathbf{H}^*$ a $p\times r$ diagonal matrix, $r\le m,n,p$. $\mathbf{H}^*$ can also be expressed as p $r\times r$ diagonal matrices $\mathbf{H}_k$, where $k = 1,\ldots,p$. This is the form of $\mathbf{H}^*$ which will be used below. Recall it is known (Kroonenberg 1983) that $\mathbf{H}_k = \text{diag}(\mathbf{F}^{*\prime}\mathbf{C}_k\mathbf{G}^*)$.

Proposition 3.2. The least squares estimates of PARAFAC (orth.) are $\mathbf{F}^*$ and $\mathbf{G}^*$ such that $\sum_{k=1}^{g} \text{trace}\left(\mathbf{H}_k^2\right)$ is maximized.

Proof: By definition PARAFAC (orth.) minimizes the total sums of squares lack of fit. Since the PARAFAC (orth.) is in the class of orthogonal models as defined above, by **Proposition 3.1** this is equivalent to maximizing the sums of squares fit. Further, by (3.2) $\sum_{k=1}^{g} \text{trace}\left(\mathbf{H}_k^2\right)$ represents that sums of squares fit. Ž

3.3 THE NESTEDNESS PROPERTY OF PARAFAC SOLUTIONS WITH ORTHOGONALITY CONSTRAINTS

A solution is called nested if the rank-$(f-1)$ solution is a subset of the rank-*f* solution, for any realizable *f*. This implies that one can find any rank-*f* solution recursively by finding *f* rank-one solutions. An example of the nestedness of solutions is the singular value decomposition (Eckart & Young) of a real matrix. A property of real matrices is that the least squares rank-*p* approximation to a matrix can be found by determining *p* rank-one approximations. For example, the best rank-two approximation of a matrix is the sum of its rank-one approximation plus the rank-one approximation to the matrix obtained by subtracting the first rank-one approximation from the original matrix.

The nestedness property is important to modeling with PARAFAC (orth.) because it allows for straightforward comparisons between solutions of different ranks, enabling one to examine the fit attributable to each component. For example, without the nestedness property the component of a rank-one solution may bear no relation to the components of a rank-two solution.

A limited result pertaining to the nestedness of PARAFAC solutions has already been achieved by Leurgans and Ross (1992). They show that a necessary condition for the existence of the nestedness property for a PARAFAC model with an exact rank-two solution is that at least two of the three pairs of components be orthonormal. The following result goes further in that it states that a sufficient condition for the existence of the nestedness property for a PARAFAC model of any rank and of imperfect fit is that two of the three matrices of components be orthonormal.

The subsequent proof will take advantage of conditional linearity. The property of conditional linearity is said to exist if the set of all parameters can be divided into subsets such that in the model each subset is linear in terms of the rest. In the case of the PARAFAC (orth.) model one such division would be the matrices of parameters **F**, **G** and **H**. Because any given subset of parameters is part of the optimal least squares solution, the estimates for those parameters must be the solution to the regression problem where the estimates for the rest of the parameters are treated as fixed. For example, the estimate for **G** must be the solution to the regression problem where $\underline{\mathbf{X}}$ is the response, and **H** and **C** are fixed in the model. This fact is called conditional

linearity. Conditional linearity is the basis for the alternating least squares algorithms that are used to obtain least squares estimates for the three-mode models.

As a preliminary I will define trilinear notation that will simplify the presentation. Let **a**, **b** and **c** be vectors. Then the trilinear multiplication of **a**, **b** and **c**, denoted as $\mathbf{a}\times\mathbf{b}\times\mathbf{c}$, is the outer product of **a**, **b** and **c** and generates a three-way array, $\underline{\mathbf{U}}$, such that $\underline{\mathbf{U}}[\mathrm{i},\mathrm{j},\mathrm{k}]=\mathrm{a}_\mathrm{i}\mathrm{b}_\mathrm{j}\mathrm{c}_\mathrm{k}$, where a_i, b_j, and c_k are the i^{th}, j^{th} and k^{th} elements of **a**, **b** and **c**. Another way to think of $\mathbf{a}\times\mathbf{b}\times\mathbf{c}$ is that the k^{th} slice of $\underline{\mathbf{U}}$ is $\mathbf{ab}'$ multiplied by c_k: $\underline{\mathbf{U}}[,,\mathrm{k}]=\mathbf{ab}'\mathrm{c}_\mathrm{k}$.

Theorem 3.1: Let $\underline{\mathbf{X}}$ be a $p\times q\times m$ array. Assume **A**, **B** and **C** represent a rank-*f* PARAFAC solution to $\underline{\mathbf{X}}$ with **A** and **B** constrained to be columnwise orthonormal and the columns of **A**, **B** and **C** ordered by the norm of $\mathbf{c}_\mathrm{k}$. That is: $\hat{\underline{\mathbf{X}}}=\mathbf{a}_1\times\mathbf{b}_1\times\mathbf{c}_1+\mathbf{a}_2\times\mathbf{b}_2\times\mathbf{c}_2+\ldots+\mathbf{a}_f\times\mathbf{b}_f\times\mathbf{c}_f$, where $\mathbf{a}_\mathrm{g}=\mathbf{A}[,\mathrm{g}]$, $\mathbf{b}_\mathrm{g}=\mathbf{B}[,\mathrm{g}]$, and $\mathbf{c}_\mathrm{g}=\mathbf{C}[,\mathrm{g}]$ for $\mathrm{g}=1,\ldots,f$, and $\|\mathbf{c}_\mathrm{k}\|^2\le\|\mathbf{c}_{\mathrm{k}'}\|^2$ for $1\le\mathrm{k}<\mathrm{k}'\le f$. Then the best rank-*d* solution, $1\le d\le f$, is $\mathbf{a}_1\times\mathbf{b}_1\times\mathbf{c}_1+\mathbf{a}_2\times\mathbf{b}_2\times\mathbf{c}_2+\ldots+\mathbf{a}_d\times\mathbf{b}_d\times\mathbf{c}_d$.

Proof: Let $\hat{\mathbf{a}}_1\times\hat{\mathbf{b}}_1\times\hat{\mathbf{c}}_1+\hat{\mathbf{a}}_2\times\hat{\mathbf{b}}_2\times\hat{\mathbf{c}}_2+\ldots+\hat{\mathbf{a}}_d\times\hat{\mathbf{b}}_d\times\hat{\mathbf{c}}_d$ be the rank-*d* estimates. Start with the conditional linearity of these estimates. The estimate for any $\hat{\mathbf{a}}_\mathrm{e}$, $1\le\mathrm{e}\le d$, must be the solution to the regression problem where $\underline{\mathbf{X}}$ is the response and $\hat{\mathbf{a}}_\mathrm{h}$, $\mathrm{h}=1,\ldots,d$, $\mathrm{h}\ne\mathrm{e}$, and $\hat{\mathbf{b}}_\mathrm{h}$ and $\hat{\mathbf{c}}_\mathrm{h}$, $\mathrm{h}=1,\ldots,d$, are fixed. This is seen below in (3.3), where $\hat{\mathrm{a}}_{\mathrm{ei}}$ denotes the i^{th} element of the vector $\hat{\mathbf{a}}_\mathrm{e}$. This regression yields the normal equations given in (3.4) or (3.5). Likewise one can consider $\hat{\mathbf{a}}_\mathrm{e}$ and $\hat{\mathbf{b}}_\mathrm{e}$ to be fixed and solve for $\hat{\mathbf{c}}_\mathrm{e}$ (3.6), and one can consider $\hat{\mathbf{a}}_\mathrm{e}$ and $\hat{\mathbf{c}}_\mathrm{e}$ to be fixed and solve for $\hat{\mathbf{b}}_\mathrm{e}$ (3.7). The least squares solution must simultaneously solve these three regressions.

$$\hat{\mathrm{a}}_{ei}\begin{pmatrix}\hat{\mathrm{b}}_{e1}\hat{\mathrm{c}}_{e1}\\ \vdots\\ \hat{\mathrm{b}}_{eq}\hat{\mathrm{c}}_{e1}\\ \vdots\\ \hat{\mathrm{b}}_{e1}\hat{\mathrm{c}}_{em}\\ \vdots\\ \hat{\mathrm{b}}_{eq}\hat{\mathrm{c}}_{em}\end{pmatrix}=\begin{pmatrix}\underline{\mathbf{X}}[\mathrm{i},1,1]\\ \vdots\\ \underline{\mathbf{X}}[\mathrm{i},q,1]\\ \vdots\\ \underline{\mathbf{X}}[\mathrm{i},1,m]\\ \vdots\\ \underline{\mathbf{X}}[\mathrm{i},q,m]\end{pmatrix}-\sum_{\substack{s=1\\ s\ne e}}^{d}\hat{\mathrm{a}}_{qi}\begin{pmatrix}\hat{\mathrm{b}}_{s1}\hat{\mathrm{c}}_{s1}\\ \vdots\\ \hat{\mathrm{b}}_{sq}\hat{\mathrm{c}}_{s1}\\ \vdots\\ \hat{\mathrm{b}}_{s1}\hat{\mathrm{c}}_{sm}\\ \vdots\\ \hat{\mathrm{b}}_{sq}\hat{\mathrm{c}}_{sm}\end{pmatrix}. \tag{3.3}$$

The normal equations for solving for $\hat{\mathrm{a}}_{\mathrm{ei}}$, for $\mathrm{i}=1,\ldots,p$, are:

$$\sum_{\mathrm{j},\mathrm{k}}\left(\underline{\mathbf{X}}[\mathrm{i},\mathrm{j},\mathrm{k}]-\hat{\mathrm{a}}_{1\mathrm{i}}\hat{\mathrm{b}}_{1\mathrm{j}}\hat{\mathrm{c}}_{1\mathrm{k}}-\ldots-\hat{\mathrm{a}}_{\mathrm{ei}}\hat{\mathrm{b}}_{\mathrm{ej}}\hat{\mathrm{c}}_{\mathrm{ek}}\right)\hat{\mathrm{b}}_{\mathrm{ej}}\hat{\mathrm{c}}_{\mathrm{ek}}=0 \tag{3.4}$$

or in vector form:

$$\sum_{j,k}\left(\underline{\mathbf{X}}[,j,k]-\hat{\mathbf{a}}_1\hat{b}_{1j}\hat{c}_{1k}-\ldots-\hat{\mathbf{a}}_e\hat{b}_{ej}\hat{c}_{ek}\right)\hat{b}_{ej}\hat{c}_{ek}=\mathbf{0}. \tag{3.5}$$

Likewise, one has normal equations for solving for $\hat{\mathbf{b}}_e$ conditioned on $\hat{\mathbf{a}}_e$ and $\hat{\mathbf{c}}_e$ being fixed:

$$\sum_{i,k}\left(\underline{\mathbf{X}}[i,,k]-\hat{a}_{1i}\hat{\mathbf{b}}_1\hat{c}_{1k}-\ldots-\hat{a}_{ei}\hat{\mathbf{b}}_e\hat{c}_{ek}\right)\hat{a}_{ei}\hat{c}_{ek}=\mathbf{0} \tag{3.6}$$

and one has normal equations for solving for $\hat{\mathbf{c}}_e$ conditioned on $\hat{\mathbf{a}}_e$ and $\hat{\mathbf{b}}_e$ being fixed:

$$\sum_{i,j}\left(\underline{\mathbf{X}}[i,j,]-\hat{a}_{1i}\hat{b}_{1j}\hat{\mathbf{c}}_1-\ldots-\hat{a}_{ei}\hat{b}_{ej}\hat{\mathbf{c}}_e\right)\hat{a}_{ei}\hat{b}_{ej}=\mathbf{0}. \tag{3.7}$$

Let $\underline{\mathbf{E}}=\underline{\mathbf{X}}-\underline{\hat{\mathbf{X}}}$. Then the normal equations for $\hat{\mathbf{a}}_e$, $\hat{\mathbf{b}}_e$ and $\hat{\mathbf{c}}_e$ can be written as

$$\sum_{j,k}(\mathbf{a}_1b_{1j}c_{1k}+\mathbf{a}_2b_{2j}c_{2k}+\cdots+\mathbf{a}_fb_{fj}c_{fk}+\underline{\mathbf{E}}[,j,k]-\hat{\mathbf{a}}_1\hat{b}_{1j}\hat{c}_{1k}-\ldots-\hat{\mathbf{a}}_e\hat{b}_{ej}\hat{c}_{ek})\hat{b}_{ej}\hat{c}_{ek}=\mathbf{0},$$

$$\sum_{i,k}(a_{1i}\mathbf{b}_1c_{1k}+a_{2i}\mathbf{b}_2c_{2k}+\cdots+a_{fi}\mathbf{b}_fc_{fk}+\underline{\mathbf{E}}[i,,k]-\hat{a}_{1i}\hat{\mathbf{b}}_1\hat{c}_{1k}-\ldots-\hat{a}_{ei}\hat{\mathbf{b}}_e\hat{c}_{ek})\hat{a}_{ei}\hat{c}_{ek}=\mathbf{0},$$

$$\sum_{i,j}(a_{1i}b_{1j}\mathbf{c}_1+a_{2i}b_{2j}\mathbf{c}_2+\cdots+a_{fi}b_{fj}\mathbf{c}_f+\underline{\mathbf{E}}[i,j,]-\hat{a}_{1i}\hat{b}_{1j}\hat{\mathbf{c}}_1-\ldots-\hat{a}_{ei}\hat{b}_{ej}\hat{\mathbf{c}}_e)\hat{a}_{ei}\hat{b}_{ej}=\mathbf{0}.$$

Collecting terms in j and k allows these normal equations to be written as follows:

$$\mathbf{a}_1\left(\sum_j b_{1j}\hat{b}_{ej}\right)\left(\sum_k c_{1k}\hat{c}_{ek}\right)+\cdots+\mathbf{a}_f\left(\sum_j b_{fj}\hat{b}_{ej}\right)\left(\sum_k c_{fk}\hat{c}_{ek}\right)$$
$$-\hat{\mathbf{a}}_1\left(\sum_j \hat{b}_{1j}\hat{b}_{ej}\right)\left(\sum_k \hat{c}_{1k}\hat{c}_{ek}\right)-\ldots-\hat{\mathbf{a}}_e\left(\sum_j \hat{b}_{ej}\hat{b}_{ej}\right)\left(\sum_k \hat{c}_{ek}\hat{c}_{ek}\right)+\sum_{j,k}\underline{\mathbf{E}}[,j,k]\hat{b}_{ej}\hat{c}_{ek}=\mathbf{0} \tag{3.8}$$

$$\mathbf{b}_1\left(\sum_i a_{1j}\hat{a}_{ei}\right)\left(\sum_k c_{1k}\hat{c}_{ek}\right)+\cdots+\mathbf{b}_f\left(\sum_i a_{fi}\hat{a}_{ei}\right)\left(\sum_k c_{fk}\hat{c}_{ek}\right)$$
$$-\hat{\mathbf{b}}_1\left(\sum_i \hat{a}_{1i}\hat{a}_{ei}\right)\left(\sum_k \hat{c}_{1k}\hat{c}_{ek}\right)-\ldots-\hat{\mathbf{b}}_e\left(\sum_i \hat{a}_{ei}\hat{a}_{ei}\right)\left(\sum_k \hat{c}_{ek}\hat{c}_{ek}\right)+\sum_{i,k}\underline{\mathbf{E}}[i,,k]\hat{a}_{ei}\hat{c}_{ek}=\mathbf{0} \tag{3.9}$$

$$\mathbf{c}_1\left(\sum_i a_{1i}\hat{a}_{ei}\right)\left(\sum_j b_{1j}\hat{b}_{ej}\right)+\cdots+\mathbf{c}_f\left(\sum_i a_{fi}\hat{a}_{ei}\right)\left(\sum_j b_{fj}\hat{b}_{ej}\right)$$
$$-\hat{\mathbf{c}}_1\left(\sum_i \hat{a}_{1i}\hat{a}_{ei}\right)\left(\sum_j \hat{b}_{1j}\hat{b}_{ej}\right)-\ldots-\hat{\mathbf{c}}_e\left(\sum_i \hat{a}_{ei}\hat{a}_{ei}\right)\left(\sum_j \hat{b}_{ej}\hat{b}_{ej}\right)+\sum_{i,j}\underline{\mathbf{E}}[i,j,]\hat{a}_{ei}\hat{b}_{ej}=\mathbf{0}. \tag{3.10}$$

Several of the terms in equations 3.8, 3.9 and 3.10 drop out. By the definition of the normal equations $\sum_{j,k}\underline{\mathbf{E}}[,j,k]\hat{b}_{ej}\hat{c}_{ek}=\sum_{i,k}\underline{\mathbf{E}}[i,,k]\hat{a}_{ej}\hat{c}_{ek}=\sum_{i,j}\underline{\mathbf{E}}[i,j,]\hat{a}_{ei}\hat{b}_{ek}=\mathbf{0}$. Because of the orthonormality of $\hat{\mathbf{a}}_g$ and of $\hat{\mathbf{b}}_g$, $g=1,\ldots,d$, one has $\left(\sum_i \hat{a}_{gi}\hat{a}_{ei}\right)=\left(\sum_j \hat{b}_{gj}\hat{b}_{ej}\right)=0$ if $g\neq e$ or 1 if $g=e$.

Consider that $\sum_{i} a_{mi}\hat{a}_{ni} = \cos ine(\alpha_{mn})$, where α_{mn} is the angle subtended between $\mathbf{a}_m$ and $\hat{\mathbf{a}}_n$, and $\sum_{j} b_{mj}\hat{b}_{nj} = \cos ine(\beta_{mn})$, where β_{mn} is the angle subtended between $\mathbf{b}_m$ and $\hat{\mathbf{b}}_n$. Then from (3.10) $\hat{\mathbf{c}}_e$ is seen to be a weighted sum of $\mathbf{c}_1,\ldots,\mathbf{c}_f$:

$$\hat{\mathbf{c}}_e = \mathbf{c}_1 \cos ine(\alpha_{1e})\cos ine(\beta_{1e}) + \mathbf{c}_2 \cos ine(\alpha_{2e})\cos ine(\beta_{2e}) + \ldots + \mathbf{c}_f \cos ine(\alpha_{fe})\cos ine(\beta_{fe}). \quad (3.11)$$

Now, the relation in (3.11) is true for $e = 1,\ldots,d$, thus one has

$$\begin{aligned}
\hat{\mathbf{c}}_1 &= \mathbf{c}_1 \cos ine(\alpha_{11})\cos ine(\beta_{11}) + \mathbf{c}_2 \cos ine(\alpha_{21})\cos ine(\beta_{21}) + \ldots + \mathbf{c}_f \cos ine(\alpha_{f1})\cos ine(\beta_{f1}) \\
\hat{\mathbf{c}}_2 &= \mathbf{c}_1 \cos ine(\alpha_{12})\cos ine(\beta_{12}) + \mathbf{c}_2 \cos ine(\alpha_{22})\cos ine(\beta_{22}) + \ldots + \mathbf{c}_f \cos ine(\alpha_{f2})\cos ine(\beta_{f2}) \\
&\vdots \\
\hat{\mathbf{c}}_d &= \mathbf{c}_1 \cos ine(\alpha_{1d})\cos ine(\beta_{1d}) + \mathbf{c}_2 \cos ine(\alpha_{2d})\cos ine(\beta_{2d}) + \ldots + \mathbf{c}_f \cos ine(\alpha_{fd})\cos ine(\beta_{fd}).
\end{aligned}$$

Clearly the upper bound for $\sum_{k=1}^{d} \|\hat{\mathbf{c}}_k\|^2$ is $\sum_{k=1}^{d} \|\mathbf{c}_k\|^2$, and this is achieved when $\cos ine(\alpha_{11}) = \cos ine(\beta_{11}) = 1$, $\cos ine(\alpha_{22}) = \cos ine(\beta_{22}) = 1$, …, and $\cos ine(\alpha_{ee}) = \cos ine(\beta_{ee}) = 1$, or when $\hat{\mathbf{a}}_1 = \mathbf{a}_1$, $\hat{\mathbf{b}}_1 = \mathbf{b}_1$, $\hat{\mathbf{a}}_2 = \mathbf{a}_2$, $\hat{\mathbf{b}}_2 = \mathbf{b}_2$,..., $\hat{\mathbf{a}}_d = \mathbf{a}_d$, and $\hat{\mathbf{b}}_d = \mathbf{b}_d$. Note this solution satisfies the normal equations, as $\sum_{i} a_{gi}a_{ei} = 0$ and $\sum_{j} b_{gj}b_{ej} = 0$ for $e \neq g$. Now by **Proposition 3.2**, the PARAFAC (orth.) solution is **A** and **B** such that the sums of squares of $\sum_{k=1}^{d} \|\hat{\mathbf{c}}_k\|^2$ is maximized. Since the solution $\hat{\mathbf{a}}_1 = \mathbf{a}_1$, $\hat{\mathbf{b}}_1 = \mathbf{b}_1$, $\hat{\mathbf{a}}_2 = \mathbf{a}_2$, $\hat{\mathbf{b}}_2 = \mathbf{b}_2$,..., $\hat{\mathbf{a}}_d = \mathbf{a}_d$, $\hat{\mathbf{b}}_d = \mathbf{b}_d$, and $\hat{\mathbf{c}}_1 = \mathbf{c}_1$, $\hat{\mathbf{c}}_2 = \mathbf{c}_2$,..., $\hat{\mathbf{c}}_d = \mathbf{c}_d$, maximizes this sums of squares and satisfies the least squares normal equations, it is the rank-*d* solution.

CHAPTER FOUR

COMMON PRINCIPAL COMPONENTS

The common principal components (CPC) model hypothesizes that the same principal components exist in multiple datasets, although the associated eigenvalues may vary. It shares with the methods developed in later chapters the concept of the common component. Flury (1988) developed the maximum likelihood approach to CPC. In this chapter I show how CPC can be approached by least squares methods. While an exposition on CPC is not strictly necessary to develop the concepts of CVA, CC and RA over time, what I do in this chapter is closely related to what I do in later chapters. The CPC model introduces in a clear way the idea of a common variate. The use of three-mode principal components for CPC presages its use for generalizing CVA, CC, RA and PR. Also of interest is the relationship between maximum likelihood and least squares methodologies.

Section **4.1** presents background material, defining common principal components and two related models, partial common principal components and common space analysis. Section **4.2** shows how to achieve the common principal components model using three-mode principal components. In Section **4.3** it is shown how to approach the partial common principal components and common space analysis with least squares. Section **4.4** has a comparison of the

maximum likelihood and least squares approaches to CPC. Lastly, in Section **4.5** an alternative formulation of common principal components is proposed.

4.1 COMMON PRINCIPAL COMPONENTS

The common principal components (Flury 1988) model hypothesizes that multiple datasets share common components, though each dataset has different eigenvalues associated with those components. The CPC hypothesis for k $p \times p$ covariance matrices, $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_k$, is:

$$\boldsymbol{\Sigma}_i = \mathbf{B}\boldsymbol{\Lambda}_i\mathbf{B}', \qquad i = 1, \ldots, k,$$

where $\mathbf{B}$ is an orthogonal $p \times p$ matrix, and $\boldsymbol{\Lambda}_i = \text{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$. Note that a component may have a large eigenvalue associated with one dataset, but a small eigenvalue associated with another dataset. Hence there is no canonical ordering of the components by ordering them according to the size of their eigenvalues as in principal components analysis.

The common principal components model is equivalent to postulating that the covariance matrices for the datasets are simultaneously diagonalizable by the same orthogonal matrices, i.e., the matrix of common components. The elements of the resulting diagonal matrices contain the respective eigenvalues. Thus:

$$\mathbf{B}'\boldsymbol{\Sigma}_i\mathbf{B} = \boldsymbol{\Lambda}_i$$

$i = 1, \ldots, k$, where $\mathbf{B}$ and $\boldsymbol{\Lambda}_i$ are defined as above. Note that a necessary and sufficient condition for the existence of $\mathbf{B}$ is that $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_k$ are commutable, that is, $\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_j\boldsymbol{\Sigma}_i$ for all i,j.

The sample covariance matrices are modeled as

$$\mathbf{S}_i = \mathbf{B}\boldsymbol{\Lambda}_i\mathbf{B}' + \mathbf{U}_i$$

where $\mathbf{S}_i$ is the i^{th} (unbiased) sample covariance matrix and $\mathbf{U}_i$ is the i^{th} matrix of error terms. I assume that the original measurements follow a multivariate normal distribution and consequently that $(n_i - 1)\mathbf{S}_i$ follows a Wishart distribution. By maximizing the likelihood subject to the constraint of orthogonality on $\mathbf{B}$, estimating equations are derived, the solutions of which include the maximum likelihood solution for $\mathbf{B}$. The F-G algorithm (Flury & Gautschi 1986) solves these equations, though without guaranty of globally optimality. The estimating equations are, for $m, r = 1, \ldots, p, \quad m \neq r$.

$$\boldsymbol{\beta}_m' \left(\sum_{i=1}^{k} (n_i - 1) \left(\frac{\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m - \boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r}{\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r} \right) \mathbf{S}_i \right) \boldsymbol{\beta}_r = 0$$

with $\boldsymbol{\beta}_j'\boldsymbol{\beta}_j = 1$ and $\boldsymbol{\beta}_j'\boldsymbol{\beta}_w = 0$ for $j \neq w$, where $\boldsymbol{\beta}_j$ is the j^{th} column of $\mathbf{B}$. Further, a likelihood ratio statistic is derived to test for the significance of deviations from the model.

Flury extends the CPC model by developing a partial common principal components model. The partial CPC model hypothesizes that there are only q of p eigenvectors common to all $\boldsymbol{\Sigma}_i$. The remaining $p - q$ are specific to each dataset. That is

$$\mathbf{B}_i'\boldsymbol{\Sigma}_i\mathbf{B}_i = \boldsymbol{\Lambda}_i,$$

where $\mathbf{B}_i$ are orthogonal matrices such that $\mathbf{B}_i = [\mathbf{B}_1 : \mathbf{B}_{2i}]$, $\mathbf{B}_1$ is a $p \times q$ orthonormal matrix of q common eigenvectors, and $\mathbf{B}_{2i}$ are $p \times (p-q)$ matrices with $p-q$ eigenvectors specific to the i^{th} dataset.

Flury indicates that the maximum likelihood equations solving this model are extremely laborious to implement. He recommends instead an approximate solution using the CPC estimates. The approximation is based on the observation that if the partial CPC model holds exactly, then the q common components are estimated correctly in the CPC model, regardless of the specific components. The method involves first obtaining approximate maximum likelihood estimates of the common components, $\mathbf{B}_1$, from the CPC estimates. Then the $\mathbf{B}_{2i}$ are obtained by finding $\mathbf{B}_{2i}$ that diagonalize $\mathbf{S}_i$ subject to $\mathbf{B}_{2i}$ being orthogonal to $\mathbf{B}_1$.

Related to the partial CPC model is common space analysis, which hypothesizes that q eigenvectors of each covariance matrix span the same subspace. Analogous to the partial CPC model, Flury describes the maximum likelihood equations as extremely laborious to implement and recommends instead an approximation using the solution to the CPC model. The approximation is based on the observation that if q eigenvectors span the same subspace, then the CPC solution will contain q columns which span that subspace.

To illustrate the method of CPC I present **Example 4.1**, which is taken from Flury (1984). A CPC analysis is performed on Fisher's (1936) well known iris data. The four variables are sepal length, sepal width, petal length and petal width. They are measured on three species of iris: versicolor, virginica and setosa. The sample sizes are 50 for each species. (a) shows the sample covariance matrices, with the variables ordered as listed above. (b) shows the coefficients of the common principal components. The columns list the components, the rows are the weights for the variables. (c) shows the estimates for the eigenvalues associated with each common component in each dataset. In this example null hypothesis of common principal components is rejected, the chi-square test statistic being 63.9 with 12 degrees of freedom. Flury (1984) indicates that the common principal components have no obvious interpretation.

Example 4.1:

(a) Sample Covariance Matrices

Versicolor

$$\mathbf{S}_1 = \begin{bmatrix} 26.6433 & 8.5184 & 18.2898 & 5.5780 \\ 8.5184 & 9.8469 & 8.2653 & 4.1204 \\ 18.2898 & 8.2653 & 22.0816 & 7.3102 \\ 5.5780 & 4.1204 & 7.3102 & 3.9106 \end{bmatrix}$$

Virginica

$$\mathbf{S}_2 = \begin{bmatrix} 40.4343 & 9.3763 & 30.3290 & 4.9094 \\ 9.3763 & 10.4004 & 7.1380 & 4.7629 \\ 30.3290 & 7.1380 & 30.4588 & 4.8824 \\ 4.9094 & 4.7629 & 4.8824 & 7.5433 \end{bmatrix}$$

Setosa

$$\mathbf{S}_3 = \begin{bmatrix} 12.4249 & 9.9216 & 1.6355 & 1.0331 \\ 9.9216 & 14.3690 & 1.1698 & 0.9298 \\ 1.6355 & 1.1698 & 3.0159 & 0.6069 \\ 1.0331 & 0.9298 & 0.6069 & 1.1106 \end{bmatrix}$$

(b) Coefficients of Common Principal Components

$$\mathbf{B} = \begin{bmatrix} 0.7367 & -0.6471 & -0.1640 & 0.1084 \\ 0.2468 & 0.4655 & -0.8346 & -0.1607 \\ 0.6047 & 0.5002 & 0.5221 & -0.3338 \\ 0.1753 & 0.3382 & 0.0628 & 0.9225 \end{bmatrix}$$

(c) Estimated Eigenvalues Associated with the Common Principal Components

Versicolor	48.46	7.47	5.54	1.01
Virginica	69.22	6.71	7.54	5.36
Setosa	14.64	2.75	12.51	1.02

4.2 THE LEAST SQUARES APPROACH TO COMMON PRINCIPAL COMPONENTS

An alternative approach to estimating common principal components is possible through decompositions of covariance matrices. This approach uses three-mode principal components analysis and is based on a least squares solution.

I refer back to the PARAFAC model with orthogonality constraints (orth.) of Section **2.3.2**,

$$\mathbf{X}_i = \mathbf{G}\mathbf{C}_i\mathbf{H}, \qquad i = 1,\ldots,k.$$

Let $\mathbf{X}_i$ be k positive definite matrices, $\mathbf{S}_i$. Then modeling $\mathbf{S}_i$ by the PARAFAC (orth.) model is equivalent to a least squares form of the CPC model

$$\mathbf{S}_i = \mathbf{B}\mathbf{\Lambda}_i\mathbf{B}' + \mathbf{E}_i, \tag{4.1}$$

where I restrict $\mathbf{B}$ to be $p \times p$ orthogonal matrices and note that $\mathbf{B} = \mathbf{D}$. The notation is changed to indicate diagonal $\mathbf{C}_i$ as $\mathbf{\Lambda}_i$, and $\mathbf{E}_i$ is defined to be the i^{th} matrix of lack of fit terms.

There is a second procedure equivalent to least squares CPC which arises in the context of analyzing three-mode principal components. In order to diagonalize the core matrices, $\mathbf{C}_i$, of the Tucker2 model, Kroonenberg and DeLeeuw (Kroonenberg 1983) present a least squares method for finding orthogonal transformation matrices that diagonalize multiple square matrices. Given multiple $p \times p$ matrices, $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_k$, their method finds orthogonal $p \times p$ matrices $\mathbf{G}$ and $\mathbf{H}$ such that one minimizes

$$\sum_{i=1}^{k} \text{trace}\left(\mathbf{G}'\mathbf{Q}_i\mathbf{H} - \text{Diag}(\mathbf{G}'\mathbf{Q}_i\mathbf{H})\right)' \left(\mathbf{G}'\mathbf{Q}_i\mathbf{H} - \text{Diag}(\mathbf{G}'\mathbf{Q}_i\mathbf{H})\right),$$

where Diag(**J**) is defined as the diagonal matrix whose diagonal elements are the diagonal elements of **J**.

This procedure is equivalent to least squares CPC when it is applied to multiple positive definite matrices. To see this, notice that the least squares solution to (4.1) is also the least squares solution to (4.2) below

$$\mathbf{B}'\mathbf{S}_i\mathbf{B} = \mathbf{\Lambda}_i + \mathbf{E}_i^*, \tag{4.2}$$

since $\mathbf{B}$ which minimizes $\sum_{i=1}^{k} \text{trace}\left(\mathbf{E}_i'\mathbf{E}_i\right)$ also minimizes $\sum_{i=1}^{k} \text{trace}\left(\mathbf{E}_i'^{*}\mathbf{E}_i^{*}\right)$, where $\mathbf{E}_i^{*} = \mathbf{B}'\mathbf{E}_i\mathbf{B}$ as $\text{trace}(\mathbf{E}_i'\mathbf{E}_i) = \text{trace}(\mathbf{E}_i'^{*}\mathbf{E}_i^{*})$. Kroonenberg and DeLeeuw's algorithm is equivalent to Harshman and Lundy's (1994) when $\mathbf{B}$ is restricted to be orthogonal.

I have shown above that CPC can be modeled as a special case of three-mode models based on a least squares solution. Next I derive estimating equations for the least squares estimates which are analogous for those of the maximum likelihood estimates given in Section **4.1**. The comparison of these equations shall bring into focus the similarities and differences of the two modes of estimation. A preliminary is necessary. Up to this point I have only discussed modeling the $\mathbf{S}_i$ matrices. However, if the sample sizes are unequal it is reasonable that covariance matrices calculated from larger samples should be given more weight in the estimation. This is accomplished by modeling $(n_i - 1)\mathbf{S}_i$ instead of $\mathbf{S}_i$. I shall choose to model these crossproduct matrices instead of the unweighted covariance matrices as this reveals how different sample sizes in the *k* groups affect the least squares estimates.

As pointed out earlier (4.2), finding the solution to least squares CPC is equivalent to finding a rotation matrix $\mathbf{B}$ that minimizes the sums of squares lack of fit to the model of simultaneous diagonalizability. I denote this sum of squares lack of fit by $f(\mathbf{B})$. Then

$$f(\mathbf{B}) = \sum_{i=1}^{k} (n_i - 1)^2 \,\text{trace}\left((\mathbf{B}'\mathbf{S}_i\mathbf{B} - \Lambda_i)'(\mathbf{B}'\mathbf{S}_i\mathbf{B} - \Lambda_i)\right). \tag{4.3}$$

It is apparent from (4.3) that the least squares solution for Λ_i is $\Lambda_i = \text{Diag}(\mathbf{B}'\mathbf{S}_i\mathbf{B})$. This result is also true for the maximum likelihood estimation of CPC (Flury 1984). Thus

$$f(\mathbf{B}) = \sum_{i=1}^{k} (n_i - 1)^2 \,\text{trace}\left((\mathbf{B}'\mathbf{S}_i\mathbf{B} - \text{Diag}(\mathbf{B}'\mathbf{S}_i\mathbf{B}))'(\mathbf{B}'\mathbf{S}_i\mathbf{B} - \text{Diag}(\mathbf{B}'\mathbf{S}_i\mathbf{B}))\right).$$

Expanding yields

$$f(\mathbf{B}) = \sum_{i=1}^{k} (n_i - 1)^2 \,\text{trace}(\mathbf{B}'\mathbf{S}_i\mathbf{B})'(\mathbf{B}'\mathbf{S}_i\mathbf{B}) + (n_i - 1)^2 \,\text{trace}(\text{Diag}(\mathbf{B}'\mathbf{S}_i\mathbf{B}))^2$$
$$-2(n_i - 1)^2 \,\text{trace}\left(\text{Diag}(\mathbf{B}'\mathbf{S}_i\mathbf{B})(\mathbf{B}'\mathbf{S}_i\mathbf{B})\right).$$

Since the first term in the sum is constant, minimizing the above reduces to maximizing

$$g(\mathbf{B}) = \sum_{i=1}^{k} (n_i - 1)^2 \,\text{trace}(\text{Diag}(\mathbf{B}'\mathbf{S}_i\mathbf{B}))^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} (n_i - 1)^2 (\beta_j'\mathbf{S}_i\beta_j)^2. \tag{4.4}$$

From this point the problem is equivalent to maximizing

$$G(\mathbf{B}) = \sum_{i=1}^{k}\sum_{j=1}^{p} (n_i - 1)^2 (\beta_j'\mathbf{S}_i\beta_j)^2 - 2\sum_{j=2}^{p}\sum_{h=1}^{j-1} \ell_{hj}\beta_h'\beta_j - \sum_{h=1}^{p} \ell_h(\beta_h'\beta_h - 1)$$

where ℓ_{hj} $(1 \le h < j \le p)$ and ℓ_h $(1 \le h \le p)$ are $p(p+1)/2$ Lagrange multipliers. The vector of partial derivatives of $G(\mathbf{B})$ with respect to β_r, set equal to zero, yields

$$\frac{\delta}{\delta\beta_r} G(\mathbf{B}) = 2\sum_{i=1}^{k} (n_i - 1)^2 (\beta_r'\mathbf{S}_i\beta_r)\mathbf{S}_i\beta_r - 2\sum_{\substack{h=1 \\ h\neq r}}^{p} \ell_{rh}\beta_h - 2\ell_r\beta_r = \mathbf{0} \tag{4.5}$$

where I put $\ell_{rh} = \ell_{hr}$ if r > h. Multiplying (4.5) from the left by $(\frac{1}{2})\boldsymbol{\beta}_r'$ gives

$$\sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)^2 - \sum_{\substack{h=1\\h\neq r}}^{p}\boldsymbol{\beta}_r'\boldsymbol{\beta}_h\ell_{rh} - \boldsymbol{\beta}_r'\boldsymbol{\beta}_r\ell_r = 0$$

implying $\ell_r = \sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)^2$. Substituting for ℓ_r back into (4.5) one has

$$\sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)\mathbf{S}_i\boldsymbol{\beta}_r - \sum_{\substack{h=1\\h\neq r}}^{p}\ell_{rh}\boldsymbol{\beta}_h - \sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)^2\boldsymbol{\beta}_r = \mathbf{0}.$$

Multiplying the above from the left by $\boldsymbol{\beta}_m'$ $(m \neq r)$ implies

$$\sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r - \sum_{\substack{h=1\\h\neq r}}^{p}\ell_{rh}\boldsymbol{\beta}_m'\boldsymbol{\beta}_h - \sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)^2\boldsymbol{\beta}_m'\boldsymbol{\beta}_r = 0.$$

Thus for $m \neq r$

$$\ell_{rm} = \sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r).$$

Interchanging the indices r and m and noting that $\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_m = \boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r$ and $\ell_{rm} = \ell_{mr}$, it follows that

$$\ell_{rm} = \sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r).$$

Hence

$$\sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r) = \sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r),$$

which implies

$$\boldsymbol{\beta}_m'\left(\sum_{i=1}^{k}(n_i-1)^2(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m - \boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)\mathbf{S}_i\right)\boldsymbol{\beta}_r = 0 \qquad (4.6)$$

for $m, r = 1, \ldots, p \quad m \neq r$.

Equations (4.6) are the estimating equations for the least squares solution to CPC. With the exception of a different term involving sample sizes and the lack of the denominator term, they are the same as the estimating equations for the maximum likelihood estimates. As mentioned earlier in this section, the least squares estimates can be obtained by an alternating least squares algorithm (Kroonenberg 1983, Harshman & Lundy 1994). However, as an alternative, Flury's and Gautschi's F-G algorithm is easily adapted to solve equations (4.6). SAS programs for both the alternating least squares algorithm and the F-G algorithm are found in Appendix Two. **Example 4.2** illustrates the application of least squares common principal components to Fisher's iris data. Note how close these estimates are to the maximum likelihood estimates presented in **Example 4.1**.

Example 4.2

Estimated Coefficients

$$\mathbf{B} = \begin{bmatrix} 0.7274 & -0.6145 & -0.1998 & 0.2310 \\ 0.2385 & 0.4519 & -0.8199 & -0.2581 \\ 0.6245 & 0.4215 & 0.5346 & -0.3828 \\ 0.1548 & 0.4904 & 0.0457 & 0.8564 \end{bmatrix}$$

Estimated Eigenvalues

Versicolor	48.37	7.36	5.59	1.16
Virginica	69.38	7.59	7.45	4.41
Setosa	14.29	2.56	12.83	1.24

The comparison of the estimates in examples **4.1** and **4.2** leads to the question of when the least squares and maximum likelihood estimates will be similar and when they will differ. To answer this one must compare closely the least squares estimating equations with the maximum likelihood estimating equations. With equal sample sizes, the $(n_i - 1)^2$ and $(n_i - 1)$ terms cancel out of both sets of equations. With unequal sample sizes, both least squares and maximum likelihood estimating equations put greater weight on the samples with larger sizes. However, the least squares equations weight the larger samples more heavily than the maximum likelihood equations do.

The difference in the denominators of the estimating equations also has implications. Flury views the k terms $(n_i - 1)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m - \boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r) / (\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)$ as weights for $\mathbf{S}_i$ in the $\mathrm{m,r}^{\text{th}}$ estimating equation. The closer $\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m$ and $\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r$ are to each other, the smaller the weight on $\mathbf{S}_i$ is. When $\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m = \boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r$ there is sphericity in the plane spanned by $\boldsymbol{\beta}_m$ and $\boldsymbol{\beta}_r$ for the i^{th} dataset and the influence of $\mathbf{S}_i$ in the $\mathrm{m,r}^{\text{th}}$ equation vanishes. The same property is apparent for the least squares equations. However, unlike the least squares estimating equations, the weights for $\mathbf{S}_i$ in the maximum likelihood equations also include the product $\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r$ in the denominator. Thus when $\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m - \boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r$ is small in absolute magnitude, but large in comparison to $(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m)(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)$, maximum likelihood estimation gives more weight to that $\mathbf{S}_i$ in the $\mathrm{m,r}^{\text{th}}$ estimating equations. Except for this circumstance the two estimators yield similar transformations given equal sample sizes.

The following example shows two matrices for which the estimated transformations differ substantially despite equal sample sizes because of the condition described in the previous paragraph. Both $\mathbf{S}_1$ and $\mathbf{S}_2$ are the product of diagonal matrices pre-multiplied and post-multiplied by an orthogonal matrix and its transpose.

Example 4.3:

$$\mathbf{S}_1 = \begin{bmatrix} 1000 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.01 \end{bmatrix}$$

$$\mathbf{S}_2 = \begin{bmatrix} 823.20 & 169.28 & -273.73 \\ 169.28 & 467.71 & -109.09 \\ -273.73 & -109.09 & 209.09 \end{bmatrix} = \begin{bmatrix} 0.8704 & -0.3714 & 0.3233 \\ 0.3482 & 0.9285 & 0.1293 \\ -0.3482 & 0.0000 & 0.9374 \end{bmatrix} \begin{bmatrix} 1000 & 0 & 0 \\ 0 & 400 & 0 \\ 0 & 0 & 100 \end{bmatrix} \begin{bmatrix} 0.8704 & 0.3482 & -0.3482 \\ -0.3714 & 0.9285 & 0.0000 \\ 0.3233 & 0.1293 & 0.9374 \end{bmatrix}$$

Assuming $n_1 = n_2$, the transformations estimated by maximum likelihood and least squares are

Maximum Likelihood

$$\mathbf{B} = \begin{bmatrix} 1.0000 & -0.0002 & 0.0000 \\ 0.0002 & 1.0000 & 0.0029 \\ 0.0000 & -0.0029 & 1.0000 \end{bmatrix}$$

Least Squares

$$\mathbf{B} = \begin{bmatrix} 0.9890 & -0.0893 & 0.1180 \\ 0.0607 & 0.9721 & 0.2266 \\ -0.1350 & -0.2169 & 0.9668 \end{bmatrix}$$

Next I modify the above example so that the least squares and the maximum likelihood estimates will differ less substantially. I only change $\mathbf{S}_1$.

$$\mathbf{S}_1 = \begin{bmatrix} 1000 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{bmatrix}$$

Maximum Likelihood

$$\mathbf{B} = \begin{bmatrix} 0.9957 & -0.0186 & 0.0909 \\ 0.0122 & 0.9974 & 0.0703 \\ -0.0919 & -0.0689 & 0.9934 \end{bmatrix}$$

Least Squares

$$\mathbf{B} = \begin{bmatrix} 0.9939 & -0.0612 & 0.0918 \\ 0.0560 & 0.9967 & 0.0587 \\ -0.0950 & -0.0532 & 0.9940 \end{bmatrix}$$

I have made clear under what circumstances the maximum likelihood and least squares solutions are different or similar. The following theorem strengthens the comparison of the two approaches by showing that their solutions are asymptotically equivalent as the sample sizes become large.

Theorem 4.1. Let $\mathbf{S}_i$ be k covariance matrices with sample sizes n_i such that $\mathbf{S}_i = \mathbf{B}\mathbf{\Lambda}_i\mathbf{B}' + \mathbf{E}_i$, where $\mathbf{B}$ and $\mathbf{\Lambda}_i$ are defined as for (4.1), and $\mathbf{E}_i$ is an error matrix whose elements have zero expectation and finite covariances. Then as $n_i \to \infty$ for $i = 1, \ldots, k$, $\mathbf{B}$ solves both the maximum likelihood and the least squares estimating equations.

Proof: Both sets of estimating equations can be written in the form

$$a_{mr1}\boldsymbol{\beta}_m'\mathbf{S}_1\boldsymbol{\beta}_r + \ldots + a_{mrk}\boldsymbol{\beta}_m'\mathbf{S}_k\boldsymbol{\beta}_r = 0, \tag{4.7}$$

$1 \le m < r \le p$. For the least squares estimating equations, $a_{mri} = (n_i - 1)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m - \boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)$, since

$$\boldsymbol{\beta}_m'\left(\sum_{i=1}^{k}(n_i - 1)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m - \boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)\mathbf{S}_i\right)\boldsymbol{\beta}_r = \left(\sum_{i=1}^{k}(n_i - 1)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m - \boldsymbol{\beta}_r\mathbf{S}_i\boldsymbol{\beta}_r)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r)\right).$$

For the maximum likelihood estimation the scalar terms are $a_{mri} = (n_i - 1)\dfrac{\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m - \boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r}{\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_m\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r}$.

From (4.7) it is clear that when $\mathbf{S}_i = \mathbf{\Sigma}_i = \mathbf{B}\mathbf{\Lambda}_i\mathbf{B}'$, $i = 1, \ldots, k$, that $\mathbf{B}$ is a solution for both the

maximum likelihood and least squares estimating equations. Since as $n_i \to \infty$, $\mathbf{S}_i \to \boldsymbol{\Sigma}_i$ (Anderson 1984), **B** asymptotically solves both sets of equations. •

What **Theorem 4.1** says is simply that as the sample size is become larger the $\mathbf{S}_i$ $i = 1,\ldots,k$ approach simultaneous diagonizability, assuming the hypothesis of common principal components is true.

4.3 LEAST SQUARES APPROACHES TO PARTIAL COMMON PRINCIPAL COMPONENTS AND COMMON SPACE ANALYSIS

In this section I show first how the partial common principal components model and then how common space analysis (Flury 1987) can be approached with least squares methods.

An exact least squares solution to partial CPC is not attempted due to its complexity. However, an approximate solution is readily available by using the least squares estimate of the full CPC model. Analogous to Flury's approximation for estimating partial CPC, this approximation is based on the observation that if there are q eigenvectors common to each dataset, then the full least squares CPC correctly estimates these common eigenvectors. This observation is formally stated in the following theorem:

Theorem 4.2. Assume that the $p \times p$ positive definite matrices $\mathbf{S}_i$ have $q < p$ common eigenvectors. Denote these by $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_q$, and let them comprise the columns of $\mathbf{B}_1$. Hence $\mathbf{S}_i = \mathbf{B}_1\boldsymbol{\Lambda}_{1i}\mathbf{B}_1' + \mathbf{B}_{2i}\boldsymbol{\Lambda}_{2i}\mathbf{B}_{2i}'$, where $\boldsymbol{\Lambda}_{1i}$ is a $q \times q$ diagonal matrix and $\boldsymbol{\Lambda}_{2i}$ is a $(p-q)\times(p-q)$ diagonal matrix. Then the $p \times p$ orthogonal matrix $\hat{\mathbf{B}}$ that maximizes the function $g(\hat{\mathbf{B}})$ (4.4) has $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_q$ among its columns, or can be chosen to if $\hat{\mathbf{B}}$ is not uniquely defined.

Proof: The main part of the proof is to show that $\hat{\mathbf{B}} = [\mathbf{B}_1 : \mathbf{B}_{21}]\mathbf{A}$, where **A** is orthogonal and of the form $\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{bmatrix}$, with $\mathbf{A}_1$ $q \times q$ and $\mathbf{A}_2$ $(p-q)\times(p-q)$. To achieve this I will determine $\hat{\mathbf{B}}$ one column vector at a time. The $\hat{\boldsymbol{\beta}}_j$ can be estimated successively because by **Theorem 4.1** the least squares solutions to PARAFAC with orthogonality constraints are nested. That is, if the $\hat{\boldsymbol{\beta}}_j$ are ordered by the sums of squares fitted, then the u^{th} vector of any optimal *m*-vector solution equals the u^{th} vector of any *n*-vector solution, $u \le m,n \le p$. The first column vector that I will determine is the one that yields the largest value of $g(\hat{\boldsymbol{\beta}}_j)$. I denote this vector by $\hat{\boldsymbol{\beta}}_d$. Define $\mathbf{a}_d = [\mathbf{B}_1 : \mathbf{B}_{21}]'\hat{\boldsymbol{\beta}}_d$, $h(\mathbf{a}_d) = g(\hat{\boldsymbol{\beta}}_d)$ and $\mathbf{J}_i = \mathbf{B}_{21}'\mathbf{B}_{2i}\boldsymbol{\Lambda}_{2i}\mathbf{B}_{2i}'\mathbf{B}_{21}$. Then

$$g(\hat{\boldsymbol{\beta}}_d) = h(\mathbf{a}_d) = \sum_{i=1}^{k} (n_i - 1)^2 \left(\mathbf{a}_d' \begin{bmatrix} \boldsymbol{\Lambda}_{1i} & 0 \\ 0 & \mathbf{J}_i \end{bmatrix} \mathbf{a}_d \right)^2 .$$

Since $\|\mathbf{a}_d\|^2 = 1$ I can partition $\mathbf{a}_d$ as $\mathbf{a}_d = \begin{pmatrix} c\mathbf{a}_{1d} \\ \cdots \\ f\mathbf{a}_{2d} \end{pmatrix}$, where $\mathbf{a}_{1d}$ is a $q \times 1$ vector, $\mathbf{a}_{2d}$ is a $(p-q)\times 1$ vector, c and f are scalars, and $\|\mathbf{a}_{1d}\|^2 = \|\mathbf{a}_{1d}\|^2 = c^2 + f^2 = 1$. Let $t_{id} = \mathbf{a}'_{1d}\mathbf{\Lambda}_{1i}\mathbf{a}_{1d}$ and $u_{id} = \mathbf{a}'_{2d}\mathbf{J}_i\mathbf{a}_{2d}$. Then

$$h(\mathbf{a}_d) = (n_i - 1)^2\left(c^4 \sum_{i=1}^{k} t_{id}^2 + c^2 f^2 \sum_{i=1}^{k} t_{id} u_{id} + f^4 \sum_{i=1}^{k} u_{id}^2 \right).$$

Define $\hat{t}$ as the maximum attainable value of $\sum_{i=1}^{k} t_{id}^2$ and $\hat{\mathbf{a}}_{1d}$ as the vector that attains it. Likewise define $\hat{u}$ as the maximum value attainable for $\sum_{i=1}^{k} u_{id}^2$ and $\hat{\mathbf{a}}_{2d}$ as the vector that attains it. If either $\hat{\mathbf{a}}_{1d}$ or $\hat{\mathbf{a}}_{2d}$ is not uniquely defined one can define $\hat{\mathbf{a}}_{1d}$ or $\hat{\mathbf{a}}_{2d}$ as any vector of that attains $\hat{t}$ or $\hat{u}$. Since $\mathbf{a}_d$ is the vector such that $h(\mathbf{a}_d)$ is at a maximum, if $\hat{t} > \hat{u}$ then $c = 1$ and $\mathbf{a}_d = \begin{pmatrix} \hat{\mathbf{a}}_{1d} \\ \cdots \\ \mathbf{0} \end{pmatrix}$; if $\hat{t} < \hat{u}$ then $f = 1$ and $\mathbf{a}_d = \begin{pmatrix} \mathbf{0} \\ \cdots \\ \hat{\mathbf{a}}_{2d} \end{pmatrix}$; if $\hat{t} = \hat{u}$ then one can arbitrarily choose between $c = 1$ or $f = 1$. Thus I have determined $\mathbf{a}_d$ and $\hat{\boldsymbol{\beta}}_d$.

Further vectors, $\hat{\boldsymbol{\beta}}_{d'}$, $d' \neq d$, are determined in a manner analogous to how $\hat{\boldsymbol{\beta}}_d$ was determined, subject to the constraint of orthogonality to the previously derived vectors. Because there exist $\hat{\mathbf{a}}_{1d'}$ and $\hat{\mathbf{a}}_{2d'}$ that are orthogonal to previously derived $\hat{\mathbf{a}}_{1d}$ and $\hat{\mathbf{a}}_{2d}$, there also exist $\mathbf{a}_{d'}$ and hence $\hat{\boldsymbol{\beta}}_{d'}$ that satisfy the orthogonality constraints. Successively finding the remaining $p-1$ $\hat{\boldsymbol{\beta}}_{d'}$ to determine $\hat{\mathbf{B}}$ yields further $\mathbf{a}_{d'}$ of the form $\mathbf{a}_{d'} = \begin{pmatrix} c\mathbf{a}_{1d'} \\ f\mathbf{a}_{2d'} \end{pmatrix}$ with $c = 1$ or $f = 1$. Let $\mathbf{A} = [\mathbf{a}_j]$, putting the columns corresponding to $c = 1$ first. Then $\mathbf{A}$ is of the form $\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{bmatrix}$.

The conclusion of the proof is to note that the $\mathbf{\Lambda}_{1i}$ are diagonal. Hence $\mathbf{a}_1$ through $\mathbf{a}_q$ can be chosen to be the first *q* unit vectors and $\hat{\mathbf{B}}$ has $\boldsymbol{\beta}_1$ through $\boldsymbol{\beta}_q$ as columns. Ž

Related to the partial CPC model is common space analysis, which hypothesizes that *q* eigenvectors of each covariance matrix span the same subspace. As with partial CPC, an exact least squares solution is not attempted due to its complexity. However, an approximate solution is likewise available by using the least squares estimate of the full CPC model. Analogous to Flury's approximation for estimating common space analysis, this approximation is based on the observation that if there are *q* eigenvectors spanning the same subspace in all the datasets, then

the least squares estimate of the full CPC solution will contain q columns which span that subspace. This observation is stated in **Theorem 4.3**.

Theorem 4.3. Assume that the positive definite symmetric matrices $\mathbf{S}_i$ of dimension $p \times p$ have $q < p$ eigenvectors each that span the same q-dimensional subspace as $\mathbf{B}_1$. Then the $p \times p$ orthogonal matrix $\hat{\mathbf{B}}$ that maximizes $g(\hat{\mathbf{B}})$ from equation (4.4) has q columns which span $\mathbf{B}_1$.

Proof: Refer to the main part of the proof of **Theorem 4.2**, substituting $\mathbf{D}_i$ for Λ_{1i}, where $\mathbf{D}_i$ is positive definite. •

4.4 COMPARING THE LEAST SQUARES AND MAXIMUM LIKELIHOOD APPROACHES

The results of the previous sections suggest a straightforward exploratory approach to modeling CPC, partial CPC and common space analysis. One performs a least squares CPC and examines the $\mathbf{B}'\mathbf{S}_i\mathbf{B}$, the covariance matrices of the estimated common principal components $\mathbf{B}$ for each dataset. To determine if the full CPC model is appropriate, one examines the off-diagonal elements of the $\mathbf{B}'\mathbf{S}_i\mathbf{B}$. If they are small in comparison to the diagonal elements then the CPC model is appropriate. If off-diagonal elements are small compared to diagonal elements only for a subset of components, then the partial CPC model is indicated, with that subset of components as the common components. The common space model is appropriate if the ordering of the components can be arranged so that the $\mathbf{B}'\mathbf{S}_i\mathbf{B}$ matrices take the form of two block diagonal matrices, where the elements off the diagonal blocks are small compared to those on the block diagonals. An attractive feature of this exploratory approach is that the squares of the off-diagonal elements (or off-block diagonal elements) of the $\mathbf{B}'\mathbf{S}_i\mathbf{B}$, $(\beta'_m\mathbf{S}_i\beta_r)^2$, represent the model lack of fit of the m,r^{th} components for the i^{th} dataset.

Ultimately, the choice whether to use maximum likelihood or least squares estimation is not obvious and perhaps not necessary as they yield similar results given equal sample sizes. Maximum likelihood estimation has the advantage of allowing the user to perform hypothesis tests. However, the value of tests in this situation may be questionable as the datasets one would analyze are typically large enough so that small deviations from the model would reject the hypothesis of common principal components. Further, CPC is exploratory in spirit and strict tests of preformulated research hypothesis may not be appropriate in such a context. The least squares approach has the advantage that no distributional assumptions are made, and that the model lack of fit is readily related to deviations from diagonalizability. Hence the least squares approach may have the advantage as an exploratory technique.

In conclusion, three-mode principal components presents a rich class of models of which common principal components is a special case. There exist other three-mode models related to CPC which may be of interest. For example, least squares estimation can be extended to include different weightings for the $\mathbf{S}_i$. Or one can perform common principal components analysis on

multiple covariance matrices derived from a data set measuring the same subjects on multiple occasions. Another possibility for data over time is to analyze the subject by measurements data and model common subject components in addition to modeling common principal components.

4.5 COMMON COMPONENTS WHICH MAXIMIZE VARIANCE

The previous sections of this chapter compare the maximum likelihood and the least squares approaches to CPC. In those sections the CPC model was presented as a common variate model that extended a principal components type analysis to multiple datasets. However, there are other possible common variate models that also achieve this. This section discusses one such alternative that turns out to be equivalent to an approximation to CPC given by Krzanowski (1984). It is also of interest because it shows that when one generalizes PCA to multiple datasets one is required to define the model of interest more carefully. It will be seen that several models for multiple datasets which reduce to standard PCA with just one dataset differ subtly in meaning when applied to multiple datasets.

In particular, CPC makes certain hypotheses about the nature of the variates. CPC, whether estimated by least squares or maximum likelihood, hypothesizes that the components should be orthogonal and that they should diagonalize the covariance matrices. The latter is equivalent to the components being uncorrelated. This model derives its justification from the definition of principal components that they be orthogonal in their weights and uncorrelated. However, this is only one of several criteria that characterize principal components. Another is that the components be orthogonal in their weights while maximizing the variance accounted for (Krzanowski 1988). I present in this section a method which finds common orthogonal components which maximize the total variance over the datasets. I shall refer to these estimates as the maximum variance estimates. I show that they are derived simply by performing a singular value decomposition on the sum of the covariance matrices, or of the crossproducts matrices if one wants to weight by sample size. The derivation of the latter result follows.

The objective is to choose orthogonal $\mathbf{B}$ to maximize $w(\mathbf{B})$, where $w(\mathbf{B}) = \sum_{i=1}^{k}\sum_{j=1}^{p}(n_i - 1)\boldsymbol{\beta}_j'\mathbf{S}_i\boldsymbol{\beta}_j$. This is equivalent to maximizing

$$W(\mathbf{B}) = \sum_{i=1}^{k}\sum_{j=1}^{p}(n_i - 1)\boldsymbol{\beta}_j'\mathbf{S}_i\boldsymbol{\beta}_j - 2\sum_{j=2}^{p}\sum_{h=1}^{j-1}\ell_{hj}\boldsymbol{\beta}_h'\boldsymbol{\beta}_j - \sum_{h=1}^{p}\ell_h(\boldsymbol{\beta}_h'\boldsymbol{\beta}_h - 1)$$

where the ℓ_{hj} $(1 \le h < j \le p)$ and ℓ_h $(1 \le h \le p)$ are $p(p+1)/2$ Lagrange multipliers. The vector of partial derivatives of $W(\mathbf{B})$ with respect to $\boldsymbol{\beta}_r$, set equal to zero, yields

$$\frac{\delta}{\delta\boldsymbol{\beta}_r}W(\mathbf{B}) = 2\sum_{i=1}^{k}(n_i - 1)\mathbf{S}_i\boldsymbol{\beta}_r - 2\sum_{\substack{h=1\\h\neq r}}^{p}\ell_{rh}\boldsymbol{\beta}_h - 2\ell_r\boldsymbol{\beta}_r = \mathbf{0} \qquad (4.8)$$

where I put $\ell_{rh} = \ell_{hr}$ if $r > h$. Multiplying the above from the left by $(\frac{1}{2})\boldsymbol{\beta}_r'$ gives

$$\sum_{i=1}^{k}(n_i-1)\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r-\sum_{\substack{h=1\\h\neq r}}^{p}\boldsymbol{\beta}_r'\boldsymbol{\beta}_h\ell_{rh}-\boldsymbol{\beta}_r'\boldsymbol{\beta}_r\ell_r=0$$

implying $\ell_r=\sum_{i=1}^{k}(n_i-1)\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r$. Substituting for ℓ_r into (4.8) and factoring out a two I have

$$\sum_{i=1}^{k}(n_i-1)\mathbf{S}_i\boldsymbol{\beta}_r-\sum_{\substack{h=1\\h\neq r}}^{p}\ell_{rh}\boldsymbol{\beta}_h-\sum_{i=1}^{k}(n_i-1)(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)\boldsymbol{\beta}_r=\mathbf{0}.$$

Multiplying the above from the left by $\boldsymbol{\beta}_m'$ $(m\neq r)$ implies

$$\sum_{i=1}^{k}(n_i-1)\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r-\sum_{\substack{h=1\\h\neq r}}^{p}\ell_{rh}\boldsymbol{\beta}_m'\boldsymbol{\beta}_h-\sum_{i=1}^{k}(n_i-1)(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)\boldsymbol{\beta}_m'\boldsymbol{\beta}_r=0.$$

Thus for $m\neq r$

$$\ell_{rm}=\sum_{i=1}^{k}(n_i-1)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r).$$

Substituting for ℓ_r and ℓ_{rm} into (4.8) and factoring out a two gives

$$\sum_{i=1}^{k}(n_i-1)\mathbf{S}_i\boldsymbol{\beta}_r-\sum_{\substack{h=1\\h\neq r}}^{p}\left(\sum_{i=1}^{k}(n_i-1)(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r)\right)\boldsymbol{\beta}_h-\sum_{i=1}^{k}(n_i-1)(\boldsymbol{\beta}_r'\mathbf{S}_i\boldsymbol{\beta}_r)\boldsymbol{\beta}_r=\mathbf{0}.$$

Differentiating with respect to $\boldsymbol{\beta}_s'$, $s\neq r,m$ yields

$$\boldsymbol{\beta}_m'\left(\sum_{i=1}^{k}(n_i-1)\mathbf{S}_i\right)\boldsymbol{\beta}_r=0. \tag{4.9}$$

Equation (4.9) and the orthogonality constraints on **B** imply that **B** is obtained by the singular value decomposition of $\sum_{i=1}^{k}(n_i-1)\mathbf{S}_i$. If one prefers not to weight by sample size the maximum variance estimate is obtained by a singular value decomposition of $\sum_{i=1}^{k}\mathbf{S}_i$, which is shown using the above argument leaving out the (n_i-1) terms.

These maximum variance common principal components estimates are identical to estimates obtained by an approximation to the maximum likelihood estimates to CPC detailed by Krzanowski (1984). It is easily shown that if the $\mathbf{S}_i$ follow the CPC model exactly, then the maximum variance estimates for **B** equal the maximum likelihood estimates for **B**.

Like the CPC methods, the maximal variance method is illustrated by its application to Fisher's iris data. The coefficients for these components and their corresponding variances bear similarity to those for the least squares and maximum likelihood CPC estimates, however this set of estimates is clearly the most different of the three. This should not be surprising, as these estimates are for parameters of a different model.

Example 4.4:

Estimated Coefficients

$$\mathbf{B} = \begin{bmatrix} 0.7378 & -0.6324 & 0.0561 & 0.2295 \\ 0.3206 & 0.1806 & -0.8732 & -0.3195 \\ 0.5729 & 0.5818 & 0.4588 & -0.3504 \\ 0.1575 & 0.4785 & -0.1543 & 0.8500 \end{bmatrix}$$

Variances

Versicolor	48.41	7.09	5.80	1.19
Virginia	58.45	6.16	10.08	4.15
Setosa	16.20	3.36	9.98	1.37

In conclusion, this section shows that there is more than one model that extends principal components to multiple datasets via a common variate model. Note that the CPC model, estimated by least squares or by maximum likelihood, and the maximal variance model reduce to standard PCA when the data is taken at only one occasion.

CHAPTER FIVE

RELATING TWO SETS OF VARIABLES OVER A THIRD MODE

5.1 INTRODUCTION

In this chapter I present least squares methods for modeling CCA, CVA, RA and PR over a third mode. The term third mode refers to either multiple datasets or multiple occasions. These methods are generalizations of CCA, CVA, RA or PR in that they maintain some of their distinguishing features while reducing to the standard method when the number of occasions is one. The methods developed will be put in the framework of the PARAFAC (orth.) and the Tucker2 models. Like the PARAFAC (orth.) and Tucker2 models these methods are not inferential but exploratory. They can be used well in conjunction with graphical methods, although this topic is deferred until Chapter Six. The methods are flexible. One can model categorical data as well. Hence one has a generalization of correspondence analysis.

Let the variables be divided into two sets, X-variables and Y-variables. The motif running through all the models is that there are two sets of variates which model the linear relationship

between the two sets of variables. These two sets of variates are hypothesized to be common over the third mode, though the strength of the relationships is allowed to change.

Now, these models can be put in the framework of the PARAFAC (orth.) or Tucker2 models. In PARAFAC (orth.) based methods, the variates form pairs, one from each set, and the linear relationship is modeled as occurring strictly between the members of these pairs. The Tucker2 based methods model the linear relationship between all possible pairings of the members of the two sets of variates.

The chapter is organized as follows. Section **5.2** introduces the three-mode models, which I will call CCA/third, CVA/third, RA/third and PR/third. It is shown that they maximize sums of squares regression, sums of squares separation, sums of squared correlations or sums of squared covariances between the two sets of variates. Section **5.3** discusses how to evaluate the fit of a model and how to choose between the PARAFAC (orth.) model and the Tucker2. Section **5.4** presents an example based on the data from the study "Sensitivity of Stream Basins in Shenandoah National Park to Acid Deposition" (Lynch & Dise 1985). Lastly, Section **5.5** discusses several considerations of the methods presented in this chapter. These include autocorrelation, the covariances between observations at different occasions, and the invariance of solutions to different choices of transformation of the X-variables and Y-variables.

5.2 RELATING TWO SETS OF VARIABLES OVER A THIRD MODE

In this section I develop the three-mode models for extending CCA, RA and PR to three-mode data. Consider that CCA, RA and PR are defined in Section **2.2** as singular value decompositions on the transformed matrices of covariances between the X-variables and Y-variables, that is $\mathbf{S}_{XX}^{-½}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-½}$, $\mathbf{S}_{XX}^{-½}\mathbf{S}_{XY}$, or $\mathbf{S}_{XY}$. But three-mode models such as the Tucker2 and the PARAFAC (orth.) are generalizations of the SVD to three-mode data. As such they suggest a framework for modeling CCA, RA and PR for three-mode data by modeling the decomposition of $\mathbf{S}_{XX}^{-½}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-½}$, $\mathbf{S}_{XX}^{-½}\mathbf{S}_{XY}$, or $\mathbf{S}_{XY}$ over a third mode. This is the approach that shall be taken in this chapter.

Define $\mathbf{X}_k$ and $\mathbf{Y}_k$ as the $n_k \times m$ and $n_k \times p$ data matrices of the X-variables and Y-variables over the third mode, $k = 1,\ldots,g$. Then, in the most general sense what is modeled will be $\mathbf{X}_k^{+\prime}\mathbf{Y}_k^{+}$, where $\mathbf{X}_k^{+}$ and $\mathbf{Y}_k^{+}$ are $\mathbf{X}_k$ and $\mathbf{Y}_k$ multiplied by the appropriate transformations; i.e., $\mathbf{S}_{XX}^{-½}$, $\mathbf{S}_{YY}^{-½}$ or $\mathbf{I}$. The relationship between the two sets of variables is harbored in these $\mathbf{X}_k^{+\prime}\mathbf{Y}_k^{+}$ terms. On the other hand, $\mathbf{S}_{XX}$ and $\mathbf{S}_{YY}$ are viewed as suggesting the metric in which to perform the model fitting through transformations of $\mathbf{X}_k$ and $\mathbf{Y}_k$.

It is important to make the proper choice of which transformations to apply to $\mathbf{X}_k'\mathbf{Y}_k$. When relating two sets of variables at just one occasion, one has the choice of using CCA, RA or PR. The choice one makes depends on what the researcher wants to emphasize in his analysis (van de Geer 1984). However, when relating two sets of variables over a third mode, there will usually be just one appropriate three-mode analog method. The choice of transformations defines

the metric in which one wants to maximize fit and minimize lack of fit, and it defines CCA/third, CVA/third, RA/third or PR/third. This topic will be discussed in more detail in Section **5.2.5**.

5.2.1 Redundancy Analysis over a Third Mode

The first three-mode method I will discuss will be redundancy analysis over a third mode, or RA/third. RA/third directly generalizes RA to three-mode data. The RA/third model hypothesizes that there are redundancy variates which are constant or stable over occasion or group. However, the root variance explained between each pair of X-variates and Y-variates varies over occasion. To model RA/third a necessary assumption is that $\mathbf{S}_{XX}$ is common over a third mode. If $\mathbf{S}_{XX}$ is not derived from constant **X** over the third mode then it must be estimated from the data.

The RA/third model is

$$\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k = \mathbf{W}\mathbf{H}_k\mathbf{V}', \tag{5.1}$$

for $k = 1,\ldots,g$, where **W** is the $m \times q$ matrix of uncorrelated common canonical variates for the X-variables, **V** is an $p \times r$ orthonormal matrix of redundancy variates for the Y-variates, and $\mathbf{H}_k$ is $q \times r$ matrix whose elements are root of the variance explained of the Y-variates, $\mathbf{V}'\mathbf{y}$, by the X-variates, $\mathbf{W}'\mathbf{x}$ (**y** and **x** represent vectors of random variables). If the X-variables are constant over the third mode, such as may be the case with longitudinal data, replace $\mathbf{X}_k$ by **X**.

Note the choice of weight for $\mathbf{X}_k'\mathbf{Y}_k$ in (5.1) of $\frac{1}{n_k - 1}$. This weighting implies one is modeling $\mathbf{S}_{XY}$. If one is modeling data from multiple datasets with different sample sizes this gives each dataset the same weight in the analysis. However, one may just as easily model $\mathbf{X}_k'\mathbf{Y}_k$ instead, yielding an analysis that effectively weights by sample size. All of the results in this chapter and in Chapter Six apply with minor modifications to this alternative weighting. Also, if one has longitudinal data n_k may be replaced by a constant *n*.

Finding **V** and **W** such that the total variance explained of $\mathbf{V}'\mathbf{y}$ when regressed on $\mathbf{W}'\mathbf{x}$ is maximized is equivalent to modeling the following in the Tucker2 or PARAFAC (orth.) framework:

$$\frac{1}{n_k - 1}\mathbf{X}_k^{*\prime}\mathbf{Y}_k = \mathbf{W}^*\mathbf{H}_k\mathbf{V}' + \text{Error}_k, \tag{5.2}$$

where $\mathbf{W}^*$ is the $m \times r$ orthonormal matrix $\mathbf{W}^* = \mathbf{S}_{XX}^{1/2}\mathbf{W}$, $\mathbf{X}_k^* = \mathbf{X}_k\mathbf{S}_{XX}^{-1/2}$, and **V** and $\mathbf{H}_k$ are defined as for (5.1). To see that minimizing the sums of squares of the error term in (5.2) maximizes the sum of the variances explained of the Y-variables, recognize first that $\mathbf{H}_k[i, j]$ is indeed the root of the variation of the j^{th} Y-variate, $\mathbf{v}_j'\mathbf{y}$ explained by the i^{th} X-variate, $\mathbf{w}_i'\mathbf{x}$.

This is so because for the Tucker2 solution $\mathbf{H}_k = \mathbf{W}^{*\prime}\mathbf{C}_k\mathbf{V}$, and for the PARAFAC (orth.) solution $\mathbf{H}_k = \text{diag}\left(\mathbf{W}^{*\prime}\mathbf{C}_k\mathbf{V}\right)$ (Kroonenberg 1983), where $\mathbf{C}_k = \frac{1}{n_k - 1}\mathbf{X}_k^{*\prime}\mathbf{Y}_k$. Thus

$$\mathbf{H}_k[i, j] = \frac{1}{n_k - 1}\mathbf{w}_i^{*\prime}\mathbf{X}_k^{*\prime}\mathbf{Y}_k\mathbf{v}_j = \frac{1}{n_k - 1}\mathbf{w}_i'\mathbf{X}_k'\mathbf{Y}_k\mathbf{v}_j.$$

But the variance of $\mathbf{v}_j'\mathbf{y}$ explained in a regression against $\mathbf{w}_i'\mathbf{x}$ is

$$\frac{1}{n_k - 1}\mathbf{v}_j'\mathbf{Y}_k'\mathbf{X}_k\mathbf{w}_i(\mathbf{w}_i'\mathbf{X}_k'\mathbf{X}_k\mathbf{w}_i)^{-1}\mathbf{w}_i'\mathbf{X}_k'\mathbf{Y}_k\mathbf{v}_j = \left(\frac{1}{n_k - 1}\right)^2\left(\mathbf{w}_i'\mathbf{X}_k'\mathbf{Y}_k\mathbf{v}_j\right)^2,$$

noting that $(\mathbf{w}_i'\mathbf{X}_k'\mathbf{X}_k\mathbf{w}_i)^{-1} = \frac{1}{n_k - 1}$. Next, by **Proposition 3.2** the sums of squares of the $\mathbf{H}_k$ is maximized. Since the sums of squares of the $\frac{1}{n_k - 1}\mathbf{X}_k^{*\prime}\mathbf{Y}_k$ terms represent the total sums of squares of the Y-variables explainable by the X-variables, the lack of fit being minimized in (5.2) is just these sums of squares explainable by the relationship that are not being fit by the model over the third mode.

5.2.2 Canonical Variate Analysis over a Third Mode

CVA/third is a direct generalization of CVA. It is appropriate in the following situation; when the data have unchanging group structure, that is, the X-variables are group indicators; and the within-groups covariance matrix, $\mathbf{S}_{\text{YY(WITHIN)}}$, is stable over the third mode ($\mathbf{S}_{\text{YY(WITHIN)}}$ is distinguished from $\mathbf{S}_{\text{YY(TOTAL)}}$). Note that $\mathbf{S}_{\text{YY(WITHIN)}}$ has to be estimated from the data. The CVA/third scenario outlined here could also be approached by Campbell and Tomenson's (1983) model if the data are from multiple datasets or groups, or by the CVA/time model of Chapter Eight if the data are longitudinal.

Recall from Section **2.2.2** that CVA is equivalent to modeling the group means in the space of the variables transformed by the Mahalanobis transformation. CVA/third extends this conception of CVA to finding planes (components) that maximize the total dispersion over the third mode in the transformed space of the transformed variables.

The arguments for CVA/third are analogous to those of RA/third, except here one maximizes the variance explained in the transformed space of the Y-variables, i.e., the dispersion. The CVA/third model is

$$\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k = \mathbf{W}\mathbf{H}_k\mathbf{V}',$$

where $\mathbf{W}$ is the $m \times q$ matrix of uncorrelated common canonical variates for the X-variables, $\mathbf{V}$ is an $p \times r$ matrix of uncorrelated variates for the Y-variates, ($\mathbf{V}'\mathbf{S}_{\text{YY(WITHIN)}}\mathbf{V} = \mathbf{I}$) and $\mathbf{H}_k$ is a

$q \times r$ matrix of root transformed variances of the Y-variates, $\mathbf{V}'\mathbf{y}$, explained by the X-variates, $\mathbf{W}'\mathbf{x}$. If the X-variables are constant over the third mode, replace $\mathbf{X}_k$ with $\mathbf{X}$.

Finding $\mathbf{W}$ and $\mathbf{V}$ that maximize the sum of the transformed variance explained of the Y-variables by the X-variables is equivalent to modeling the following in a Tucker2 or PARAFAC (orth.) framework:

$$\frac{1}{n_k - 1}\mathbf{X}_k^{*\prime}\mathbf{Y}_k^{*} = \mathbf{W}^{*\prime}\mathbf{H}_k\mathbf{V}^{*} + \text{Error}_k, \tag{5.3}$$

where $\mathbf{W}^{*}$ is an $m \times r$ orthonormal matrix such that $\mathbf{W}^{*} = \mathbf{S}_{XX}^{1/2}\mathbf{W}$, $\mathbf{V}^{*}$ an $n \times r$ orthonormal matrix such that $\mathbf{V}^{*} = \mathbf{S}_{YY(\text{WITHIN})}^{1/2}\mathbf{V}$, $\mathbf{H}_k$ is a $q \times r$ matrix, $\mathbf{X}_k^{*} = \mathbf{X}_k\mathbf{S}_{XX}^{-1/2}$ and $\mathbf{Y}_k^{*} = \mathbf{Y}_k\mathbf{S}_{YY}^{-1/2}$.

To see that minimizing the sums of squares error term of (5.3) maximizes the sum of the explained of the transformed variance, note first that $\mathbf{H}_k[i, j]$ is indeed the root of the transformed variation of the j^{th} Y-variate, $\mathbf{v}_j'\mathbf{y}$, explained by the i^{th} X-variate, $\mathbf{w}_i'\mathbf{x}$. This is so because for the Tucker2 solution $\mathbf{H}_k = \mathbf{W}'\mathbf{C}_k\mathbf{V}$, and for the PARAFAC (orth.) solution $\mathbf{H}_k = \text{diag}(\mathbf{W}'\mathbf{C}_k\mathbf{V})$ (Kroonenberg 1983), where $\mathbf{C}_k = \frac{1}{1 - n_k}\mathbf{X}_k^{*\prime}\mathbf{Y}_k^{*}$. Thus

$$\mathbf{H}_k[i, j] = \mathbf{w}_i^{*\prime}\mathbf{X}_k^{*\prime}\mathbf{Y}_k^{*}\mathbf{v}_j = \mathbf{w}_i'\mathbf{X}_k^{*\prime}\mathbf{Y}_k^{*}\mathbf{v}_j.$$

Now the variance (in the transformed space) of $\mathbf{v}_j'\mathbf{y}$ explained by $\mathbf{w}_i'\mathbf{x}$ is

$$\frac{1}{1 - n_k}\mathbf{v}_j'\mathbf{Y}_k^{*\prime}\mathbf{X}_k\mathbf{w}_i(\mathbf{w}_i'\mathbf{X}_k'\mathbf{X}_k\mathbf{w}_i)^{-1}\mathbf{w}_i'\mathbf{X}_k'\mathbf{Y}_k^{*}\mathbf{v}_j = \left(\frac{1}{1 - n_k}\right)^2\left(\mathbf{w}_i'\mathbf{X}_k'\mathbf{Y}_k^{*}\mathbf{v}_j\right)^2,$$

noting that $(\mathbf{w}_i'\mathbf{X}_k'\mathbf{X}_k\mathbf{w}_i)^{-1} = \frac{1}{n_k - 1}$. Next, by **Proposition 3.2** the sums of squares of the $\mathbf{H}_k$ are being maximized. Hence the error being minimized is the sums of squares explainable by the relationship that is not being fit by the model over a third mode.

5.2.3 Canonical Correlation Analysis over a Third Mode

This section defines two generalizations of CCA, based on two different distinguishing features of CCA. CCA generates variates $\mathbf{V}$, which are both uncorrelated with respect to the total covariance of the Y-variables, and with respect to the matrix of error terms, where the error is the total variation less the sums of squares explained by a multivariate regression. That is: $\mathbf{V}'\mathbf{S}_{YY}\mathbf{V} = \mathbf{I}$ and $\mathbf{V}'(\mathbf{S}_{YY} - \mathbf{S}_{XY}'\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY})\mathbf{V} = \mathbf{I} - \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with the squared canonical correlations. Thus one can generalize CCA to the third mode by defining Y-variates which are uncorrelated with respect to a stable total variation, or by defining Y-variates which are uncorrelated with respect to a stable error term. On the other hand, taking the former approach leads to a method which can be shown to maximize the sums of the squared correlations between the variates. However it requires the awkward assumption that both $\mathbf{S}_{XX}$ and $\mathbf{S}_{YY}$ are stable over the third mode. Taking the latter approach leads to a method which has

the identical form of that of CVA/third (5.3) in Section **5.2.2**, except that the X-variables are no longer restricted to be indicators, and the error is defined as above. This method requires one to assume a stable error matrix and maximizes the sum of a weighted variance explained.

5.2.4 Procrustes Rotation over a Third Mode

Procrustes rotation is traditionally defined as a method that finds an orthogonal transformation **Q** such the point configuration of **YQ** is similar to that of **X** (see Section **2.2.4**). However, Procrustes rotation can also be defined as a method that finds orthogonal X-variates and orthogonal Y-variates such that the sum of the squared covariances between the pairs of corresponding X-variates and Y-variates is maximized (see Section **2.2.4)**. In this section I generalize PR to three-mode data along the lines of the former definition. Putting PR/third in the PARAFAC and Tucker2 frameworks allows me to maximize the sum of the squared covariances over the third mode. Note that one can also define Procrustes rotation as a method that finds pairs of X-variates and Y-variates such that sum (not squared) of the covariances is maximized. However, this definition has the disadvantage that covariances of differing signs would cancel each other out in a summation, so that a strong relationship that changed in sign would receive less weight. The attractive feature about PR/time is that it is not necessary to assume either $\mathbf{S}_{XX}$ or $\mathbf{S}_{YY}$ is constant.

The PR/third model is

$$\frac{1}{n_k - 1}\mathbf{X}_K'\mathbf{Y}_k = \mathbf{W}\mathbf{H}_k\mathbf{V}' + \text{Error}_k$$

where **W** is a $m \times q$ orthonormal matrix of X-variates and **V** is a $n \times r$ orthonormal matrices and $\mathbf{H}_k$ is a $q \times r$ matrix.

To see that PR/third maximizes the sums of the squared covariances, note that $\mathbf{H}_k[i, j]$ is indeed the root of the covariance of the j^{th} Y-variate, $\mathbf{v}_j'\mathbf{y}$, and the i^{th} X-variate, $\mathbf{w}_i'\mathbf{x}$. This is so because for the Tucker2 solution, $\mathbf{H}_k = \mathbf{W}'\mathbf{C}_k\mathbf{V}$, and for the PARAFAC (orth.) solution $\mathbf{H}_k = \text{diag}(\mathbf{W}'\mathbf{C}_k\mathbf{V})$ (Kroonenberg 1983), where $\mathbf{C}_k = \frac{1}{1 - n_k}\mathbf{X}_k'\mathbf{Y}_k$. Thus $\mathbf{H}_k[i, j] = \mathbf{w}_i'\mathbf{X}_k'\mathbf{Y}_k\mathbf{v}_j$. Now by **Proposition 3.2** $\sum_{k=1}^{g} \text{trace}(\mathbf{H}_k^2)$ is maximized.

For comparison's sake, if one wanted to generalize Procrustes rotation in the sense of maximizing the sum of the covariances, then it is simple to show that one obtains **W** and **V** by performing a singular value decomposition on the sum of the $\mathbf{X}_k'\mathbf{Y}_k$ matrices. That is, $\mathbf{W}\mathbf{J}\mathbf{V}' = \sum_{k=1}^{g}\mathbf{X}_k'\mathbf{Y}_k$. Further, say one wanted to generalize PR less in the spirit of finding common variates, but more in the spirit of finding an orthogonal rotation **Q** by which to rotate $\mathbf{Y}_k$

to maximum similarity to $\mathbf{X}_k$. That is, one wished to find $\mathbf{Q}$ that minimizes the following expression:

$$\sum_{k=1}^{g} \text{trace}(\mathbf{QY}_k - \mathbf{X}_k)'(\mathbf{QY}_k - \mathbf{X}_k).$$

Then again it is easy to show that one performs a singular value decomposition on the sum of the $\mathbf{X}_k' \mathbf{Y}_k$ matrices.

5.2.5 Which Transformations to Use

The decision of whether to transform the X-variables by $\mathbf{S}_{XX}^{-1/2}$ and/or to transform the Y-variables by $\mathbf{S}_{YY}^{-1/2}$ is a central one. It determines the metric in which the fit is being maximized and the lack of fit minimized. In the standard case it is equivalent to choosing between CCA, RA and PR. With three-mode data it is equivalent to choosing between CVA/third, CCA/third, RA/third and PR/third. This choice is determined by several factors; whether a set of variables stays constant; whether a covariance matrix of either the X-variables or Y-variables is hypothesized to be stable over the third mode; the hypothesized nature of cause and effect between the X-variables and the Y-variables; and what one wants to emphasize in his analysis.

For example, one is probably not interested in modeling the variation of a set of variables which are constant. Suppose the X-variables are group indicators that are constant over time. To choose PR/third over RA/third or CVA/third would imply one is maximizing the covariance between the two sets. The covariance, however, involves variation of the X-variables, and one is likely not interested in modeling the variation of the X-variables since they are constant over the third mode anyway. One is likely to be more interested in modeling just the variation of the Y-variables. Thus RA/third or CVA/third would be more appropriate. A similar logic would hold if the X-variables were not constant but the covariance of the X-variables, $\mathbf{S}_{XX}$, were stable.

Also, one may think in terms of cause and effect. One may hypothesize the X-variables cause variation in the Y-variables in a regression sense. If the X-variables are not constant over time, one may nevertheless choose to estimate of $\mathbf{S}_{XX}$ and to transform the X-variables by $\mathbf{S}_{XX}^{-1/2}$. This reduces the importance of the variation of the X-variables in the estimation, though in an uneven way; that is, since $\mathbf{S}_{XXk} \neq \mathbf{S}_{XX}$, for $k = 1,\ldots,g$, $\mathbf{S}_{X^*X^*k} \neq \mathbf{I}$ where $\mathbf{X}_k^* = \mathbf{X}_k \mathbf{S}_{XX}^{-1/2}$, though for some occasions $\mathbf{S}_{X^*X^*k}$ may be closer to $\mathbf{I}$ than others.

What one wishes to emphasize may also play a role in choosing the transformation. For example, with grouped data with stable within-group covariances, one could use the within-groups covariance to transform the Y-variables to get scale invariance. However, if the Y-variables are directly comparable one may prefer not to scale the data, but rather perform a RA/third which analyzes the Y-variables in their raw form. As another example, RA/third would usually be preferred over CVA/third if the within-groups covariance matrix was not hypothesized to be stable. In this case one may wish to standardize the variance to make each Y-variable of equal importance by scaling each variable such that they all have an equal total variation.

However, if the within-group covariances were not too dissimilar, one could estimate a common within-groups covariance matrix anyway and transform the Y-variables to achieve a crude scale invariance.

In summary, the methods introduced in this chapter are exploratory. When deciding what kind of analysis to perform the researcher needs to consider the nature of his data and the phenomena, and what he wants to bring out in an analysis.

5.3 HOW TO EVALUATE THE FIT OF THE MODEL

The choices in fitting the model are whether to use the PARAFAC (orth.) or Tucker2 framework, and how many components to include in the model. While there is no systematic approach to fitting the model such as hypothesis tests, there are certain principles and pointers to guide in the decision. Basically, one evaluates the sums of squares lack of fit and the interpretability of the model terms.

It may help to view the modeling frameworks as a hierarchy going from the most complex model that explains the relationship perfectly to the least complex which would have the greatest lack of fit. The most complex model is a separate CCA, RA or PR at each occasion. These will explain the relationship perfectly at each occasion or for each dataset. The next most complex model is the Tucker2. The Tucker2 models two sets of stable variates. A linear relationship which varies in strength is assumed to exist between each variate of the X-set and each variate of the Y-set. Then there is the PARAFAC (orth.) model, which hypothesizes pairs of stable variates with the linear relationship varying in strength. Unlike the Tucker2 the relationship is modeled strictly between members of a pair. Lastly, the simplest model would be to model identical relationships at each occasion, with identical variates and equal strength of relationship.

As an aid to the evaluation and interpretation of a three-mode model one can examine what I shall call the matrix of explained sums of squares. This matrix shows the sums of squares explained of each Y-variable by each Y-component. To determine how much of the sums of squares of the i^{th} Y-variable is explained by the j^{th} Y-component, $\mathbf{v}_j'\mathbf{y}$, first consider the weight of the (orthonormal) component corresponding to that variable. If one is performing RA/third or PR/third this is $\mathbf{v}_j[i]$; if CVA/third or CCA/third this is $\mathbf{v}_j^*[i]$. Square this weight and multiply it by the sum of the squared core elements corresponding to the j^{th} Y-component. The matrix one obtains allows one to see in a simple way what variables are well explained by what components. The interpretation of what the sums of squares means depends on the method. For example, for RA/third, the sums of squares represent the variance of a given variable explained by a given component.

When examining fit to compare the PARAFAC (orth.) and Tucker2, the values of the off-diagonal elements are worth looking at. Small off-diagonal elements suggest the PARAFAC (orth.) model is more appropriate. The Tucker2 will always explain more sums of square, but the PARAFAC (orth.) requires fewer parameters to be estimated, and it has an advantage in interpretability because each X-variate is related only to one Y-variate.

When deciding on the number of components to include one can use a scree plot such as that used in multidimensional scaling. One plots the sums of squares lack of fit against the number of components. The point where the curve levels out suggests the number of components to include in the model.

Other points to consider when evaluating the fit of a three-mode model are the nestedness and scale invariance of the solutions. In Chapter Three I show that the PARAFAC (orth.) solutions are nested. The nestedness property implies that a rank-*f* solution is always a subset of a rank-(*f*+1) solution. This allows one to evaluate the fit contributed by each pair of components when comparing solutions of differing rank. The Tucker2 solutions do not have this property. In Chapter Nine I discuss the topic of scale invariance. While three-mode methods are generally not scale invariant, the Tucker2 and PARAFAC (orth.) models do have approximate scale invariance properties if the fit is very good.

In summary, it should be clear from the discussion that how much complexity the researcher decides to model will be in part subjective.

5.4 AN EXAMPLE

I will go into more details about the models in Section **5.5**, and about the interpretation of the models in Chapter Six. But first I present an example to illustrate what has been laid out so far. The following example is an application of RA/third. The U.S. Geological Survey in cooperation with the University of Virginia's Department of Environmental Sciences performed a study, "Sensitivity of Stream Basins in Shenandoah National Park to Acid Deposition" (Lynch & Dise 1985). This study investigated the acidification of the streams in the said park. The acid presumably came in the form of acid rain from man-made sources such as sulfur bearing pollutants. The purpose of the study was to identify and evaluate geological factors relating to the sensitivity of basins to acid deposition. There were 56 streams located throughout the park, which was divided into 56 corresponding basins. Per recommendation of the authors I discarded three streams which ran parallel to roads, leaving 53 streams in my analysis.

Geological measurements were taken once over the whole basin and are assumed to be unchanging over time. These I designate the X-variables. The geological measurements taken included the classification of the types of underlying bedrock. Since a stream basin frequently contained more than one type of underlying bedrock, the basins are assigned a percentage for each bedrock type. The bedrock classifications included Catoctin, Pedlar, Old Rag, Hampton, Antietam, Swift Run and Weverton. To avoid colinearity among classification variables, I did not assign variables for the Swift Run and Weverton bedrocks, which taken together were still less common than any of the other bedrock types. Other geological measurements were: the percent of the basin above 2400 feet in elevation; an indicator variable for whether the site was on the east or west slope (E/W); an indicator variable for whether 5% or more of the basin was developed; and the drainage density (DD), calculated by dividing the length of the stream by the area of the basin.

For each stream, streamwater measurements were taken on six occasions, always at the same site. These were the Y-variables. The measurements taken were pH; alkalinity, which was

defined as a measure of ability to neutralize or buffer strong acids; conductivity, which is an indirect measure of how much ionization has occurred; temperature; stream discharge; the base cations calcium (Ca^{++}), magnesium (Mg^{++}), potassium (K^{+}), sodium (Na^{+}) and ammonium (NH_4^{+}); and the acidic anions chloride (Cl^{-}), sulfate ($SO_4^{=}$), nitrate (NO_3^{-}), and silica ($SiO_4^{=}$).

The researchers have analyzed their data carefully. The intention of my analysis is not to replicate their work, but to illustrate my method, which will emphasize the investigation of what is changing or not changing over time. The researchers' statistical analysis was largely concentrated in two parts. First, they averaged the data over the six occasions and performed a multiple regression for each chemical measurement against the geological measures. The multiple regression with alkalinity as a response showed that the geological variables explained 0.962 of the variation. Bedrock alone had an R^2 of 0.947. These were the key statistical findings of their study. Geology was also important for predicting amount of base cations and silica, with R^2's ranging from 0.855 to 0.944. Sulfate, nitrate and chlorine concentrations were modeled with R^2's of 0.535, 0.695 and 0.600 respectively.

Next, in an attempt to examine time differences with respect to different bedrocks the researchers subtracted the data for January 1982 from that of May 1982, and then that of June 1982 from September 1981. They do this for only the 28 sites that are classified 75% or more as one type of bedrock, and for only four of the more common bedrocks. Then they compare the mean differences between the sites classified to the various bedrocks. These pairwise comparisons revealed that in warmer months there was in general more alkalinity as well as more base cations, and in particular there was relatively more in certain bedrocks susceptible to carbonic weathering. These increases are explained by more carbonic acid weathering due to higher levels of carbon dioxide in the soil that result from greater microbial and plant activity during these times of year.

My analysis attempts to model the relationship between the geological variables and the stream measurements over time. Since the geological measurements are constant it is appropriate to transform these measurements to uncorrelated variates by multiplying by the Mahalanobis transformation ($\mathbf{S}_{XX}^{-1/2}$). **Table 5.1** shows the error variances of the fourteen water variables at each occasion. (I define error variance as the sum of squared residuals obtained if one performs a separate multivariate regression at each occasion). One sees the error variances of the streamwater measurements are not constant over time. Note for example the changes in the variation in pH and alkalinity. To transform based on an averaged covariance matrix would put the between-group differences in a metric that may not make sense. Hence I choose not to transform the water chemistry variables. (See Section **5.2.5** for a discussion on the appropriate choice of transformations.) However, the responses are

Table 5.1. Error Variance of Responses

Measurement	Aug. 1981	Sept. 1981	Jan. 1982	March 1982	May 1982	June 1982
discharge	0.08	0.01	0.07	0.07	2.19	0.11
conductivity	0.13	0.55	0.05	0.05	0.48	0.10
pH	0.01	0.01	0	0	**4.19**	0.01
temperature	0.99	0.64	0.28	0.28	0.69	0.54
Ca^{++}	0.19	0.60	0.05	0.05	0.25	0.09
Mg^{++}	0.14	0.36	0.05	0.05	0.14	0.09
Na^{+}	0.13	0.24	0.12	0.12	0.28	0.12
K^{+}	0.10	0.14	0.03	0.03	1.90	0.05
alkalinity	**0.22**	**0.61**	0.06	0.06	0.06	0.08
$SO_4^{=}$	0.18	0.35	0.30	0.30	0.59	0.30
Cl^{-}	0.21	0.28	0.24	0.24	2.04	0.16
$SiO_4^{=}$	0.19	0.27	0.10	0.10	0.18	0.19
NO_3^{-}	0.08	0.06	0.30	0.30	2.96	0.12
NH_4^{+}	0.01	0	0	0	4.50	0.01

standardized to have an equal variance over the six occasions. This standardization gives each of the variables equal weight in the analysis while allowing the measurements to vary over time. The choice of transforming the geological variables but not the water measurement variables defines the analysis as an RA/third analysis. The RA/third model finds uncorrelated weighted sums of the geological variables that explain the maximum amount of variation of the streamwater variables over time.

The next step in the analysis is to determine whether the PARAFAC (orth.) or Tucker2 model is appropriate and how many pairs of components are appropriate. I do this by comparing the fit and the interpretability of the competing models. I start by presenting the estimates of the core elements PARAFAC (orth.) estimates in **Table 5.2**. To save space they are not presented as

Table 5.2 PARAFAC (orth.) Core

Date / Component	Aug. 1981	Sept. 1981	Jan. 1982	Mar. 1982	May 1982	June 1982
1st	2.25	2.52	1.83	1.61	2.2	2.06
2nd	0.93	0.95	0.95	0.99	1.03	1.15
3rd	0.90	1.05	0.56	0.76	0.71	0.76
4th	-0.03	0.01	0.01	-1.61	0.02	0.02
5th	-0.13	-0.20	-0.30	-0.83	-0.25	-0.3
6th	-0.23	-0.27	-0.28	-0.35	-0.31	-0.42
7th	0.14	0.06	0.16	0.3	0.28	0.27
8th	0.22	0.28	0.07	-0.1	-0.07	0.09
9th	0.18	0.13	0.12	0.03	0.13	0.12

a diagonal matrix for each occasion, rather as vectors for each occasion. Only the first four components have core elements greater than one in magnitude. Also, as will be discussed later, the first four components have interpretations that lend credence to their selection in the model. Thus the PAFARAC (orth.) model with four pairs of components is a competitive model. The SAS programming codes for the Tucker2 and PARAFAC (orth.) models are found in Appendix Three.

The four-component Tucker2 solution is examined for comparison. The estimates of its core matrices are shown in **Table 5.3**. The sums of squares explained is equal to the sums of squares of the core matrix. The Tucker2 explains 41.8 (81.3%) of the variation, which is only modestly more than the

Table 5.3 Core Matrices for the Tucker2

August 1981			
2.26	-0.02	-0.11	-0.21
-0.63	**0.90**	0.13	0.17
-0.01	0.23	**0.81**	-0.26
0.29	-0.10	0.22	**-0.10**

September 1981			
2.54	-0.13	-0.09	-0.22
-0.10	**0.89**	-0.01	0.20
-0.06	0.31	**0.96**	-0.34
-0.10	-0.14	0.13	**0.01**

January 1982			
1.82	0.03	-0	0.16
-0.02	**0.91**	0.11	0.18
-0.03	0.16	**0.45**	-0.26
0.24	-0.1	0.2	**-0.05**

March 1982			
1.55	-0.35	0.36	**1.25**
0.50	**1.05**	-0.17	-0.10
-0.26	-0.30	**0.90**	-0.01
0.41	0.27	-0.54	**-1.13**

May 1982			
2.19	-0.03	0.13	-0.22
0.42	**0.99**	0.05	0.21
-0.02	0.13	**0.59**	-0.32
-0.03	-0.10	0.20	**-0.08**

June 1982			
2.06	0.06	0.1	-0.12
0.37	**1.12**	-0.08	0.29
0.07	0.01	**0.63**	-0.35
0.04	-0.1	0.24	**-0.08**

rank-four PARAFAC (orth.) solution which explains 39.0 (75.9%). Note that the total variation that can be explained by the relationship by separate multivariate regressions at each occasion is 51.37. Further, the components are similar to those of the PARAFAC (orth.) model, as one can

see by comparing the weights for geological variables in **Table 5.4** for the PARAFAC (orth.) model to those in **Table 5.5** for the Tucker2, and the weights for the water measurements for the PARAFAC (orth.) model in **Table 5.6** to those for the Tucker2 in **Table 5.7**.

Table 5.4 Canonical Variate X-weights for PARAFAC (orth.)

Geological Variable	First Comp.	Second Comp.	Third Comp.	Fourth Comp.
Antietam	-0.084	-0.311	0.682	0.588
Hampton	0.188	-0.938	1.167	1.025
Catoctin	1.229	-0.750	1.327	1.373
Pedlar	0.838	-0.489	1.888	1.475
Old Rag	0.436	-0.238	1.542	1.052
ab2400	-0.445	0.406	-0.586	-0.468
DD	0.124	-0.131	0.345	-0.947
E/W	0.228	-0.396	0.101	0.585
Dev.	0.256	-0.173	-0.261	-0.313

Table 5.5 Canonical Variate X-weights for the Tucker2

Geological Variable	First Comp.	Second Comp.	Third Comp.	Fourth Comp.
Antietam	-0.105	-0.314	0.540	0.695
Hampton	0.148	-0.907	0.858	1.331
Catoctin	1.169	-0.688	0.943	1.668
Pedlar	0.766	-0.433	1.489	1.871
Old Rag	0.382	-0.284	1.299	1.393
ab2400	-0.425	0.342	-0.409	-0.648
DD	0.193	-0.246	0.527	-0.851
E/W	0.193	-0.382	-0.031	0.597
Dev.	0.271	-0.204	-0.189	-0.250

Also, the core elements are similar. The diagonal elements of the Tucker2 core in **Table 5.3** are close to that of the PARAFAC (orth.) of **Table 5.2**. For example, for the first occasion they are 2.26, 0.9, 0.81 and -0.1 versus 2.25, 0.93, 0.9 and -0.03. One can see in **Table 5.3** that the off-diagonal elements of the core matrices for the Tucker2 are generally small, except for those at occasion four. Given the similarity in fit and interpretation, the PARAFAC (orth.) model is preferable to the Tucker2 because it has less terms and is simpler to interpret. For the rest of this section I discuss the PARAFAC (orth.) model and its estimates.

The interpretation of the first four components of the PARAFAC (orth.) model is consistent with the analysis given by the researchers in their study. The first component relates to the level of alkalinity and the concentrations of base cations. First inspect the matrix of sums of squares explained, **Table 5.8**, which indicates how much variance a given geological component explains of a given water measurement variable. Note that the total variance of alkalinity explained by the first component is 4.11. Compare this with the total variance explainable of 4.86 (if separate regression were performed at each occasion), and a total variation of 6.0. Also, there are large variances explained for the base cations Ca^{++}, Mg^{++} and Na^{+}, and the acid silica ($SiO_4^{=}$). These are all products of the same process, the carbonic weathering of minerals high in silica and in base anions. This process tends to increase the alkalinity through the production of carbonic acid, which is a buffer against strong acids. The corresponding geology variate, the X-variate or

Table 5.6 Y-weights for PARAFAC

Measurement	First Comp.	Second Comp.	Third Comp.	Fourth Comp.
discharge	0.036	0.297	-0.271	-0.113
conductivity	**0.369**	-0.293	-0.108	-0.156
pH	0.113	0.058	-0.002	0.593
temperature	0.132	-0.08	0.351	0.048
Ca^{++}	**0.383**	0.018	-0.241	0.002
Mg^{++}	**0.365**	-0.261	-0.43	-0.006
Na^{+}	**0.364**	0.256	0.419	-0.063
K^{+}	-0.158	-0.566	0.182	0.308
alkalinity	**0.394**	0.012	-0.235	0.12
$SO_4^{=}$	0.14	-0.543	0.22	-0.231
Cl^{-}	0.265	-0.06	0.133	-0.065
$SiO_4^{=}$	**0.363**	0.229	0.458	0.003
NO_3^{-}	0.133	0.098	-0.089	0.321
NH_4^{+}	0.035	-0.029	0.034	0.579

Table 5.7. Y-weights for Tucker2

Measure-ment	First Comp.	Second Comp.	Third Comp.	Fourth Comp.
discharge	0.057	0.311	-0.256	0.260
conductivity	**0.373**	-0.274	-0.177	-0.143
pH	0.094	-0.050	0.239	0.513
temperature	0.134	-0.065	0.311	-0.269
Ca^{++}	**0.381**	-0.003	-0.206	0.074
Mg^{++}	**0.363**	-0.267	-0.410	0.087
Na^{+}	**0.367**	0.258	0.379	-0.139
K^{+}	-0.173	-0.592	0.243	0.079
alkalinity	**0.393**	-0.019	-0.193	0.059
$SO_4^{=}$	0.129	-0.508	0.138	-0.144
Cl^{-}	0.268	-0.049	0.116	-0.002
$SiO_4^{=}$	**0.361**	0.234	0.437	-0.125
NO_3^{-}	0.122	0.023	0.101	0.529
NH_4^{+}	0.017	-0.129	0.255	0.467

Table 5.8. Matrix of Sums of Squares Explained by Variable and Component

Measurement	First Comp.	Second Comp.	Third Comp.	Fourth Comp.	Total
discharge	0.03	0.53	0.29	0.03	0.88
conductivity	**3.60**	0.52	0.05	0.06	4.23
pH	0.34	0.02	0	**0.91**	1.27
temperature	0.46	0.04	0.48	0.01	0.99
Ca^{++}	**3.88**	0	0.23	0	4.11
Mg^{++}	**3.53**	0.41	**0.72**	0	4.63
Na^{+}	**3.51**	0.39	**0.68**	0.01	4.49
K^{+}	0.66	**1.93**	0.13	0.25	2.95
alkalinity	**4.11**	0	0.21	0.04	4.36
$SO_4^{=}$	0.52	**1.78**	0.19	0.14	2.66
Cl^{-}	**1.86**	0.02	0.07	0.01	1.96
$SiO_4^{=}$	**3.48**	0.31	**0.81**	0	4.60
NO_3^{-}	0.47	0.06	0.03	0.27	0.83
NH_4^{+}	0.03	0	0	**0.87**	0.90
Total Variation Explained	26.47	6.02	3.88	2.59	39.0

predictor variate, is seen in **Table 5.4**. This variate is uncorrelated with the other X-variates. Its weights are interpreted in the same sense that weights in a regression equation are. First, one sees the weights for type of bedrock are ordered as Catoctin (1.22), Pedlar (0.84), Old Rag (0.44), Hampton (0.18) and Antietam (-0.08). This ordering is the same as that of the regression weights predicted by the researchers and found in their regression equation for alkalinity. The other weights are also consistent with their regression equation for alkalinity.

This first and largest component has several implications. First it affirms the researchers' decision to average the data over time. Analyzing data averaged over time is after all a crude way to get common components. Indeed, it only makes sense if there are common components, otherwise averaging muddles the analysis. Second, it shows the advantage of the multivariate approach over the univariate approach in that it models the responses simultaneously. Alkalinity, the base cations Ca^{++}, Mg^{++}, Na^{+} and silica, which are all modeled individually as univariate responses in the researchers' analysis, are modeled in RA/time in a way that reveals their interrelationship. Further, what the analysis over time reveals is that although this process was roughly stable in strength over time, there were differences. At occasion four the total variance explained is 2.6 (one squares the core element to obtain the variance explained, which is 1.61 from **Table 5.2**). This small value is likely due to heavy rain during that month which would increase the proportion of runoff in the stream as opposed to ground discharge. Runoff has less of the chemicals that are formed by reactions in the soil and bedrock than does ground discharge, weakening the relative strength of these alkalinity and base cations and silica. The occasion where

this process is strongest is September 1981 where the variate explains a variation of 6.36 (2.52 squared), followed by August 1981 with 5.06. These higher values relate to the fact that in warmer weather there is more microbial and plant activity in the soil creating carbon dioxide and initiating carbonic acid weathering.

The second component pair is likewise interpretable as a process predicted and observed by the researchers in their analysis. It is related to precipitation of sulfurous compounds and the ions that result from the ensuing reactions with the bedrock. This is the process that researchers refer to as "acidification". In the table of variance explained, **Table 5.8**, one sees that 1.78 of the variation of sulfate, $SO_4^=$, is explained, and also 1.93 of potassium (K). From **Table 5.4** one sees that there is more sulfate on the western slope and less at higher altitudes. This can be accounted for by the effect of the prevailing winds bearing pollution from the west. Also, higher elevations have more rain and consequently more acid deposition. The high concentration of potassium may be due to greater reactivity with sulfur in bedrock consisting of minerals with high potassium content such as Hampton. What the analysis over time reveals is first, that the researchers' averaging of data over time was again plausible. Second, though the strength of the relationship does seem to be relatively stable, there is a weak increasing trend over time.

The variance explained by the third variate pair is a little less than that explained by the second (**Table 5.2**). The geological variate shows higher weights for bedrock that is granitous, such as Pedlar and Old Rag. The streamwater variate has larger weights for Silica and Na^{++}, which are byproducts of the plagioclastic weathering of granites, which also happen to be low in Mg^{++}. Hence this variate pair is interpretable as indicating plagioclastic weathering. The differences over time in the strength of the variances would need to be interpreted by the researchers for significance.

The fourth component pair is related to runoff effects of rainwater. It is one that would not fair well in an averaging over time. Indeed it was not mentioned by the researchers. One sees that the variance explained at the fourth occasion is 2.59 (-1.61 squared, from **Table 5.2**), but at other occasions it is close to zero. This is likely due to the unusually heavy runoff during March due to heavy rains and perhaps to melting snows. The mean stream flow was 3.2 cubic feet per second per square mile in March, versus 0.2 to 1.3 at the other occasions. The salient weight among the geology variates is drainage density, though altitude, east/west and bedrock type all play a role. In areas with poor drainage, the proportion of runoff will be greater, hence there will be greater runoff effects. Also, basins at higher altitudes receive more rain and have a greater runoff. The streamwater variables explained by this canonical variate are pH and NH_4^+. These both reflect the fact that streamwater with a higher proportion of runoff from rain or snow melts is more like rainwater. Rainwater has a pH of 4.22, lower than the lowest soils in the park which is about pH of 5. Also, ammonia is not found in ground discharge but rather derives strictly from atmospheric precipitation.

In summary, the RA/time analysis confirms the researchers conclusion found using conventional methods. However, it gave a more integrated view of the processes by using both a multivariate approach and modeling over time. It also raised some questions about the relationships over time that the researchers might profitably address.

5.5 SOME FURTHER CONSIDERATIONS

Having presented the CCA, RA and PR/third models and gone over in detail an application, I discuss some issues that shed further light on the nature of the modeling and the problems it solves. These include autocorrelation, the covariances between the Y-variables at different occasions and the (lack of) invariance of solutions to non-singular transformations.

5.5.1 Autocorrelation

Autocorrelation is a common phenomena with measurements made over time. In this section I attempt to answer the question of what effect autocorrelation has on models relating two sets of variables over time. The situation is clearest when one examines RA/third where the X-variables indicate group membership. Then $\frac{1}{n_k - 1}\mathbf{X}'\mathbf{Y}_k$ is a matrix whose i^{th} row is the group means for each Y-variate for the k^{th} occasion. Now it is easy to see that the autocorrelation for $\overline{Y}$ equals the autocorrelation for Y, and that the effect of autocorrelation weakens over time, i.e., $\text{corr}(\mathbf{Y}_k, \mathbf{Y}_{k+m}) = \sigma^m$. Thus a strong autocorrelation tends to make the structure of the data static; that is, little changes over time. Otherwise its presence is observed in the within-groups covariance matrix. The effect of autocorrelation is similar if the X-variables are continuous. In summary, the possible presence of autocorrelation does not require any extra model terms when modeling CCA, RA and PR over time.

5.5.2 Cross Occasion Covariances

When one has longitudinal data one will observe covariances of variables at different occasions. Take, for example, $\mathbf{Y}_r'\mathbf{Y}_s$, $r \neq s$. By modeling only $\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k$ or its transformations, it seems that one is ignoring information by not modeling these cross-occasion covariances. However, for certain important situations this is not so. Consider when one has constant $\mathbf{X}$ over time and is modeling RA/third. Then the sums of squares regression of the Y-variables explained by the X-variables at occasions r and s is $\mathbf{Y}_r'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_s$. But this is just the product of $(n_k - 1)\mathbf{S}_{xx}^{-½}\mathbf{X}'\mathbf{Y}_r$ and $(n_k - 1)\mathbf{S}_{xx}^{-½}\mathbf{X}'\mathbf{Y}_s$, two matrices which are already modeled in RA/third (with a different weighting). Further, $\mathbf{Y}_r'\mathbf{Y}_s - \mathbf{Y}_r'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_s$ is the sum of squares error, which is not modeled in RA/third. In this sense the covariance between variables at different occasions offers no new information.

An analogous argument can be made for CVA/third with longitudinal data. However, with data where $\mathbf{X}$ is not constant the situation is not clear. Later chapters approach the issue of modeling cross occasion covariances. Chapter Seven presents least squares methods that attempt to model some of these cross-occasion terms. Chapter Eight presents maximum likelihood methods that model the error terms.

5.5.3 Invariance of the Rank of the Solution to Non-Singular Transformations

In Section **5.2.5** I discussed which transformations to apply. In this section I add to that a brief discussion on the invariance of the solutions to the choice of transformation. It is known that the rank of a matrix is invariant to non-singular transformations, and that the column space of a matrix is invariant to non-singular transformations of the row space, and vice versa. Since CCA, RA and PR are based on the SVD of $\mathbf{S}_{XY}$ with the appropriate transformations (see Section **2.2**), one can draw two implications; first, that for a given dataset the solutions for CCA, RA and PR all are of the same rank; second, the X-variates for CCA and RA span the same space, as do the Y-variates for RA and PR.

These features hold true for the three-mode extensions of CCA, RA and PR if they are put in the framework of the Tucker2, but not if they are in the PARAFAC (orth.) framework. To see this, consider a series of *g* matrices, $\mathbf{Q}_1,\ldots,\mathbf{Q}_g$, which are modeled exactly by the Tucker2. That is: $\mathbf{Q}_1 = \mathbf{RS}_1\mathbf{T}'$, $\mathbf{Q}_2 = \mathbf{RS}_2\mathbf{T}'$, with **R** and **T** orthonormal, etc. Now consider an arbitrary, non-singular transformation **A** for the row space. Perform a SVD on **A**, $\mathbf{A} = \mathbf{MNP}'$. Then $\mathbf{AQ}_k = \mathbf{MNP}'\mathbf{RS}_k\mathbf{T}'$, for $k = 1,\ldots,g$. Now one can perform a SVD on $\mathbf{MNP}'\mathbf{R}$ to get $\mathbf{MNP}'\mathbf{R} = \mathbf{DEF}'$ and consequently $\mathbf{AQ}_k = \mathbf{DEF}'\mathbf{S}_k\mathbf{T}'$, or $\mathbf{AQ}_k = \mathbf{DS}_k^*\mathbf{T}'$, where $\mathbf{S}_k^* = \mathbf{EF}'\mathbf{S}_k$. One sees that one has a Tucker2 solution with the same column space **T**. On the other hand, if $\mathbf{S}_k$ is now restricted to be diagonal, one cannot generally find $\mathbf{S}_k^*$ that is diagonal to yield a PARAFAC (orth.) solution.

The main implication of this to modeling is that if one is uncertain about which transformation to apply to the data, then the Tucker2 is a safer model than the PARAFAC (orth.). Also, if the Tucker2 has a lower rank model that fits better than a higher rank PARAFAC (orth.) model, one might consider another transformation.

5.5.4 Concluding Comments

First note that categorical data are handled the same way as continuous data. Hence correspondence analysis is generalized to the third mode. See Section **2.2.2** for the interpretation of categorical data with CCA type analyses.

In summary, the methods of this chapter are flexible and exploratory. They require some subjective choices as to the choice of transformation, choosing the Tucker2 versus the PARAFAC (orth.) models and determining the number of components in the solution. Ultimately, as seen in the example in Section **5.4**, one may choose a model which does not explain all of what is going on in the data, but explains some of what is going on; that is, finds some structure to the data over the third mode.

CHAPTER SIX

GRAPHICAL METHODS

6.1 INTRODUCTION

In some circumstances a visual display more easily yields insights than the inspection of tables of parameter estimates. This is often the case with three-mode models as they have many variables and complex relationships between components. In this chapter I present the use of graphical methods in modeling two sets of variables over a third mode. The methods are based on the three-mode models of Chapter Five. These methods offer the user graphical views of the relationships between variables, components and modes.

The chapter is organized as follows: in Section **6.2** I introduce background on biplots and joint plots. Biplots for canonical correlation analysis (CCA) were developed by Ter Braak (1990), who also suggested biplots for redundancy analysis (RA), which I develop in more detail. I also develop biplots for Procrustes rotation (PR). In Section **6.3** I extend these biplots to

three-mode data with joint plots. In Section **6.4** I discuss plots of components scores for three-mode methods. Lastly, in Section **6.5** I present the use of residual plots in three-mode modeling. Note that all the programming code that generated the graphs is found in Appendix Four.

6.2 BIPLOTS

The biplot is a technique devised by Gabriel (1971) to represent a matrix approximately by a two-dimensional plot of two sets of two vectors. If the rank of the matrix is two, the representation is exact. A biplot is often, but not necessarily, derived from the two pairs of components corresponding to the largest singular values from the SVD. Eckart and Young (1936) proved that such a two-dimensional approximation to a matrix is optimal in a least squares sense. Generally, any rank-two approximation to a matrix **X** can be decomposed as $\mathbf{AB}'$:

$$\mathbf{X} \cong \mathbf{AB}' = \left[\mathbf{u}_1, \mathbf{u}_2\right]\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}\mathbf{v}_1'\\ \mathbf{v}_2'\end{bmatrix}.$$

There are three common equivalent ways of displaying this two-dimensional approximation to **X**.

1. $\mathbf{A} = \left[\mathbf{u}_1\sqrt{\lambda_1} \quad \mathbf{u}_2\sqrt{\lambda_2}\right]$ and $\mathbf{B} = \left[\mathbf{v}_1\sqrt{\lambda_1} \quad \mathbf{v}_2\sqrt{\lambda_2}\right]$ (6.1)
2. $\mathbf{A} = \left[\mathbf{u}_1\lambda_1 \quad \mathbf{u}_2\lambda_2\right]$ and $\mathbf{B} = \left[\mathbf{v}_1 \quad \mathbf{v}_2\right]$
3. $\mathbf{A} = \left[\mathbf{u}_1 \quad \mathbf{u}_2\right]$ and $\mathbf{B} = \left[\mathbf{v}_1\lambda_1 \quad \mathbf{v}_2\lambda_2\right]$.

Which of the three displays one chooses depends on whether one prefers to emphasize the rows or the columns.

If one thinks of the rows as corresponding to subjects and the columns as corresponding to variables, then **A** plots the subjects and **B** the variables. The first axis on the two-dimensional graph represents both the first component for **A** and the first component for **B**, while the second axis represents both the second component of **A** and the second component of **B**. Thus a biplot is two graphs superimposed upon each other: a graph of subject scores and a graph of variable scores. The approximate value of any variable for any subject can be determined by the inner product of the subject and variable vectors. Further, the inner product between vectors of two variables or two subjects is an approximate measure of the closeness or similarity of those variables or subjects. For example, in the third type of biplot the covariance between any two variables is approximated by their inner product on the biplot.

6.2.1 Biplots for Canonical Correlation Analysis

In this section and in the subsequent ones I outline biplots for CCA, RA and PR. These plots have in common that they provide an approximation to $\mathbf{S}_{XY}$, the matrix of covariances between the X-variables and Y-variables. They will differ in the invariance properties that they invoke. The resulting biplots will be based on vectors whose values are directly related to the variate weights of the associated method.

As a preliminary I shall review the interpretation of $\mathbf{S}_{XY}$. If the X-variables and Y-variables are standardized to unit length, then $\mathbf{S}_{XY}[i,j]$ is the correlation between the i^{th} X-variable, x_i, and the j^{th} Y-variable, y_j. If only the X-variables are standardized to unit length, then $\mathbf{S}_{XY}[i,j]$ is the square root of the regression variance of y_j explained by x_i in a simple linear regression of y_j on x_i (by the regression variance I refer to the sums of squares regression divided by the $n-1$, which is also sometimes known as the mean square regression). If neither the X-variables nor the Y-variables are standardized, then $\mathbf{S}_{XY}[i,j]$ is the covariance between x_i and y_j.

In this section I describe biplots of structure coefficients associated with CCA. These biplots will also have the property that they approximate $\mathbf{S}_{XY}$ with invariance to linear transformations of both the X-variables and the Y-variables. Plots of structure coefficients have been proposed as a means to graphically interpret canonical correlation analysis (Caillez & Pagès, 1976; Israëls, 1987; van der Geer, 1986). Structure coefficients are defined as the correlations between the variables and the canonical variates. Ter Braak (1990) puts plots of structure coefficients in the framework of biplots. He shows their optimality property and discusses their interpretation. The following development of biplots for CCA is from Ter Braak (1990).

Let the matrix $\mathbf{AB}'$ denote a rank-two weighted least squares approximation to $\mathbf{S}_{XY}$ where the variables are standardized, thus $\mathbf{S}_{XY}$ is a correlation matrix. The ways to factor $\mathbf{AB}'$ are of the form:

$$\mathbf{A} = \left[\mathbf{S}_{XX}^{½}\mathbf{W}^{*}\mathbf{E}^{\alpha-1}\right]_2 \qquad \mathbf{B} = \left[\mathbf{S}_{YY}^{½}\mathbf{V}^{*}\mathbf{E}^{\alpha}\right]_2, \tag{6.2}$$

where $[\mathbf{Z}]_2$ indicates the first two columns of a matrix $\mathbf{Z}$; $\mathbf{V}^{*}$ is a $p \times r$ orthonormal matrix such that $\mathbf{V}^{*} = \mathbf{S}_{YY}^{½}\mathbf{V}$, where $\mathbf{V}$ is the $p \times r$ matrix of Y-canonical variates; $\mathbf{W}^{*}$ is an $m \times r$ orthonormal matrix such that $\mathbf{W}^{*} = \mathbf{S}_{XX}^{½}\mathbf{W}$, where $\mathbf{W}$ is the $m \times r$ matrix of X-canonical variates; and $m, p \le r$, where r is the rank of $\mathbf{R}_{XY}$. Here $\mathbf{E}$ is an $r \times r$ diagonal matrix with diagonal entries that are the canonical correlations $\mu_1 \ge \mu_2 \ge \ldots \ge \mu_r \ge 0$. The most common ways to factor $\mathbf{AB}'$ are to set $\alpha = 1,\ ½$ or 0.

The vectors on the biplot have the following interpretations. $\mathbf{S}_{YY}^{½}\mathbf{V}^{*}$ or equivalently, $\mathbf{S}_{YY}\mathbf{V}$, is the matrix of structure coefficients (correlations) of the Y-variables with the Y-canonical variates. Similarly, $\mathbf{S}_{XX}^{½}\mathbf{W}^{*}$ or $\mathbf{S}_{XX}\mathbf{W}$ is the matrix of structure coefficients of the X-variables on the X-canonical variates. If $\alpha = 1$, then $\mathbf{B} = \mathbf{S}_{YY}^{½}\mathbf{V}^{*}\mathbf{E}$ is the matrix of correlations between the Y-variables and the X-canonical variates. In the biplot display, the inner product of the vector corresponding to the i^{th} X-variable with the vector corresponding to the j^{th} Y-variable is a rank-two approximation of the $(i,j)^{th}$ element of $\mathbf{S}_{XY}$. This is because $\mathbf{S}_{XX}^{½}\mathbf{W}^{*}\mathbf{E}^{1-\alpha}\mathbf{E}^{1}\mathbf{V}^{*}\mathbf{S}_{YY}^{½} = \mathbf{S}_{XX}^{½}\mathbf{S}_{XX}^{-½}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-½}\mathbf{S}_{YY}^{½} = \mathbf{S}_{XY}$.

These biplots are optimal in the sense that one obtains an optimal rank-two approximation to $\mathbf{S}_{XY}$ which is invariant to linear transformations of **X** and **Y**. To see this, define the problem as finding **A** and **B**, each of rank-two, such that the expression below is minimized:

$$\left\|\mathbf{S}_{XX}^{-½}(\mathbf{S}_{XY} - \mathbf{AB}')\mathbf{S}_{YY}^{-½}\right\|^2, \tag{6.3}$$

where for a matrix **Z**, $\|\mathbf{Z}\|^2 = \text{trace}(\mathbf{Z}'\mathbf{Z})$. The solution to (6.3) is provided by the singular value decomposition of $\mathbf{S}_{XX}^{-½}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-½}$,

$$\mathbf{W}^*\mathbf{E}\mathbf{V}^{*\prime} = \mathbf{S}_{XX}^{-½}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-½},$$

where **A** and **B** are defined as in (6.2).

Ter Braak (1990) also gives an alternative version of the biplot for CCA. Instead of plotting **A** and **B** of (6.2), one plots

$$\mathbf{A} = \left[\mathbf{S}_{XX}^{½}\mathbf{W}^*\mathbf{E}^{\alpha-1}\right]_2 \quad \mathbf{B} = \left[\mathbf{S}_e^{½}\mathbf{V}^*\mathbf{E}^{\alpha}\right]_2 \tag{6.4}$$

where $\mathbf{S}_e$ is the residual sum of squares when the products matrix of **Y** with respect to **X** is subtracted from the total variation for **Y**. That is:

$$\mathbf{S}_e = \mathbf{S}_{YY} - \mathbf{S}_{XY}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}. \tag{6.5}$$

Ter Braak justifies this biplot in terms of approximating the matrix of regression coefficients in a multivariate regression of **Y** on **X**. However, it can be justified in the same way that (6.2) is justified, by finding an optimal approximation to $\mathbf{S}_{XY}$ that is invariant to non-singular transformations of the X-variables and the Y-variables. That is, one finds **A** and **B** such that

$$\left\|\mathbf{S}_{XX}^{-½\prime}(\mathbf{S}_{XY} - \mathbf{AB}')\mathbf{S}_e^{-½}\right\|^2, \tag{6.6}$$

is minimized.

When the X-variables are thought to cause the Y-variables in a regression sense then $\mathbf{S}_e$ is a more natural choice than $\mathbf{S}_{YY}$. For example, if the X-variables are group indicators then $\mathbf{S}_e$ becomes the within-groups covariance matrix. The biplot then indicates which groups are well discriminated, and which responses contribute to the discrimination.

6.2.2 Biplots for Redundancy Analysis

For an analogous biplot for redundancy analysis Ter Braak suggests finding **A** and **B** such that one has an optimal approximation to $\mathbf{S}_{XY}$ which is invariant only to non-singular transformations of the X-variables. This definition leads to **A** and **B** that are functions of the redundancy variables. Find **A** and **B** such that

$$\left\|\mathbf{S}_{XX}^{-½}(\mathbf{S}_{XY} - \mathbf{AB}')\right\|^2$$

is minimized. Which leads to

$$\mathbf{A} = \left[\mathbf{S}_{XX}^{½}\mathbf{W}^*\mathbf{E}^{\alpha-1}\right]_2 \qquad \mathbf{B} = \left[\mathbf{V}\mathbf{E}^{\alpha}\right]_2,$$

where $\mathbf{W}^*$ is an $m \times r$ orthonormal matrix such that $\mathbf{W}^* = \mathbf{S}_{XX}^{½}\mathbf{W}$, where $\mathbf{W}$ is the $m \times r$ matrix of canonical coefficients for the X-variables; $\mathbf{V}$ is a $p \times r$ orthonormal matrix of redundancy variates; and $\mathbf{E}$ is an $r \times r$ diagonal matrix whose j^{th} element is $(n-1)^{-1}$ times the root of the variation of the j^{th} Y-variate explained by the j^{th} X-variate, where n is the sample size.

The matrices of biplot vectors, $\mathbf{A}$ and $\mathbf{B}$, are functions of the redundancy weights. However, Ter Braak's interpretation of these weights is incorrect. He states (1990) correctly that with $\alpha = 1$ $\mathbf{A}$ is the matrix of structure coefficients. However, it is not correct that "The elements of $\mathbf{B}$ are not only correlations but also canonical coefficients". The correct interpretation is that $\mathbf{B} = [\mathbf{VE}]_2$ is a matrix whose $(i, j)^{th}$ element is $(n-1)^{-1}$ times the root of the variation of y_i explained by a simple linear regression on the j^{th} redundancy variate, $\mathbf{w}_j'\mathbf{x}$. To see this, notice that if one regresses y_i on $\mathbf{w}_j'\mathbf{x}$, the regression variance is $\mathbf{y}_i'\mathbf{X}\mathbf{w}_j(\mathbf{w}_j'\mathbf{X}'\mathbf{X}\mathbf{w}_j)^{-1}\mathbf{w}_j'\mathbf{X}'\mathbf{y}_i = \left(\mathbf{y}_i'\mathbf{X}\mathbf{w}_j\right)^2$, since $\mathbf{w}_j'\mathbf{X}'\mathbf{X}\mathbf{w}_j = 1$. Hence $\mathbf{y}_i'\mathbf{X}\mathbf{w}_j$ is the root of the regression variance. The matrix of root variances of the y_i regressed against the $\mathbf{w}_j'\mathbf{x}$ is $\mathbf{Y}'\mathbf{XW}$, and $\mathbf{Y}'\mathbf{XW} = (n-1)\mathbf{VEW}'\mathbf{W} = (n-1)\mathbf{VE}$.

An alternative factorization is worth mentioning: $\mathbf{A} = [\mathbf{S}_{XX}^{½}\mathbf{WE}]_2$ and $\mathbf{B} = [\mathbf{V}]_2$. Here $\mathbf{B}$ is just the matrix of weights of the redundancy variates for the Y-variables. $\mathbf{A}$ is a matrix whose elements are $(n-1)^{-1}$ times the root of the variation explained by each X-variable of each Y-variate.

6.2.3 Biplots for Procrustes Rotation

Biplots for PR which are analogous to those for CCA and RA can also be developed. One wishes to find $\mathbf{A}$ and $\mathbf{B}$ that minimize $\left\|(\mathbf{S}_{XY} - \mathbf{AB}')\right\|^2$. This method yields a biplot approximation to $\mathbf{S}_{XY}$ that is optimal in a least squares sense. Consistent with PR, this plot is invariant neither to non-singular transformations of the X-variables nor the Y-variables. As with the biplots for CCA and RA, the matrices of vectors $\mathbf{A}$ and $\mathbf{B}$ are interpretable in terms of a PR analysis. $\mathbf{A} = [\mathbf{W}]_2$ is a matrix of coefficients for the X-variates. $\mathbf{B} = [\mathbf{VE}]_2$ is a matrix showing the covariance of each Y-variable with the Procrustes rotation variates, as $\mathbf{W}'\mathbf{S}_{XY}\mathbf{I} = \mathbf{W}'\mathbf{WEV} = \mathbf{EV}$.

The biplot for PR can be interpreted as showing which X-variables covary with which Y-variables. It also can be interpreted as showing which variables are fit well in the Procrustes rotation. That is, it shows which X-variables match the pattern of the Y-variables when rotated.

In summary, there are biplots for displaying CCA, RA and PR that are optimal in a weighted least squares sense and which yield markers related to the estimated model parameters.

6.3 JOINT PLOTS

Chapter Five extended the CCA, RA and PR models to three-mode data, that is, multiple occasions and multiple datasets. Analogously, this section extends biplots for CCA, RA and PR to three-mode data. The biplots of Section **6.2** were based on singular value decompositions of a matrix. Joint plots (Kroonenberg 1983) are based on the Tucker2 or PARAFAC (orth.) decomposition of a three-mode array.

Denote an $m \times p \times g$ three-mode array as $\underline{\mathbf{C}}$. To examine the relationship between the component weights of two modes one can make joint plots. One possibility is to create a series of joint plots, one joint plot for each component of the non-examined mode; i.e., one for each occasion. Another possibility is to make one averaged joint plot. Each joint plot is analogous to a biplot. If one wants to examine the relationship between the subject and variable modes, then what is plotted is based on either $\mathbf{G}\mathbf{C}_k\mathbf{H}'$, $k = 1,\ldots,g$, or $\mathbf{G}\overline{\mathbf{C}}\mathbf{H}'$, where $\mathbf{G}$ is the matrix of components for the subject mode, $\mathbf{H}$ is the matrix of components for the variables mode, $\mathbf{C}_k$ is the k^{th} slice of the core box, that is the matrix $\underline{\mathbf{C}}[,,k]$, and $\overline{\mathbf{C}}$ is the average of the $\mathbf{C}_k$. As $\mathbf{C}_k$ and $\overline{\mathbf{C}}$ are generally not diagonal, a SVD is performed on $\mathbf{C}_k$ to factor it. So what is plotted is $\mathbf{A}_k$ and $\mathbf{B}_k$ where:

$$\mathbf{A}_k\mathbf{B}_k' = \mathbf{G}\mathbf{C}_k\mathbf{H}' = \mathbf{G}(\mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{V}_k')\mathbf{H}' = \left(\tfrac{m}{p}\right)^{1/4}(\mathbf{G}\mathbf{U}_k\boldsymbol{\Lambda}_k^{\alpha})(\boldsymbol{\Lambda}_k^{1-\alpha}\mathbf{V}_k'\mathbf{H}')\left(\tfrac{p}{m}\right)^{1/4},$$

or $\mathbf{A}_k = \left(\tfrac{m}{p}\right)^{1/4}(\mathbf{G}\mathbf{U}_k\boldsymbol{\Lambda}_k^{\alpha})$ and $\mathbf{B}_k = (\boldsymbol{\Lambda}_k^{1-\alpha}\mathbf{V}_k'\mathbf{H}')\left(\tfrac{p}{m}\right)^{1/4}$, where m is the number of measurements in the subjects mode and p the number measurements in the variables mode, $\mathbf{G}$ and $\mathbf{H}$ are defined to be the two specific pairs of components, thus $\mathbf{G}$ is an $m \times 2$ matrix and $\mathbf{H}$ is a $p \times 2$ matrix; and $\mathbf{C}_k$ and $\overline{\mathbf{C}}$ are the 2×2 submatrices of the core matrices corresponding to the selected components. One typically chooses the two components for $\mathbf{G}$ and $\mathbf{H}$ with the largest associated core elements. However, one can make a joint plot between any two components.

Further details: α is chosen typically to be 1/2. The weightings $\left(\tfrac{m}{p}\right)^{1/4}$ and $\left(\tfrac{p}{m}\right)^{1/4}$ put the distances from the origin of the subject vectors on the same scale as those of the variable vectors (Kroonenberg 1983). This makes the plots easier to view, though one may choose other weightings (see Section **6.2**, in particular (6.1), for a discussion on the weighting of biplots).

6.3.1 Joint Plots for Canonical Correlation Analysis

In this section I extend the biplots of structure coefficients for CCA at one occasion to joint plots of structure coefficients for multiple occasions. Here one models $\mathbf{S}_{XY}$ over the third mode. The assumptions necessary for this particular biplot are the same as for CCA/third, that $\mathbf{S}_{XX}$ is constant over the third mode, and that either $\mathbf{S}_{YY}$ or $\mathbf{S}_e$ (6.5) is also constant over the third mode (see Section **5.5.3** for a discussion of these assumptions). The subsequent developments apply to both $\mathbf{S}_e$ and $\mathbf{S}_{YY}$, although I use $\mathbf{S}_{YY}$.

Plot $\mathbf{A}_k$ and $\mathbf{B}_k$ for each occasion k, where $\mathbf{A}_k$ and $\mathbf{B}_k$:

$$\mathbf{A}_k = \mathbf{S}_{XX}^{1/2}\mathbf{W}^*\mathbf{U}_k\mathbf{\Lambda}_k^{1/2}, \quad \mathbf{B}_k = \mathbf{S}_{YY}^{1/2}\mathbf{V}^*\mathbf{Z}_k\mathbf{\Lambda}_k^{1/2},$$

where $\mathbf{V}^*$ is a $p \times 2$ orthonormal matrix such that $\mathbf{V}^* = \mathbf{S}_{YY}^{1/2}[\mathbf{V}]_2$, where $\mathbf{V}$ is the $p \times r$ matrix of Y-canonical variates from the CCA/third model; $\mathbf{W}^*$ is a $m \times 2$ orthonormal matrix such that $\mathbf{W}^* = \mathbf{S}_{XX}^{1/2}[\mathbf{W}]_2$, where $\mathbf{W}$ is the $m \times q$ matrix of X-canonical variates from the CCA/third model; and $\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{Z}_k' = \mathbf{C}_k$ is the singular value decomposition of $\mathbf{C}_k$, the k^{th} core matrix from the CCA/third model. An alternative is to plot $\mathbf{A} = \mathbf{S}_{XX}^{1/2}\mathbf{W}^*\mathbf{U}\mathbf{\Lambda}^{1/2}$ and $\mathbf{B} = \mathbf{S}_{YY}^{1/2}\mathbf{V}^*\mathbf{Z}\mathbf{\Lambda}^{1/2}$, where $\mathbf{U}\mathbf{\Lambda}\mathbf{Z}' = \overline{\mathbf{C}}$.

Note that the $\mathbf{S}_{XX}^{1/2}\mathbf{W}^*$ and the $\mathbf{S}_{YY}^{1/2}\mathbf{V}^*$ are matrices of structure coefficients.

Next I discuss the sense in which these joint plots are optimal. The function one minimizes is the least squares lack of fit of a rank-two approximation to the matrix of covariances between the X-variables and the Y-variables, $\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k$, $k = 1, \ldots, g$. The loss function should be invariant to non-singular transformations of the X-variables and the Y-variables, as is CCA. Further, $\mathbf{A}_k$ should span the same space for $k = 1, \ldots, g$, as should $\mathbf{B}_k$. Thus the problem reduces to finding $\mathbf{A}_k$ and $\mathbf{B}_k$ of rank-two such that

$$\sum_{k=1}^{g} \left\| \mathbf{S}_{XX}^{-1/2}\left(\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k - \mathbf{A}_k\mathbf{B}_k' \right)\mathbf{S}_{YY}^{-1/2} \right\|^2$$

is minimized. Clearly the optimal solution is given by a two-component Tucker2 decomposition of $\frac{1}{n_k - 1}\mathbf{S}_{XX}^{-1/2}\mathbf{X}_k'\mathbf{Y}_k\mathbf{S}_{YY}^{-1/2}$, $k = 1, \ldots, g$. That is

$$\mathbf{W}^*\mathbf{C}_k\mathbf{V}^{*\prime} = \frac{1}{n_k - 1}\mathbf{S}_{XX}^{-1/2}\mathbf{X}_k'\mathbf{Y}_k\mathbf{S}_{YY}^{-1/2},$$

for $k = 1, \ldots, g$. Thus $\mathbf{W}^*\mathbf{C}_k\mathbf{V}^{*\prime} = \mathbf{S}_{XX}^{-1/2}\mathbf{A}_k\mathbf{B}_k'\mathbf{S}_{YY}^{-1/2}$ and

$$\mathbf{A}_k = \mathbf{S}_{XX}^{1/2}\mathbf{W}^*\mathbf{U}_k\mathbf{\Lambda}_k^{\alpha}, \quad \mathbf{B}_k = \mathbf{S}_{YY}^{1/2}\mathbf{V}^*\mathbf{Z}_k\mathbf{\Lambda}_k^{1-\alpha}, \tag{6.7}$$

where one performs a singular value decomposition on $\mathbf{C}_k$ to get $\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{Z}_k' = \mathbf{C}_k$.

If one restricts the columns of $\mathbf{A}_k$ to be proportional over $k = 1, \ldots, g$, i.e., $\mathbf{A}_c[, i] = f\mathbf{A}_d[, i]$, for $c \neq d$, and likewise makes the same restriction for $\mathbf{B}_k$, then one can use an argument similar to the one above to show that the optimal joint plots are based on the PARAFAC (orth.) solution. Such joint plots are easier to interpret as the axes in the joint plots correspond to the two pairs of components. This is so because the core matrices are diagonal, and thus the matrices of structure coefficients are not rotated (by $\mathbf{U}_k$ or $\mathbf{Z}_k$ in (6.7)).

6.3.2 Joint Plots for Redundancy Analysis

The biplot for RA, which plotted structure coefficients and redundancy variates, can be extended to joint plots for multimode data in a manner analogous to how the biplot for CCA was extended to joint plots for CCA. The multimode extension of the biplot for RA is

$$\mathbf{A}_k = \mathbf{S}_{XX}^{1/2}\mathbf{W}^*\mathbf{U}_k\mathbf{\Lambda}_k^{1/2}, \quad \mathbf{B}_k = \mathbf{V}^*\mathbf{Z}_k\mathbf{\Lambda}_k^{1/2}, \tag{6.8}$$

where $\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{Z}_k = \mathbf{C}_k$ is the singular value decomposition of the k^{th} core matrix $\mathbf{C}_k$. $\mathbf{W}^*$ is an $m \times 2$ orthonormal matrix such that, $\mathbf{W}^* = \mathbf{S}_{XX}^{1/2}[\mathbf{W}]_2$, where $\mathbf{W}$ is the $m \times q$ matrix of canonical coefficients for the X-variables; and $\mathbf{V}^* = [\mathbf{V}]_2$, where $\mathbf{V}$ is a $p \times r$ orthonormal matrix of redundancy variates. One can also plot $\mathbf{A} = \mathbf{S}_{XX}^{1/2}\mathbf{W}^*\mathbf{U}\mathbf{\Lambda}^{1/2}$ and $\mathbf{B} = \mathbf{V}^*\mathbf{Z}\mathbf{\Lambda}^{1/2}$, where $\mathbf{U}\mathbf{\Lambda}\mathbf{Z}' = \overline{\mathbf{C}}$.

Notice that $\mathbf{S}_{XX}^{1/2}\mathbf{W}^*$ is the matrix of structure coefficients and $\mathbf{V}$ is the matrix of redundancy coefficients.

These biplots are optimal in that they yield a display approximating the matrices $\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k$, $k = 1,\ldots,g$, based on a loss function that is invariant to non-singular linear transformations of the X-variables, as is RA. Further, $\mathbf{A}_k$ should span the same space for $k = 1,\ldots,g$, as should $\mathbf{B}_k$. The problem reduces to finding $\mathbf{A}_k$ and $\mathbf{B}_k$ of rank-two such that

$$\sum_{k=1}^{g}\left\|\mathbf{S}_{XX}^{-1/2}\left(\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k - \mathbf{A}_k\mathbf{B}_k'\right)\right\|^2$$

is minimized. Clearly the optimal solution is given by a two-component Tucker2 decomposition of $\frac{1}{n_k - 1}\mathbf{S}_{XX}^{-1/2}\mathbf{X}_k'\mathbf{Y}_k$, $k = 1,\ldots,g$. That is $\mathbf{W}^*\mathbf{C}_k\mathbf{V}^{*'} = \frac{1}{n_k - 1}\mathbf{S}_{XX}^{-1/2}\mathbf{X}_k'\mathbf{Y}_k$. Thus $\mathbf{W}^*\mathbf{C}_k\mathbf{V}^{*'} = \mathbf{S}_{XX}^{-1/2}\mathbf{A}_k'\mathbf{B}_k$ and $\mathbf{A}_k = \mathbf{S}_{XX}^{1/2}\mathbf{W}^*\mathbf{U}_k\mathbf{\Lambda}_k^{\alpha}$, $\mathbf{B}_k = \mathbf{V}^*\mathbf{Z}_k\mathbf{\Lambda}_k^{1-\alpha}$, where $\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{Z}_k' = \mathbf{C}_k$.

If one restricts the columns of $\mathbf{A}_k$ to be proportional over $k = 1,\ldots,g$, i.e. $\mathbf{A}_c[,i] = f\mathbf{A}_d[,i]$, for $c \neq d$, and likewise makes the same restriction for $\mathbf{B}_k$, then one can use an argument similar to the one above to show that the optimal joint plots are based on the PARAFAC (orth.) solution. Such joint plots are easier to interpret as the axes in the joint plots correspond to the two pairs of components. This is so because the core matrices are diagonal, and thus the matrices of structure coefficients are not rotated (by $\mathbf{U}_k$ or $\mathbf{Z}_k$ in (6.8)).

Example 6.1

Figure 6.1 shows the joint plot for the Shenendoah data from Section **5.4**. The plot is based on the RA/third solution for the first two components of the PARAFAC (orth.) model. Since for this example the plots are roughly similar across time, instead of showing the joint plots for each occasion, the plot based on the sum of the core matrices is shown. This gives a general picture of how the two modes, geological variables and streamwater variables, relate.

The vectors corresponding to the nine geological variables are numbered one through nine. Recall that the first five geological variables refer to bedrock types, ab2400 refers to

altitude, DD refers to drainage density, E/W refers to east or west slope, and Dev. refers to the presence of development. The vectors corresponding to the fourteen streamwater variables are lettered from a to p, skipping l and o; they are more self-explanatory. The origin is labeled with a zero. The key to the numbering and lettering is given in **Figure 6.2**.

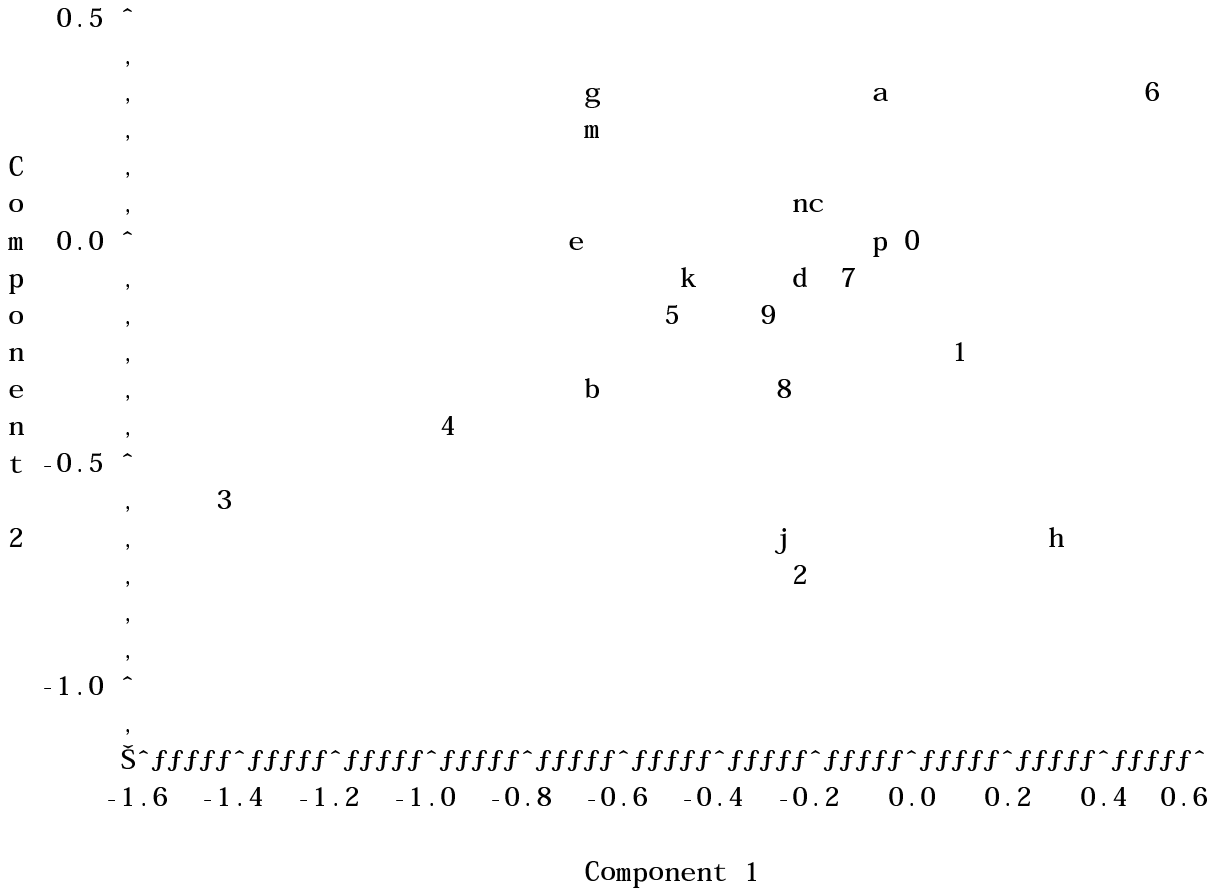

Figure 6.1 Joint Plot for the Sum of Core Matrices for PARAFAC (orth.)

a	discharge	h	K^+	1	Antietam
b	conductivity	i	alkalinity	2	Hampton
c	pH	j	$SO_4^=$	3	Catoctin
d	temperature	k	Cl^-	4	Pedlar
e	Ca^{++}	m	$SiO_4^=$	5	Old Rag
f	Mg^{++}	n	NO_3^-	6	ab2400
g	Na^+	p	NH_4^+	7	DD
				8	E/W
				9	Dev.

Figure 6.2 Key to Symbols

The interpretation of the joint plot is aided by consideration of what the components are and how the vectors relate. The first axis corresponds to the first geological and streamwater variables component pair, which relates to carbonic acid weathering. The second axis corresponds to the second component pair, which relates to acidification. Next, consider that since the geological variables are standardized, the $(\mathrm{i},\mathrm{j})^{\mathrm{th}}$ element of $\mathbf{S}_{\mathrm{XY}}$ for a given occasion shows $(n-1)^{-1}$ times the root of the variance of the j^{th} Y-variable explained in a simple linear regression on the i^{th} X-variable. Consequently, the inner product between a geological variable's vector and a streamwater variable's vector yields an approximation to $(n-1)^{-1}$ times the root of the regression variance of the streamwater variable predicted by the geological variable. This is in an averaged sense since the joint plot is based on an averaged core matrix. Thus if a geological and streamwater variable are near each other on the plot, such as Hampton bedrock (2) and $SO_4^{=}$ (j), then the geological variable is a strong predictor of the streamwater variable in a simple linear regression.

The advantage of the joint plot is that it lets one view all of the variables in relation to each other. Geological variables which are located near one another are similar in the variation of the streamwater variables they explain. For example, Catoctin (3) and Pedlar (4) bedrock types explain the streamwater variables similarly, though Catoctin (3) has a stronger effect since it is further from the origin. Likewise, streamwater variables which are near to each other are similar in that they are explained by the same geological variables. For example, the sodium (g), silica (m) and calcium (e) ions are near each other. They result from the same processes, and thus are predicted by the same geological variates. Important also is the distance from the origin. Geological variables that are near the origin, such as drainage density (7), are not strong predictors in the rank-two RA/time model upon which the joint plot is based. Similarly streamwater variables near the origin, such as ammonium (p), are not well explained by the rank-two RA/time model.

A limitation of these joint plots is that they only display a two-components solution. Thus the effects of possible lower order components are not seen. For example, ammonium and drainage density are both related to the fourth pair of components in the analysis in Section **5.4**.

6.3.3 Joint Plots for Procrustes Rotation

The biplot for PR can be extended to joint plots for multimode data in a manner analogous to how the biplots for CCA and RA were extended to joint plots. The multimode extension of the biplot for PR is

$$\mathbf{A}_{\mathrm{k}} = \mathbf{W}\mathbf{U}_{\mathrm{k}}\boldsymbol{\Lambda}_{\mathrm{k}}^{1/2}, \quad \mathbf{B}_{\mathrm{k}} = \mathbf{V}\mathbf{Z}_{\mathrm{k}}\boldsymbol{\Lambda}_{\mathrm{k}}^{1/2}, \tag{6.9}$$

where $\mathbf{W}^{*} = [\mathbf{W}]_2$, where $\mathbf{W}$ is the $m \times q$ matrix of PR coefficients for the X-variables; $\mathbf{V}^{*} = [\mathbf{V}]_2$, where $\mathbf{V}$ is a $p \times r$ orthonormal matrix of PR variates; and where $\mathbf{U}_{\mathrm{k}}\boldsymbol{\Lambda}_{\mathrm{k}}\mathbf{Z}_{\mathrm{k}} = \mathbf{C}_{\mathrm{k}}$ is the singular value decomposition of the k^{th} core matrix $\mathbf{C}_{\mathrm{k}}$. One can also plot $\mathbf{A} = \mathbf{W}\mathbf{U}\boldsymbol{\Lambda}^{1/2}$ and $\mathbf{B} = \mathbf{V}\mathbf{Z}\boldsymbol{\Lambda}^{1/2}$, where $\mathbf{U}\boldsymbol{\Lambda}\mathbf{Z}' = \overline{\mathbf{C}}$.

Note that $\mathbf{W}$ is the $m \times r$ matrix of PR coefficients for the X-variables and $\mathbf{V}$ is the $p \times r$ matrix of PR coefficients for the Y-variables.

These biplots are optimal in that they yield a display approximating the matrix $\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k$. Consistent with PR, the solution is not invariant linear transformations of the variables. $\mathbf{A}_k$ should span the same space, as should the $\mathbf{B}_k$. The problem reduces to finding $\mathbf{A}_k$ and $\mathbf{B}_k$ of rank-two such that

$$\sum_{k=1}^{g} \left\| \left(\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k - \mathbf{A}_k\mathbf{B}_k' \right) \right\|^2$$

is minimized. Clearly the optimal solution is given by a Tucker2 decomposition of $\frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k$, $k = 1, \ldots, g$. That is $\mathbf{W}\mathbf{C}_k\mathbf{V}' = \frac{1}{n_k - 1}\mathbf{X}_k'\mathbf{Y}_k$. Thus $\mathbf{W}\mathbf{C}_k\mathbf{V}' = \mathbf{A}_k'\mathbf{B}_k$ and $\mathbf{A}_k = \mathbf{W}\mathbf{U}_k\Lambda_k^{\alpha}$, $\mathbf{B}_k = \mathbf{V}\mathbf{Z}_k\Lambda_k^{1-\alpha}$, where $\mathbf{U}_k\Lambda_k\mathbf{Z}_k' = \mathbf{C}_k$.

If one restricts the columns of $\mathbf{A}_k$ to be proportional over $k = 1, \ldots, g$, i.e. $\mathbf{A}_c[,i] = f\mathbf{A}_d[,i]$, for $c \neq d$, and likewise makes the same restriction for $\mathbf{B}_k$, then one can use an argument similar to the one above to show that the optimal joint plots are based on the PARAFAC (orth.) solution. Such joint plots are easier to interpret as the axes in the joint plots correspond to the two pairs of components. This is so because the core matrices are diagonal, and thus the matrices of structure coefficients are not rotated (by $\mathbf{U}_k$ or $\mathbf{Z}_k$ in (6.9)).

6.4 PLOTS OF THE COMPONENT SCORES

The score on a component for a given subject is generally defined as a weighted sum of the variables, the weights being component weights. Define the $m \times 1$ vector of scores for subjects on the b^{th} variable component at the k^{th} occasion, denoted by $\mathbf{Q}_k[,b]$, as

$$\mathbf{Q}_k[,b] = \mathbf{D}_k\mathbf{H}[,b].$$

An equivalent form is $\mathbf{Q}_k[,b] = \mathbf{G}\mathbf{C}_k[,b]$. One could also define scores for the variables on the subject components.

In some applications it may be useful to inspect the scores of all combinations of the elements of two modes on the components of the third mode. For instance, for longitudinal data the scores of each subject-time combination on the variable components can be used to inspect the development of a subject's score on the variable components over time (Kroonenberg 1983). Component scores serve as an intermediate level of condensation between the raw data and the three-mode model.

Example 6.2

In **Figures 6.3**, **6.4**, **6.5** and **6.6** I present the scores of the geological variables at the six occasions on each of the four streamwater components. These scores are based on the Tucker2 solution with four geological and four streamwater variates, as given in Section **5.4**.

The scores give a picture of how the geological variables relate to the streamwater components. I start by making two general observations on the plots. First, one notices that in the plots for the first three components (**Figures 6.3**, **6.4** and **6.5**) is that the relative positions of the geological variables are steady, even roughly proportional, over time. This is a reflection of the fact that the off-diagonal elements of the core matrices are near zero. Indeed, if one plotted scores based on the PARAFAC (orth.) model the positions of the scores would be exactly proportional over time. Only for the fourth variate does this not hold true. If one looks at the core matrix in **Table 5.3** for the March 1982 one sees a large off-diagonal element (1.25) between the fourth streamwater variate and first geological variate. Second, note that the range of the scores narrows for the subsequent components. This is because each component accounts for less of the variation than the previous.

I will just point out a few notable details about the score plots to provide a sense for how one interprets them. First, recall from Section **5.4** that the first streamwater component is interpreted to be related to the results of the weathering due to carbonic acid, with heavy weightings for, silica, alkalinity and the alkaline ions. **Figure 6.3** shows where the geological variables measure on this component over time. For example one sees that altitude (6) has a generally low score, but is particularly low at the first occasion, August 1981. Also, note in **Figure 6.5** that drainage density (7) has a high score on the third component at the fourth occasion (March 1982). This component was related in Section **5.4** to high silica and sodium, the results of plagioclastic weathering. Drainage density explains a relatively good deal of the variation of this component and of the associated streamwater variables at this occasion.

The key to the numbering is shown in Figure **6.2**. Note that some numbers may be obscured by others.

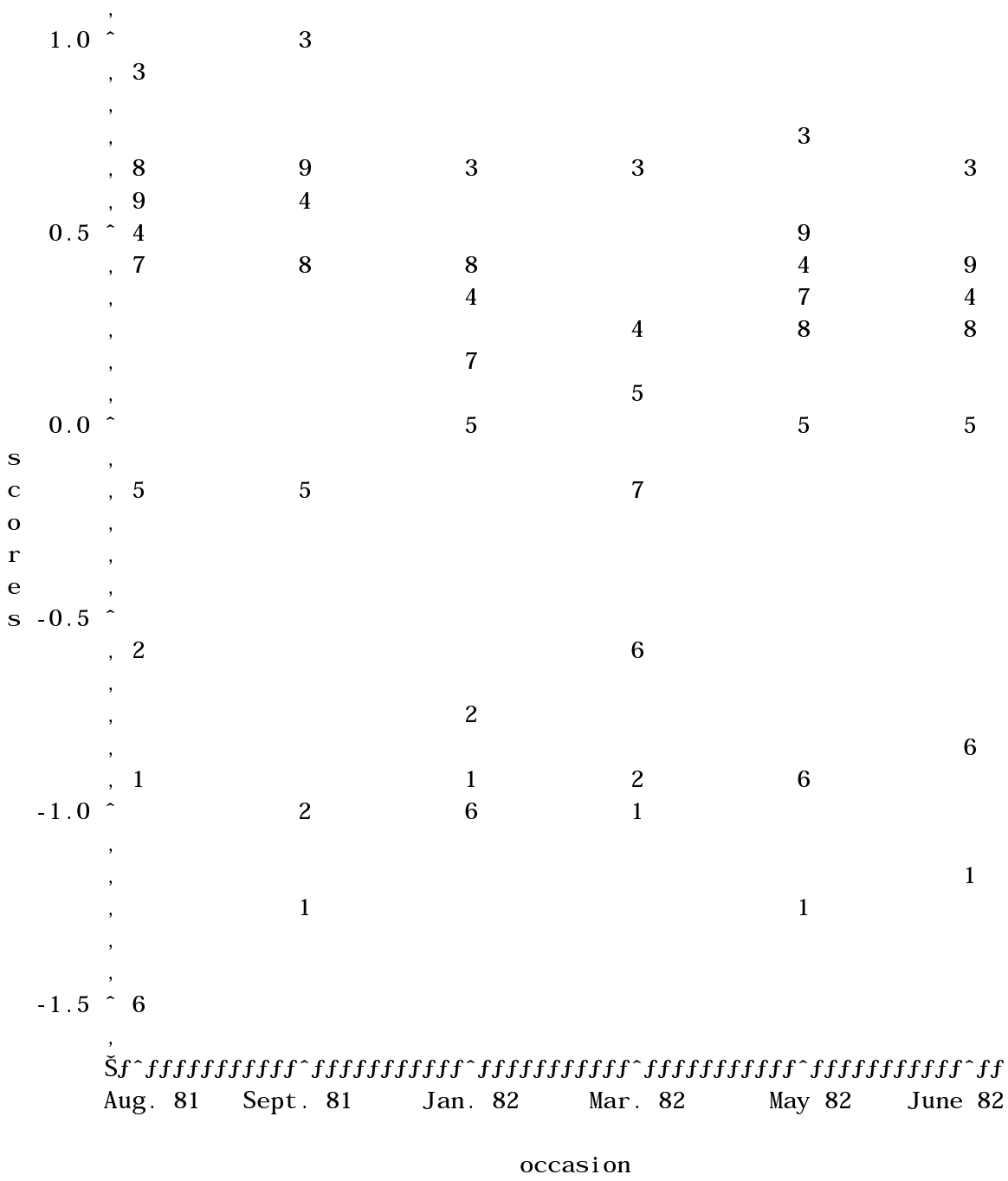

Figure 6.3 Scores on the First Streamwater Variate

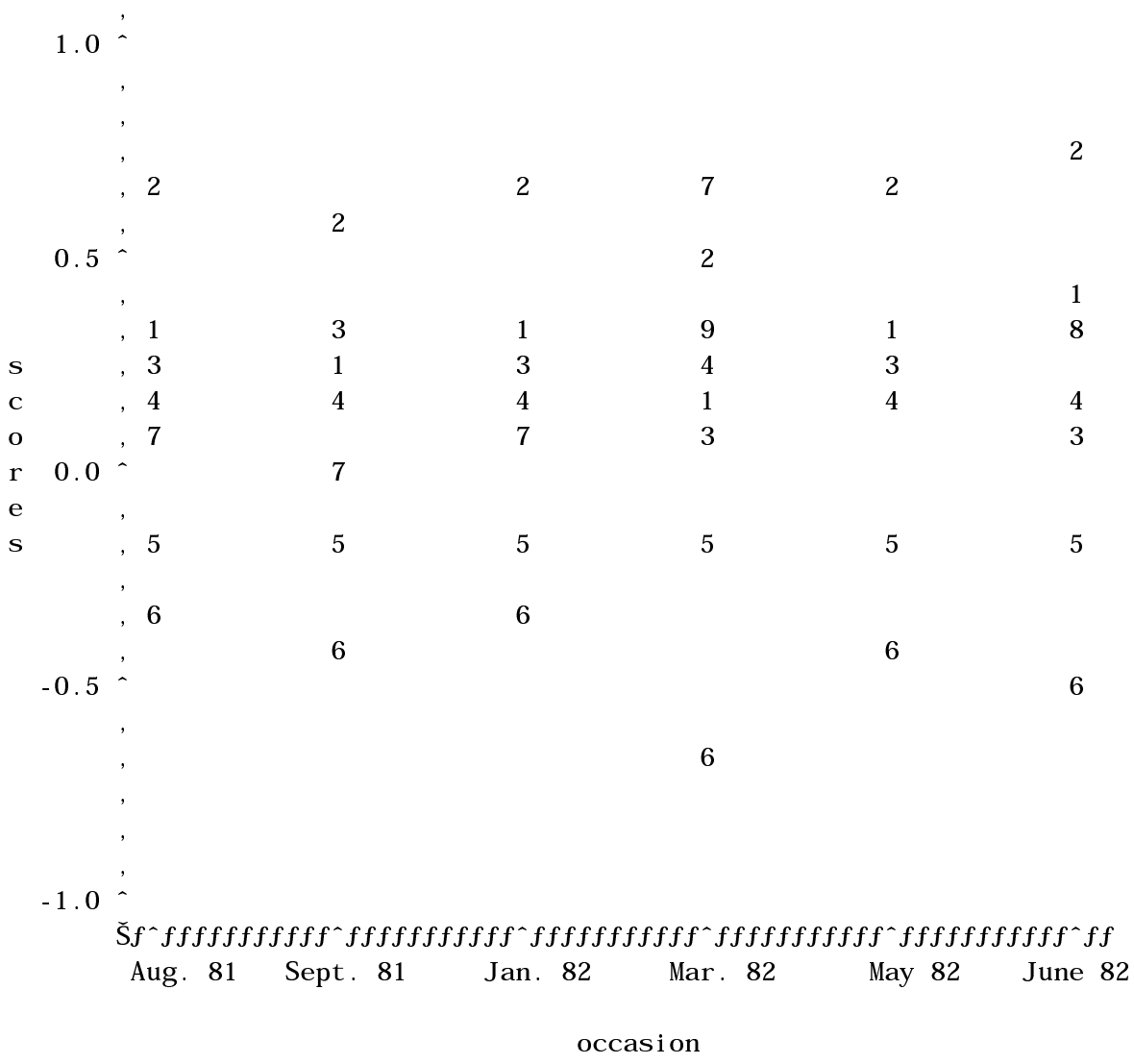

Figure 6.4 Scores on the Second Streamwater Variate

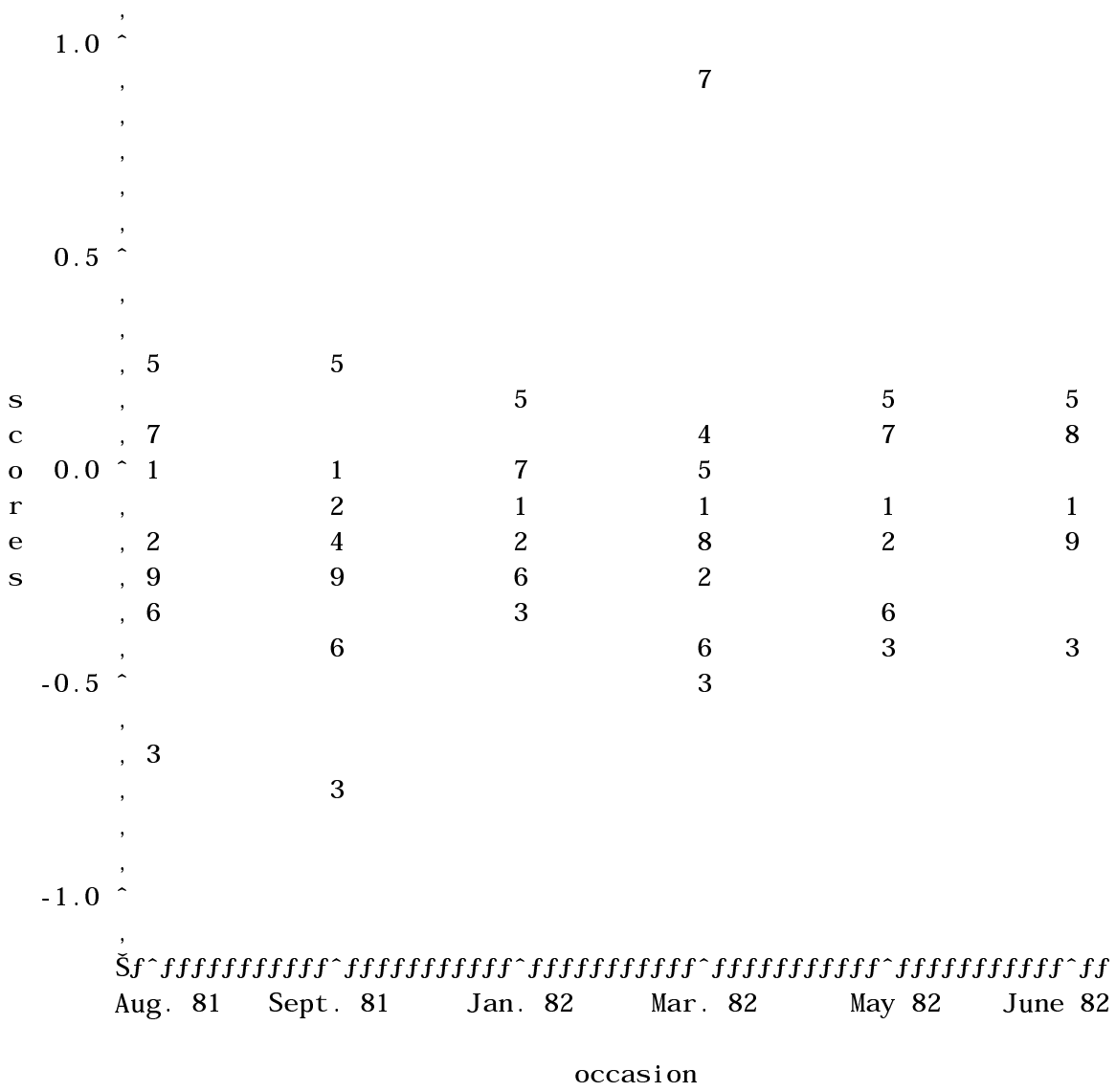

Figure 6.5 Scores on the Third Streamwater Variate

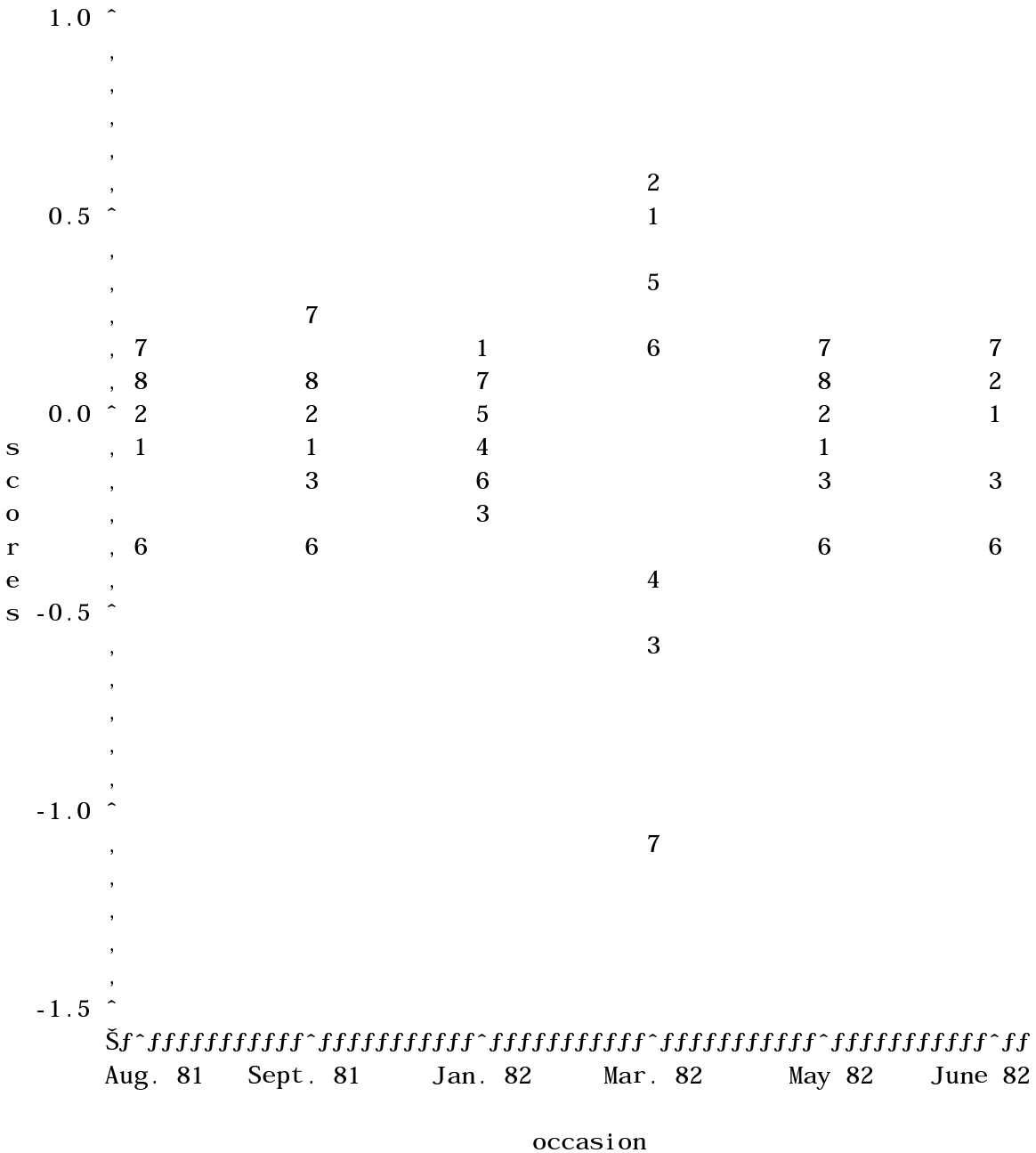

Figure 6.6 Scores on the Fourth Streamwater Variate

6.5 RESIDUAL PLOTS

Kroonenberg (1983) recommends sums of squares plots to assess the quality of fit of the elements of a mode. For each variable (or subject) one plots its sums of squares fit (sums of squares explained) against its sums of squares residuals (sums of squares lack of fit).

Example 6.3

Below in **Figure 6.7** is the residual plot for the Shenendoah data based on the estimates for the four component PARAFAC (orth.) model from Section **5.5**. Plotted are the sums of squares residuals versus sums of squares fit for the fourteen variables. Note that the total sums of squares for each variable is the total of the sums of squares explained by a separate multiple regression at each occasion.

In this plot one sees the relationship variance of variables temperature (d), discharge (a) and nitrate (n) are modeled relatively poorly by the three-mode model. If one recalls the analysis of Section **5.5**, these were variables that did not participate in any of the processes attributed to the four variate pairs. On the other hand, some streamwater variables are modeled well by the three-mode models, such as alkalinity.

The key relating the numbers to the variables is given in **Figure 6.8**.

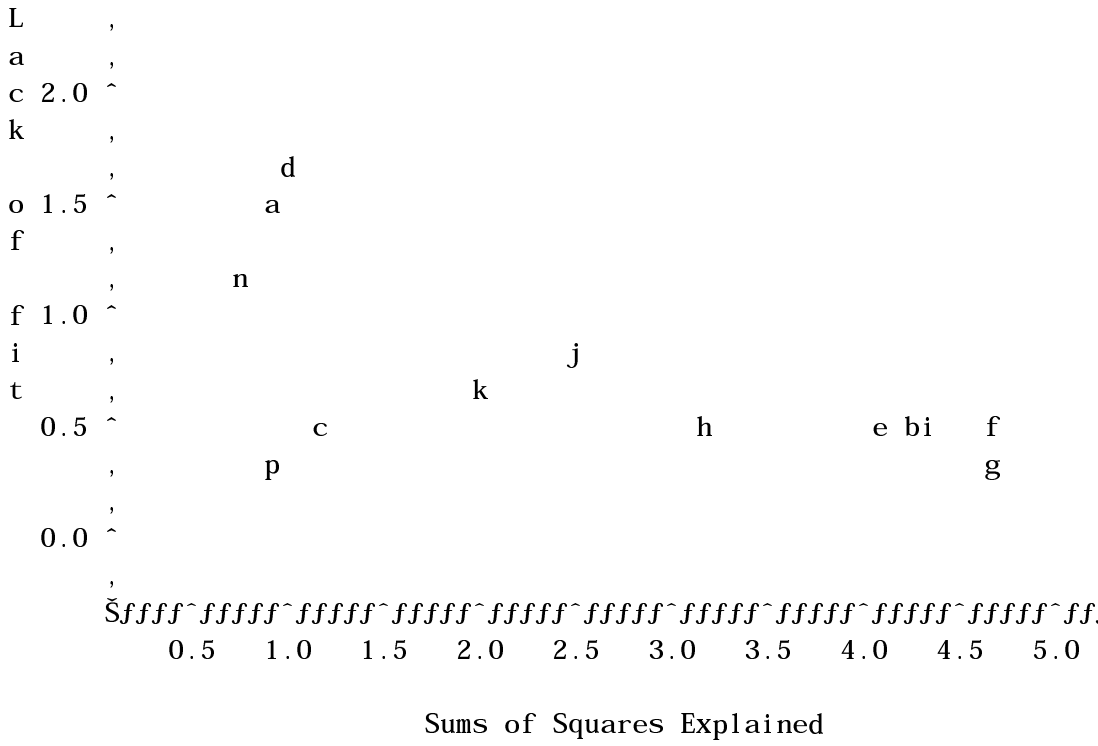

Figure 6.7 Residual Plot for the Sums of Squares Explainable of the Streamwater Variables

a	discharge	h	K^+
b	conductivity	i	alkalinity
c	pH	j	$SO_4^=$
d	temperature	k	Cl^-
e	Ca^{++}	m	$SiO_4^=$
f	Mg^{++}	n	NO_3^-
g	Na^+	p	NH_4^+

Figure 6.8 Key to Symbols

6.6 SUMMARY

In this chapter I showed how to use graphical displays to aid in the analysis of the relationship between two sets of variables over time. These displays were related to the CCA, RA, or PR parameters of the three-mode models of Chapter Six. The joint plots showed how the

variables from the two sets related to each other. The component plots were useful for showing the interactions between variables and components over time. In all, these plots allow the researcher to get a quick, visual appreciation of the relationships between many variables at different occasions.

CHAPTER SEVEN

COVARIANCE STRUCTURE ANALYSIS

7.1 INTRODUCTION

In this chapter I model canonical variate analysis (CVA) with longitudinal data using covariance structure analysis (COSAN) (McDonald 1978, 1980). If one assumes common canonical variates, then multivariate data with group structure imply a certain covariance structure. COSAN models this implied structure. An advantage of modeling with COSAN is that SAS software exists for analyzing it, obviating the need to program new algorithms.

I begin Chapter Seven with a description of the COSAN model in Section **7.2**. In Section **7.3** I express CVA with longitudinal data as a covariance structure. In Section **7.4** I show how CVA over time is parameterized in the COSAN framework. Lastly, in Section **7.5** I show a limited example based on the Shenandoah study previously described in Section **5.4.**

7.2 COVARIANCE STRUCTURE ANALYSIS

Covariance Structure Analysis (McDonald 1978, 1980, SAS 1990) is a model for analyzing positive definite or semidefinite matrices. Most commonly known models for analyzing covariance structures can be presented as special cases of COSAN, such as principal components analysis, confirmatory and exploratory factor analysis, and LISREL (Linear Structural RELations, Jöreskog 1989).

The general form of the COSAN model is

$$\mathbf{C} = \mathbf{F}_1\mathbf{P}_1\mathbf{F}_1' + \cdots + \mathbf{F}_m\mathbf{P}_m\mathbf{F}_m' \tag{7.1}$$

where $\mathbf{C}$ is a symmetric positive definite or semi-definite matrix; each $\mathbf{F}_k$, $k = 1,\ldots,m$, is the product of s_k matrices, $\mathbf{F}_k = \mathbf{F}_{k1}\ldots\mathbf{F}_{ks_k}$; and each matrix $\mathbf{P}_k$ is symmetric. The matrices $\mathbf{P}_k$ can be of the form of an inverse of a matrix $\mathbf{H}_k$, that is, $\mathbf{P}_k = \mathbf{H}_k^{-1}$. The matrices $\mathbf{F}_{ki}$ above can be of the form of an inverse of a matrix $\mathbf{H}_{ki}$, or of the inverse of an identity matrix minus $\mathbf{H}_{ki}$; that is, $\mathbf{F}_{ki}$ can be of the form $\mathbf{F}_{ki} = \mathbf{H}_{ki}^{-1}$ or $\mathbf{F}_{ki} = (\mathbf{I} - \mathbf{H}_{ki})^{-1}$. Furthermore, a matrix can contain both parameter terms and constant terms. A parameter can be constrained to be a function of other parameters. Hence COSAN is a flexible model for analyzing covariance structures.

The general idea behind COSAN is that covariance structures can often be modeled as the crossproduct of matrices whose columns consist of the weights of components or the loadings of factors. Consider for example the factor analysis model, $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}$. $\mathbf{\Sigma}$ is modeled as the crossproduct matrix of $\mathbf{L}$, the matrix of factor loadings, (plus a diagonal matrix of specific variances $\mathbf{\Psi}$). The factor analysis model is expressed in COSAN as $\mathbf{C} = \mathbf{F}_1\mathbf{P}_1\mathbf{F}_1' + \mathbf{F}_2\mathbf{P}_2\mathbf{F}_2'$, where $\mathbf{F}_1$ is the matrix of factor loadings, $\mathbf{P}_1$ and $\mathbf{F}_2$ are restricted to be identity matrices, and $\mathbf{P}_2$ is restricted to be a diagonal matrix of specific variances. One can incorporate more complexity to get LISREL models by modeling the $\mathbf{F}_k$ to be the product of matrices $\mathbf{F}_k = \mathbf{F}_{k1}\ldots\mathbf{F}_{ks_k}$, and where appropriate by constraining these $\mathbf{F}_{ki}$ to be either the inverse of a matrix of parameter and constants, or the identity matrix minus the inverse of a matrix of parameter and constants (see Jöreskog 1989).

In COSAN restricting a matrix to be orthogonal is done indirectly with the Cayley (McDonald 1978) decomposition. For example, if one wants to restrict a matrix of parameters and constants $\mathbf{L}$ to be orthogonal, then set $\mathbf{L} = (\mathbf{I} - \mathbf{H}')^{-1}(\mathbf{I} - \mathbf{H})$, where $\mathbf{H}$ is skew symmetric with zeros as its diagonal elements. (Skew symmetric means $\mathbf{H}$ is parameterized such that $\mathbf{H} = -\mathbf{H}'$).

For an example of modeling orthogonal matrices consider the principal components model. This is parameterized as

$$\mathbf{C} = (\mathbf{I} - \mathbf{H}')^{-1}(\mathbf{I} - \mathbf{H})\mathbf{P}(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H}')^{-1'}.$$

$\mathbf{P}$ is a $p \times p$ diagonal matrix whose elements are the squares of the eigenvalues associated with the principal components. The $p \times p$ orthogonal matrix of principal components $\mathbf{V}$ is found as $\mathbf{V} = (\mathbf{I} - \mathbf{H}')^{-1}(\mathbf{I} - \mathbf{H})$, where $\mathbf{H}$ is a $p \times p$ skew symmetric matrix. In terms of the model given

in (7.1), $\mathbf{C} = \mathbf{F}_{11}\mathbf{F}_{12}\mathbf{P}\mathbf{F}_{12}'\mathbf{F}_{11}'$, where $\mathbf{F}_{11}$ is the inverse of the identity minus $\mathbf{H}'$, and $\mathbf{F}_{12}$ is the identity minus $\mathbf{H}$.

The COSAN model can be estimated using several different fit functions. These include maximum likelihood if one assumes the data follow a multivariate normal distribution. Other fit functions include unweighted least squares and generalized least squares. Estimates and test statistics can be obtained using SAS's Proc Calis package, which offers a variety of fit functions and convergence algorithms. All of the fit functions mentioned previously can be fit with Proc Calis. Proc Calis also offers the user the choice of the several optimization techniques. These include conjugate-gradient techniques, the Marquardt technique, and Newton-Raphson techniques. Furthermore, parameter constraints can be specified using SAS programming statements.

7.3 MODELING CANONICAL VARIATE ANALYSIS OVER TIME AS A COVARIANCE STRUCTURE

The CVA over time model that I present in this section is more akin to an extension of RA than of CVA, as the canonical variates are orthogonal in their weights as opposed to being uncorrelated. Essentially, I will model the between-groups covariance matrix. Start with the standard (non-longitudinal) case by performing a spectral decomposition of the between-groups covariance matrix. Let the between-groups covariance matrix for p Y-variables be denoted as CovB($\mathbf{Y}$), and let $\mathbf{V}$ denote a $p \times r$ columnwise orthonormal matrix of variate weights. Then

$$\text{CovB}(\mathbf{Y}) = \mathbf{V}\mathbf{D}^2\mathbf{V}',$$

where $\mathbf{D}$ is an $r \times r$ diagonal matrix whose i^{th} diagonal element is the square root of the between-groups variation of the i^{th} column of $\mathbf{V}$, $\mathbf{v}_\text{i}$.

Now consider the multiple occasions case. (As a reminder, $\mathbf{X}$ and $\mathbf{Y}$ are assumed to be centered). The products matrix for $\mathbf{Y}_\text{i}$ and $\mathbf{Y}_\text{j}$ regressed on $\mathbf{X}$ is $\mathbf{Y}_\text{i}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_\text{j}$, where $\mathbf{Y}_\text{i}$ is the Y-data at the i^{th} occasion. But $(n-1)^{-½}(\mathbf{X}'\mathbf{X})^{-½}\mathbf{X}'\mathbf{Y}_\text{j} = \mathbf{W}^*\mathbf{D}_\text{j}\mathbf{V}'$, where $\mathbf{W}^*$, and $\mathbf{V}$ are the redundancy variates for the X-variables and Y-variables as defined in Section **2.2.4**, and $\mathbf{D}_\text{j}$ is a diagonal matrix whose i^{th} diagonal element is the square root of the variance explained by the i^{th} variate; see equation (2.2). Now if one assumes that one has common variates at each of g occasions, then

$$(n-1)^{-1}\mathbf{Y}_\text{i}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_\text{j} = \mathbf{V}\mathbf{D}_\text{i}\mathbf{D}_\text{j}\mathbf{V}'. \tag{7.2}$$

But (7.2) implies

$$\text{CovB}(\mathbf{Y}_1:\mathbf{Y}_2:\cdots:\mathbf{Y}_g) = \begin{bmatrix} \mathbf{V}\mathbf{D}_1^2\mathbf{V}' & \mathbf{V}\mathbf{D}_1\mathbf{D}_2\mathbf{V}' & \cdots & \mathbf{V}\mathbf{D}_1\mathbf{D}_g\mathbf{V}' \\ \mathbf{V}\mathbf{D}_2\mathbf{D}_1\mathbf{V}' & \mathbf{V}\mathbf{D}_2^2\mathbf{V}' & \cdots & \mathbf{V}\mathbf{D}_2\mathbf{D}_g\mathbf{V}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}\mathbf{D}_g\mathbf{D}_1\mathbf{V}' & \mathbf{V}\mathbf{D}_g\mathbf{D}_2\mathbf{V}' & \cdots & \mathbf{V}\mathbf{D}_g^2\mathbf{V}' \end{bmatrix}, \tag{7.3}$$

where $\mathrm{CovB}(\mathbf{Y}_1 : \mathbf{Y}_2 : \cdots : \mathbf{Y}_g)$ indicates the between-groups covariance matrix for the p variables over g occasions. This between-groups covariance matrix follows a non-central, deficient Wishart distribution. Thus maximum likelihood estimates are not readily obtained. However, (7.3) can be estimated by the method of least squares.

In order to estimate with the method of maximum likelihood one must include the within-groups covariance matrix in the model, as the between-groups covariance matrix plus the within-groups covariance matrix yield the overall covariance structure, which does follow a Wishart distribution. A plausible assumption is that the within-groups covariance matrices at all occasions are proportional to $\mathbf{E}$, where $\mathbf{E}$ is a $p \times p$ positive definite matrix. Then $\mathrm{CovW}(\mathbf{Y}) = \mathbf{A} \otimes \mathbf{E}$, where $\mathrm{CovW}(\mathbf{Y})$ is the $pg \times pg$ within-groups covariance matrix, and $\mathbf{A}$ is a $g \times g$ positive semi-definite matrix scaled such that $\mathrm{trace}(\mathbf{A}) = g$.

The approach outlined above does not model $\mathbf{W}$, the matrix of coefficients for the X-variables (which are group indicators). If one desires an estimate of $\mathbf{W}$, a reasonable approach is to find $\mathbf{W}$ which minimizes the sums of squares fit to $\mathbf{S}_{\mathrm{XYk}} = \mathbf{W}\mathbf{D}_\mathrm{k}\mathbf{V}'$, for $\mathrm{k} = 1, \ldots, g$. This is equivalent to a step in the alternating least squares algorithm for the PARAFAC (orth.) model.

A more complicated way to obtain $\mathbf{W}$ is to model the $\mathbf{S}_{\mathrm{XY}}$ and $\mathbf{S}_{\mathrm{XX}}$ matrices along with the $\mathbf{S}_{\mathrm{YY}}$ matrices. By definition $\mathbf{S}_{\mathrm{XX}} = \mathbf{W}^{-1\prime}\mathbf{W}^{-1}$ since $\mathbf{W}'\mathbf{S}_{\mathrm{XX}}\mathbf{W} = \mathbf{I}$ and $\mathbf{S}_{\mathrm{XY1}} = \mathbf{W}^{-1}\mathbf{D}_1\mathbf{V}'$ since $\mathbf{W}'\mathbf{S}_{\mathrm{XY1}}\mathbf{V} = \mathbf{D}_1$. Hence the resulting model is

$$\begin{bmatrix} \mathbf{S}_{\mathrm{XX}} & \mathbf{S}_{\mathrm{XY1}} & \mathbf{S}_{\mathrm{XY2}} & \cdots & \mathbf{S}_{\mathrm{XYg}} \\ \mathbf{S}_{\mathrm{Y1X}} & \mathbf{S}_{\mathrm{Y1Y1}} & \mathbf{S}_{\mathrm{Y1Y2}} & \cdots & \mathbf{S}_{\mathrm{Y1Yg}} \\ \mathbf{S}_{\mathrm{Y2X}} & \mathbf{S}_{\mathrm{Y2Y1}} & \mathbf{S}_{\mathrm{Y2Y2}} & \cdots & \mathbf{S}_{\mathrm{Y2Yg}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{\mathrm{YgX}} & \mathbf{S}_{\mathrm{YgY1}} & \mathbf{S}_{\mathrm{YgY2}} & \cdots & \mathbf{S}_{\mathrm{YgYg}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}^{-1\prime}\mathbf{W}^{-1} & \mathbf{W}^{-1\prime}\mathbf{D}_1\mathbf{V}' & \mathbf{W}^{-1\prime}\mathbf{D}_2\mathbf{V}' & \cdots & \mathbf{W}^{-1\prime}\mathbf{D}_\mathrm{k}\mathbf{V}' \\ \mathbf{W}^{-1\prime}\mathbf{D}_1\mathbf{V}' & \mathbf{V}\mathbf{D}_1^2\mathbf{V}' & \mathbf{V}\mathbf{D}_1\mathbf{D}_2\mathbf{V}' & \cdots & \mathbf{V}\mathbf{D}_1\mathbf{D}_\mathrm{k}\mathbf{V}' \\ \mathbf{W}^{-1\prime}\mathbf{D}_2\mathbf{V}' & \mathbf{V}\mathbf{D}_2\mathbf{D}_1\mathbf{V}' & \mathbf{V}\mathbf{D}_2^2\mathbf{V}' & \cdots & \mathbf{V}\mathbf{D}_2\mathbf{D}_\mathrm{k}\mathbf{V}' \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}^{-1\prime}\mathbf{D}_\mathrm{k}\mathbf{V}' & \mathbf{V}\mathbf{D}_\mathrm{k}\mathbf{D}_1\mathbf{V}' & \mathbf{V}\mathbf{D}_\mathrm{k}\mathbf{D}_2\mathbf{V}' & \cdots & \mathbf{V}\mathbf{D}_\mathrm{k}^2\mathbf{V}' \end{bmatrix}.$$

7.4 PUTTING CVA OVER TIME IN THE COSAN FRAMEWORK

Next I show how the CVA over time model is expressed in terms of the $\mathbf{F}_\mathrm{i}$ and $\mathbf{P}_\mathrm{i}$ matrices of (7.1). The model for the between-groups covariance matrix in (7.3) can be decomposed as in (7.4):

$$\begin{bmatrix} \mathbf{V} & & \\ & \ddots & \\ & & \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{D}_1 & & \\ & \ddots & \\ & & \mathbf{D}_\mathrm{k} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \cdots & \mathbf{I} \\ \vdots & \ddots & \vdots \\ \mathbf{I} & \cdots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{D}_1 & & \\ & \ddots & \\ & & \mathbf{D}_\mathrm{k} \end{bmatrix} \begin{bmatrix} \mathbf{V} & & \\ & \ddots & \\ & & \mathbf{V} \end{bmatrix}' . \qquad (7.4)$$

Note that the rank of the center matrix and thus the rank of the product of the matrices is r, where r is the number of common variates, $r \le \min(p, g-1)$. Now, the second and fourth matrices in

(7.4), that is, the ones with block diagonals of $\mathbf{D}_1$ to $\mathbf{D}_g$, and the center matrix can be expressed directly in the COSAN model. For example, one could label the matrix of block diagonals of $\mathbf{D}_1$ to $\mathbf{D}_g$ as $\mathbf{F}_{11}$, and the center matrix as $\mathbf{P}_1$. However, $\mathbf{V}$ must be modeled indirectly, using the Cayley decomposition to constrain it to orthogonality. Hence let the matrix $\mathbf{I}_{(k)} \otimes \mathbf{V}$ be $\mathbf{I}_{(k)} \otimes (\mathbf{I}_{(p)} + \mathbf{H})^{-1} \mathbf{I}_{(k)} \otimes (\mathbf{I}_{(p)} - \mathbf{H})$. Let $\mathbf{I}_{(k)} \otimes (\mathbf{I}_{(p)} + \mathbf{H})^{-1}$ be labeled $\mathbf{F}_{12}$ and let $\mathbf{I}_{(k)} \otimes (\mathbf{I}_{(p)} - \mathbf{H})$ be labeled $\mathbf{F}_{13}$. Thus the between groups covariance is modeled as

$$\mathbf{F}_{12}\mathbf{F}_{13}\mathbf{F}_{11}\mathbf{P}_1\mathbf{F}_{11}'\mathbf{F}_{13}'\mathbf{F}_{12}'.$$

To see how the error matrix, $\mathrm{CovW}(\mathbf{Y}) = \mathbf{A} \otimes \mathbf{E}$, would be expressed in the terms of the $\mathbf{F}_i$ and $\mathbf{P}_i$ matrices of (7.1), notice that:

$$\mathbf{A} \otimes \mathbf{E} = \begin{bmatrix} \mathbf{L} & & \\ & \mathbf{L} & \\ & & \ddots \\ & & \mathbf{L} \end{bmatrix} \begin{bmatrix} a_{11}\mathbf{I}(p) & a_{12}\mathbf{I}(p) & \cdots & a_{1g}\mathbf{I}(p) \\ a_{21}\mathbf{I}(p) & a_{22}\mathbf{I}(p) & \cdots & a_{2g}\mathbf{I}(p) \\ \vdots & \vdots & \ddots & \vdots \\ a_{g1}\mathbf{I}(p) & a_{g2}\mathbf{I}(p) & \cdots & a_{gg}\mathbf{I}(p) \end{bmatrix} \begin{bmatrix} \mathbf{L}' & & & \\ & \mathbf{L}' & & \\ & & \ddots & \\ & & & \mathbf{L}' \end{bmatrix}$$

where $\mathbf{L}$ is a $p \times p$ matrix. Thus $\mathbf{E}$ is not directly modeled by COSAN, but is determined as $\mathbf{E} = \mathbf{L}\mathbf{L}'$. One can call the matrix of block diagonals of $\mathbf{L}$ matrices $\mathbf{F}_{21}$, and the center matrix $\mathbf{P}_2$. These terms are added to those that model the between groups variation. Thus the model for the total covariance is the between groups covariance plus the within groups covariance:

$$\mathbf{F}_{12}\mathbf{F}_{13}\mathbf{F}_{11}\mathbf{P}_1\mathbf{F}_{11}'\mathbf{F}_{13}'\mathbf{F}_{12}' + \mathbf{F}_{21}\mathbf{P}_2\mathbf{F}_{21}'.$$

It is necessary to mention a further detail. In the CVA over time model one will usually desire a solution with fewer variates than *p*, say *r*. Thus one would like to model orthonormal $\mathbf{V}$, where $\mathbf{V}$ has $r < p$ columns. But the Cayley decomposition requires a square matrix. Hence one needs to model dummy canonical variates in the matrix $\mathbf{V}$. One fixes these dummy variates by setting certain elements of $\mathbf{V}$ equal to zero. However, this must be done indirectly because in the COSAN model one directly estimates the matrix of parameters $\mathbf{H}$. The details on how to do this are found in Appendix Five.

7.5 AN EXAMPLE

In this section I revisit the data from the Shenandoah study which was described in Section **5.4**. Because of unresolved difficulties in programming in Proc Calis and in executing programs that model large covariance matrices, the example will be limited to modeling one canonical variate over the first two occasions, August and September of 1981. Since the error variance was previously determined to be non-homogeneous (see **Table 5.1**), the between-groups covariance matrix is being modeled with the fit function unweighted least squares. The SAS program is given in Appendix Five.

The weights for the first canonical variate are presented in **Figure 7.1**, along with the first canonical variate of the PARAFAC (orth.) solution for comparison. These weights are similar, as are the estimated roots of the variances explained by each variate, which are 2.47 and 2.65 for

COSAN, in contrast to 2.25 and 2.52 for the PARAFAC (orth.). The similarity in the parameter estimates is evidence that both methods are correctly estimating the common variate, though it will be necessary to estimate a larger COSAN model with data from all the occasions to have convincing evidence of this.

Measurement	COSAN Estimates	PARAFAC (orth.) Estimates
discharge	-0.024	0.036
conductivity	**0.411**	**0.369**
pH	0.046	0.113
temperature	0.097	0.132
Ca^{++}	**0.415**	**0.383**
Mg^{++}	**0.434**	**0.365**
Na^{+}	**0.312**	**0.364**
K^{+}	-0.069	-0.158
alkalinity	**0.467**	**0.394**
$SO_4^{=}$	0.052	0.14
Cl^{-}	**0.220**	**0.265**
$SiO_4^{=}$	**0.293**	**0.363**
NO_3^{-}	0.023	0.133
NH_4^{+}	-0.001	0.035

Figure 7.1 Estimates for the COSAN and PARAFAC Models

7.6 CONCLUDING REMARKS

In summary, COSAN offers a flexible and powerful modeling tool for modeling CVA with longitudinal data. However, work needs to be done to overcome programming and estimation difficulties. One problem is that the SAS system uses up all the available memory when large covariance matrices are analyzed. Another is that when writing code each element of each matrix must be specified, making programming a laborious task for modeling large matrices. This is seen in the SAS code in Appendix Five. A way to solve this problem is to write a macro that sets up the program code.

If the difficulties mentioned in the previous paragraph are overcome, then the model can be further developed. For example, modeling uncorrelated canonical variates would be useful, as would a model that hypothesized that some variates be unique to each occasion.

CHAPTER EIGHT

CANONICAL VARIATE ANALYSIS OVER TIME

8.1 INTRODUCTION

In this chapter I present a model for CVA with measurements made over multiple occasions. In contrasts to earlier chapters, this chapter emphasizes statistical inference based on maximum likelihood methods. I shall call the models developed in this chapter CVA/time, though I distinguish between models with orthogonal canonical variates, CVA/time (orthogonal), and those with uncorrelated canonical variates, CVA/time (uncorrelated). CVA/time is suggested by Campbell and Tomenson`s (1983) model for CVA with multiple datasets, which hypothesizes that the group means lie on planes defined by canonical variates common to all datasets (see Section **2.4**). Analogously, CVA/time hypothesizes that the group means lie on planes defined by canonical variates which are common to all occasions.

The goal of the CVA/time model is to answer the question of what is and what is not changing over time when one has multivariate data with group structure. In particular it attempts to determine if the canonical variates are stable over time, and if they are, if the positions of the group means on the canonical variates are changing over time. Thus CVA/time is the only model in this dissertation that will estimate the group positions and develop hypothesis tests to determine if they are equal over time.

Chapter **Eight** is organized as follows. In Section **8.2** I make two preliminary points. In Section **8.3** I detail a model for group means in the space of orthogonal canonical variates, CVA/time (orthogonal). I also derive estimating equations for this model, discuss their solution and describe how to make statistical inferences. In Section **8.4** I use simulated data to test the methodology of Section **8.3**. In Section **8.5** I derive a model for group means in the space of uncorrelated canonical variates, CVA/time (uncorrelated). Uncorrelated variates entail assuming a particular structure for the within-groups covariance matrix. The estimation of this structure is also discussed in this section. In Section **8.6** I illustrate the methodology of Section **8.5** by analyzing a real dataset. Lastly, in Section **8.7** I compare CVA/time with several alternative methods for this type of data, with particular attention given to doubly multivariate repeated measures.

8.2 PRELIMINARIES

8.2.1 Orthogonal Versus Uncorrelated Variates

This chapter presents two models with differing assumptions about the structure of the group means. The first model to be discussed CVA/time (orth.), hypothesizes that the canonical variates are orthogonal to each other in their weights. It models the positions of the group means in the space of the untransformed data. The second model to be discussed, CVA/time (unc.), hypothesizes that the canonical variates are uncorrelated, which is consistent with the standard definition of canonical variates. It models the positions of the group means in the space transformed by the Mahalanobis transformation.. I shall present the CVA/time (orth.) model first because it is simpler.

The main reason to model orthogonal variates is that, unlike uncorrelated variates, they do not require the assumption that one has the same within-groups covariance structure at each occasion, an assumption which may be unrealistic. Beyond the issue of whether the within-groups covariance matrices are stable over time, there are important differences between the approaches whose implications the researcher needs to consider. These differences are analogous to the differences between canonical variate analysis and redundancy analysis, a topic which is discussed in Section **2.2.3**. Uncorrelated variates have the important advantage that they are scale invariant. They also are more closely related to the goal of optimizing the discrimination among the groups. Though CVA/time (unc.) does not explicitly maximize group discrimination over time (see Section **5.2.2** for something along this line), it is a generalization of CVA, which does. On the other hand, the CVA/time (orth.) model is not a true generalization of CVA, but is more akin to a generalization of redundancy analysis for grouped data (see Section **2.2.3**).

The situations where one may prefer orthogonal variates to uncorrelated variates when uncorrelated variables are feasible are the same as those where one would prefer to perform a redundancy analysis over a canonical correlation analysis. CVA may find group differences which are large in terms of discrimination but small in terms of between-groups variation explained. Or, the total variation explained may be of direct interest. For example, if the measurements made are directly comparable, such as if one had a battery of exams with the same scales, one may prefer to maximize the total variance explained by the group structure.

8.2.2 The Structure of the Data

A clarification of how the data is organized illuminates the discussion of the previous section and other issues not yet touched upon. The same variables measured at different occasions will be treated as distinct variables. Hence *tp* variables are effectively modeled, where *t* is the number of occasions and *p* is the number of variables measured at one occasion. This contrasts with Campbell & Tomenson's method which models distinct datasets of the same *p* variables.

It will be necessary to partition Σ, the $tp \times tp$ within-groups covariance matrix, into t^2 $p \times p$ matrices Σ_{qs}, where Σ_{qs} is the matrix of covariances between the measurements of the q^{th} and s^{th} occasions, $q, s = 1, \ldots, t$. The partitioning of Σ is shown below:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{21}' & \cdots & \Sigma_{t1}' \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{t2}' \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{t1} & \Sigma_{t2} & \cdots & \Sigma_{tt} \end{bmatrix}. \tag{8.1}$$

The model developed in Section **8.3** assumes no specific structure for Σ. A consequence of this flexibility in Σ is that there is no common $p \times p$ within-groups covariance matrix, Σ_{qq}, by which to transform the data. In Section **8.5.2** a model will be introduced which assumes a structure for Σ that specifies common Σ_{qq} and thus allows for modeling uncorrelated variates.

8.3 THE CVA/TIME (ORTHOGONAL) MODEL

In this section I develop a model for analyzing group structure with longitudinal multivariate data which I shall call CVA/time (orthogonal). CVA/time (orth.) is not a generalization of CVA/time, but rather a generalization of redundancy analysis. Beyond interest in its own right, the discussion of this model introduces basic ideas and methods which will be used later for the CVA/time model with uncorrelated variates. In particular, I introduce the concepts of common and unique variates, group positions, the methods of obtaining estimates, and statistical inference.

The model I develop in this section encompasses several possible cases. One basic model hypothesizes a given number of variates common to all occasions. A simple alternative to this model is one that hypothesizes that there are an equal number of variates specific or unique at

each occasion. Henceforth I shall refer to the former as common variates and to the latter as unique variates. Unique variates are the natural alternative to common variates because they hypothesize variates that change over time, and the interest is to determine what is and what is not changing over time.

The model which I will refer to as CVA/time (orth.) hypothesizes both types of variates. However, even more complex models are possible. For example, one could hypothesize variates which are common to only a subset of the groups. Estimating equations can be derived for all of the possible models using the methods of calculus, though I shall derive them only for the CVA/time (orth.) model.

It is useful to give a simple example of a common variate model. Assume the positions of two group means can be plotted on one canonical variate which is common over two occasions, and assume the positions of the group means change over time. **Figure 8.1** shows the positions at the first occasion, and **Figure** Error! Reference source not found. shows the positions at the second occasion.

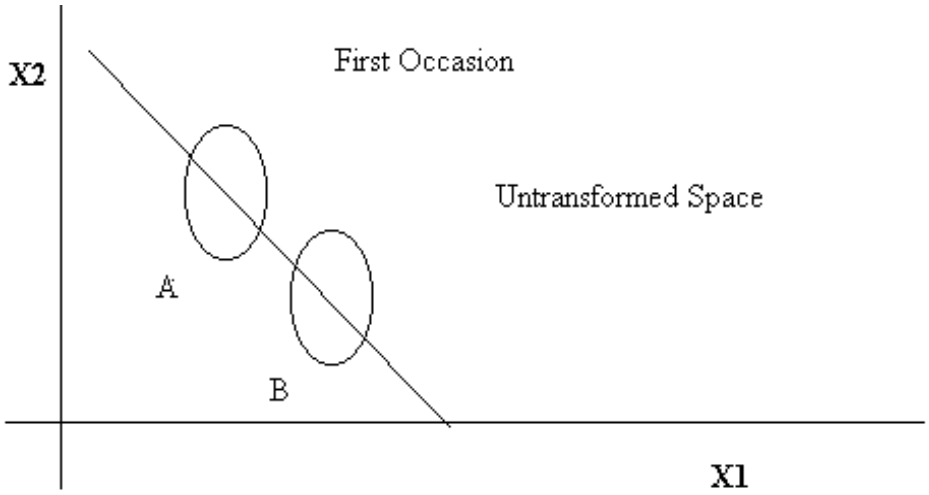

Figure 8.1

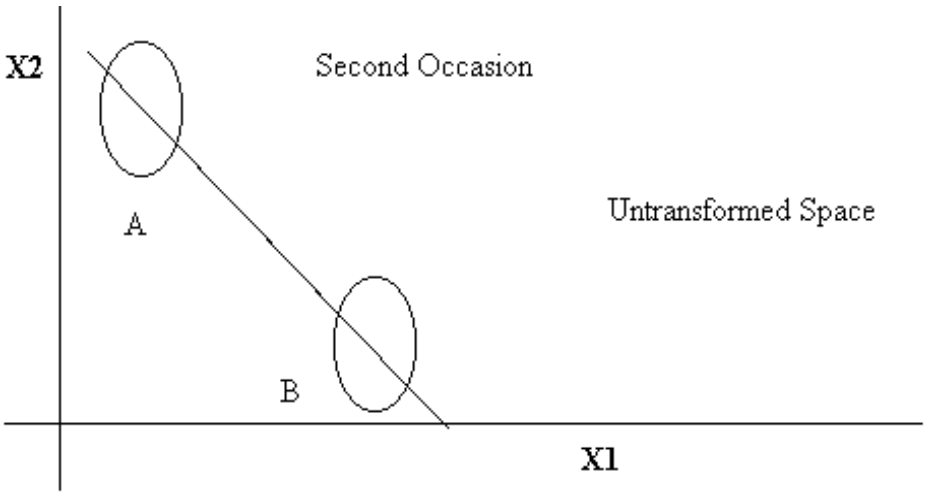

Figure 8.2

8.3.1 The CVA/Time Model with Orthogonal Variates

The CVA/time (orth.) model is specified as follows; assume the data follow the multivariate normal distribution; $\mathbf{x}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\mathbf{x}_i$ is a tp vector of random variables, $\boldsymbol{\mu}_i$ is a tp vector of means, and $\boldsymbol{\Sigma}$ is a $tp \times tp$ covariance matrix. Further assume that $\boldsymbol{\mu}_i$ is completely determined by group membership, so that $\boldsymbol{\mu}_i \in \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g\}$, depending on the group membership of the i^{th} observation.

The model for the structure of the means given in equation (8.2) below specifies u variates for each occasion q; c of these variates, $\mathbf{v}_1, \ldots, \mathbf{v}_c$, are common to all occasions, where $\mathbf{v}_i$ indicates the i^{th} common variate. $u - c$ of these variates, $\mathbf{v}^q_{c+1}, \ldots, \mathbf{v}^q_u$, are unique to the q^{th} occasion, where $\mathbf{v}^q_i$ indicates the i^{th} variate of the set of variates for the q^{th} occasion. Thus the model for the g^{th} group mean, $\boldsymbol{\mu}_g$, $g = 1, \ldots, m$, is:

$$\boldsymbol{\mu}_g = \boldsymbol{\mu}_0 + \mathbf{v}_1 \otimes \mathbf{e}_{g,1} + \ldots + \mathbf{v}_c \otimes \mathbf{e}_{g,c} + \begin{bmatrix} e^1_{g,c+1}\mathbf{v}^1_{c+1} \\ \vdots \\ e^t_{g,c+1}\mathbf{v}^t_{c+1} \end{bmatrix} + \cdots + \begin{bmatrix} e^1_{g,u}\mathbf{v}^1_u \\ \vdots \\ e^t_{g,u}\mathbf{v}^t_u \end{bmatrix}, \qquad (8.2)$$

where $\boldsymbol{\mu}_g$ is a $pt \times 1$ vector of means for the g^{th} group, $\boldsymbol{\mu}_0$ is a $pt \times 1$ vector of overall means, $\mathbf{v}_i$ are c $p \times 1$ vectors of common variates, $\mathbf{v}^k_j$ are $(u - c)$ $t \times 1$ vectors of unique variates, $e^q_{g,i}$ is the score for the g^{th} group mean on the i^{th} canonical variate at the q^{th} occasion, and $\mathbf{e}_{g,i}$ is the $t \times 1$ vector whose elements are $e^q_{g,i}$.

Note the constraints on the parameters. First, the group positions for each occasion for each variate sum to zero, i.e., $\sum_{g=1}^{m} n_g e^q_{g,i} = 0$ for $q = 1, \ldots, t$, and $i = 1, \ldots, u$. This constraint is just

a reflection of the fact that the model is centered by an overall mean, $\boldsymbol{\mu}_0$. Second, the common variates are mutually orthogonal:

$$\mathbf{V}'_{\text{com}}\mathbf{V}_{\text{com}} = \mathbf{I}_{c\times c},$$

where $\mathbf{V}_{\text{com}}$ is the matrix whose columns are the c common variates. Furthermore, within each set of unique variates for each occasion the variates are constrained to be mutually orthogonal, that is:

$$\mathbf{V}^{\text{q}\prime}\mathbf{V}^{\text{q}} = \mathbf{I}_{(u-c)\times(u-c)},$$

where $\mathbf{V}^{\text{q}}$ is the matrix whose columns are the $u-c$ unique variates for the q^{th} occasion, $\text{q} = 1,\ldots,t$. Finally, each variate in each set of unique variates is orthogonal with each common variate. Thus:

$$\mathbf{V}'_{\text{com}}\mathbf{V}^{\text{q}} = [0]_{c\times(u-c)}.$$

Note there is a limit on $u-c$, the number of unique variates one can have at each occasion. $u-c$ cannot be greater than the modular of $\frac{p}{t}$. For example, if $p = t$, then a model can hypothesize at most one unique variate at each occasion. Further, such a model is equivalent to a model with p common variates.

8.3.2 Sufficient Statistics

Before proceeding to develop the estimating equations I will show a result for grouped multivariate data that will simplify the later derivations. I will show that $\bar{\mathbf{x}}_{\text{g}}$ and $\mathbf{S}$, the sample means and within-groups covariance matrix, are sufficient statistics for $\boldsymbol{\mu}_{\text{g}}$ and $\boldsymbol{\Sigma}$, and consequently for the parameters with which I later model $\boldsymbol{\mu}_{\text{g}}$ and $\boldsymbol{\Sigma}$. The likelihood equations for multivariate grouped data are as follows below, where $\mathbf{X}$ indicates the data matrix, $\boldsymbol{\mu}$ indicates the parameters determining the mean of the variables, $\boldsymbol{\Sigma}$ indicates the parameters determining the covariance of the variables, $L(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ indicates the likelihood of the data $\mathbf{X}$ given parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and $\mathbf{x}_{\text{ig}}$ is the i^{th} observation in the g^{th} group:

$$L(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{npt}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}}\exp\left\{\frac{-1}{2}\left(\sum_{\text{g}=1}^{m}\sum_{\text{i}=1}^{n_{\text{g}}}(\mathbf{x}_{\text{gi}}-\boldsymbol{\mu}_{\text{g}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{\text{gi}}-\boldsymbol{\mu}_{\text{g}})\right)\right\}. \tag{8.3}$$

Consider the part of the likelihood equation which is a function of $\mathbf{x}_{\text{gi}}$ and $\boldsymbol{\mu}_{\text{g}}$ and call it K. Then K is:

$$\begin{aligned}\text{K} &= \sum_{\text{g}=1}^{m}\sum_{\text{i}=1}^{n_{\text{g}}}(\mathbf{x}_{\text{gi}}-\boldsymbol{\mu}_{\text{g}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{\text{gi}}-\boldsymbol{\mu}_{\text{g}})\\ &= \sum_{\text{g}=1}^{m}\sum_{\text{i}=1}^{n_{\text{g}}}\text{tr}(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{\text{gi}}-\boldsymbol{\mu}_{\text{g}})(\mathbf{x}_{\text{gi}}-\boldsymbol{\mu}_{\text{g}})')\end{aligned}$$

$$= \sum_{g=1}^{m}\sum_{i=1}^{n_g} \mathrm{tr}(\Sigma^{-1}((\mathbf{x}_{gi} - \bar{\mathbf{x}}_g) + (\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g))((\mathbf{x}_{gi} - \bar{\mathbf{x}}_g) + (\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g))')$$

Since $\sum_{i=1}^{n_g} \mathrm{tr}(\Sigma^{-1}(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g)') = 0$, for $g = 1,\ldots,m$,

$$\mathrm{K} = \sum_{g=1}^{m}\sum_{i=1}^{n_g} \mathrm{tr}(\Sigma^{-1}(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)' + \Sigma^{-1}(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g)(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g)').$$

Recognizing that $\mathbf{S} = \frac{1}{n}\sum_{g=1}^{m}\sum_{i=1}^{n_g}(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)'$ and further rearrangement gives:

$$\mathrm{K} = \tfrac{1}{n}\Sigma^{-1}\mathbf{S} + \sum_{g=1}^{m} n_g(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g)'\Sigma^{-1}(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g).$$

Replacing K back into the likelihood equation one has the desired result:

$$L(\mathbf{X}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{\frac{npt}{2}}|\Sigma|^{\frac{n}{2}}} \exp\left\{-\tfrac{1}{2}\left(\tfrac{1}{n}\Sigma^{-1}\mathbf{S} + \sum_{g=1}^{m} n_g(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g)'\Sigma^{-1}(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g)\right)\right\}. \tag{8.4}$$

From this form of the likelihood function it is clear that $\bar{\mathbf{x}}_g$ and $\mathbf{S}$ are sufficient statistics for $\boldsymbol{\mu}_g$ and Σ because the likelihood function is factored into a part which is a function of the sufficient statistics $\bar{\mathbf{x}}_g$ and $\mathbf{S}$, and the parameters $\boldsymbol{\mu}_g$ and Σ, and a part which is not a function of $\boldsymbol{\mu}_g$ and Σ.

8.3.3 Estimating Equations

In this section I develop estimating equations for the CVA/time (orth.) model. Henceforth I will work with the log-likelihood equation instead of the likelihood equation. Let $l(\mathbf{X}|,\boldsymbol{\mu},\Sigma)$ stand for the natural logarithm of the likelihood of the data $\mathbf{X}$ given parameters $\boldsymbol{\mu}$ and Σ. Then $l(\mathbf{X}|,\boldsymbol{\mu},\Sigma)$ is:

$$l(\mathbf{X}|\boldsymbol{\mu},\Sigma) = \frac{-npt}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma|$$
$$-\tfrac{1}{2}\left(\sum_{g=1}^{m}\sum_{i=1}^{n_g}(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)'\Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_g) + \sum_{g=1}^{m} n_g(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g)'\Sigma^{-1}(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g)\right). \tag{8.5}$$

First I derive the maximum likelihood estimator for $\boldsymbol{\mu}_o$. For convenience, let C denote terms not involving $\boldsymbol{\mu}_g$, and $\mathrm{F}(\mathbf{v}_i, \mathbf{v}_i^q, \mathbf{e}_{g,i})$ denote the terms in the model for $\boldsymbol{\mu}_g$ (8.2) that do not involve $\boldsymbol{\mu}_0$; hence $\boldsymbol{\mu}_g = \boldsymbol{\mu}_0 + \mathrm{F}(\mathbf{v}_i, \mathbf{v}_i^q, \mathbf{e}_{g,i})$. Then the log-likelihood is:

$$l(\mathbf{X}|\boldsymbol{\mu},\Sigma) = \mathrm{C} - \tfrac{1}{2}\sum_{g=1}^{m} n_g\left(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_0 - \mathrm{F}(\mathbf{v}_i, \mathbf{v}_i^q, \mathbf{e}_{g,i})\right)' \Sigma^{-1}\left(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_0 - \mathrm{F}(\mathbf{v}_i, \mathbf{v}_i^q, \mathbf{e}_{g,i})\right).$$

Taking the derivatives of the log-likelihood with respect to $\boldsymbol{\mu}_o$ yields the following:

$$\frac{\delta l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta \boldsymbol{\mu}_o} = -\tfrac{1}{2}\sum_{g=1}^{m}\left(n_g 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_o - 2n_g\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_g + 2n_g\boldsymbol{\Sigma}^{-1}\mathrm{F}(\mathbf{v}_i,\mathbf{v}_i^r,\mathbf{e}_{g,i})\right). \tag{8.6}$$

One sets these derivatives equal to zero to obtain the estimating equations for $\boldsymbol{\mu}_o$. The last term in (8.6) drops out as $\sum_{g=1}^{m} n_g \mathrm{F}(\mathbf{v}_i,\mathbf{v}_i^q,\mathbf{e}_{g,i}) = \mathbf{0}$, where $\mathbf{0}$ is a $pt \times 1$ vector of zeros, because $\sum_{g=1}^{m} e_{g,i}^q = 0$ for all q,i. Hence:

$$\sum_{g=1}^{m} n_g\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_o = \sum_{g=1}^{m} n_g\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_g \ .$$

Multiplying through by $\boldsymbol{\Sigma}$ gives the maximum likelihood estimate for $\boldsymbol{\mu}_o$, which I denote as $\hat{\boldsymbol{\mu}}_o$,

$$\hat{\boldsymbol{\mu}}_o = \left(\sum_{g=1}^{m} n_g\right)^{-1}\sum_{g=1}^{m} n_g\bar{\mathbf{x}}_g = n^{-1}\sum_{g=1}^{m} n_g\bar{\mathbf{x}}_g \,. \tag{8.7}$$

$\hat{\boldsymbol{\mu}}_o$ is just the average over all observations.

Next, I derive estimating equations for $\boldsymbol{\Sigma}$. Denote by C terms that are not a function of $\boldsymbol{\Sigma}$. Then $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is:

$$l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathrm{C} - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\left(\sum_{g=1}^{m}\sum_{i=1}^{n_g}(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g)' + \sum_{g=1}^{m} n_g(\bar{\mathbf{x}}_g-\boldsymbol{\mu}_g)\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_g-\boldsymbol{\mu}_g)'\right).$$

Taking the derivative of $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ yields:

$$\begin{aligned}\frac{\delta l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta\boldsymbol{\Sigma}} = &-n\boldsymbol{\Sigma}^{-1} + \tfrac{n}{2}\,\mathrm{diag}(\boldsymbol{\Sigma}^{-1}) \\ &+\sum_{g=1}^{m}\sum_{i=1}^{n_g}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g)(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g)'\boldsymbol{\Sigma}^{-1} - \tfrac{1}{2}\,\mathrm{diag}\left(\sum_{g=1}^{m}\sum_{i=1}^{n_g}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g)(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g)'\boldsymbol{\Sigma}^{-1}\right) \\ &+\sum_{g=1}^{m} n_g\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_g-\boldsymbol{\mu}_g)(\bar{\mathbf{x}}_g-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1} - \tfrac{1}{2}\,\mathrm{diag}\left(\sum_{g=1}^{m} n_g\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_g-\boldsymbol{\mu}_g)(\bar{\mathbf{x}}_g-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}\right).\end{aligned}$$

Pre-multiply and post-multiply the above by $\boldsymbol{\Sigma}$, and set the equations equal to a matrix of zeros. Then the normal equations are solved when:

$$\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{g=1}^{m}\sum_{i=1}^{n_g}(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g)(\mathbf{x}_{gi}-\bar{\mathbf{x}}_g)' + n^{-1}\sum_{g=1}^{m} n_g(\bar{\mathbf{x}}_g-\boldsymbol{\mu}_g)(\bar{\mathbf{x}}_g-\boldsymbol{\mu}_g)', \tag{8.8}$$

where $\hat{\boldsymbol{\Sigma}}$ denotes the estimate for $\boldsymbol{\Sigma}$.

One sees that $\hat{\boldsymbol{\Sigma}}$ is equal to **S**, the sample estimate of $\boldsymbol{\Sigma}$, plus an additional term which depends on the difference between the predicted and observed group means. Although the topic of obtaining estimates will be discussed later in Section **8.3.5**, it is worth mentioning now that this additional term will be small when the model is correctly specified, but inflated when the model is misspecified. **S**, on the other hand, is completely robust to model misspecification. Thus in practice using **S** may be preferable to $\hat{\boldsymbol{\Sigma}}$.

Next I derive the estimating equations for $\mathbf{v}_i$, $\mathbf{v}_i^r$ and $\mathbf{e}_{g,i}$. Substituting the means model for $\boldsymbol{\mu}_g$ from equation (8.2) into equation (8.4) gives the likelihood equations for CVA/time (orth.). Denote the log-likelihood by $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, and the terms which include neither $\mathbf{v}_i$, $\mathbf{v}_i^r$ nor $\mathbf{e}_{g,i}$ by C. Then:

$$l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathrm{C} - \tfrac{1}{2}\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\left(\sum_{a=1}^{c}\sum_{b=1}^{c} e_{g,a}^{q}\mathbf{v}_a'\boldsymbol{\Sigma}_{qs}^{-1}\mathbf{v}_b e_{g,b}^{s} + 2\sum_{a=1}^{c}\sum_{b=c+1}^{u} e_{g,a}^{q}\mathbf{v}_a'\boldsymbol{\Sigma}_{qs}^{-1}\mathbf{v}_b^{s} e_{g,b}^{s} + \sum_{a=c+1}^{u}\sum_{b=c+1}^{u} e_{g,a}^{q}\mathbf{v}_a^{q}\boldsymbol{\Sigma}_{qs}^{-1}\mathbf{v}_b^{s} e_{g,b}^{s}\right)$$
$$+ \tfrac{1}{2}\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\left(2\sum_{b=1}^{c}(\bar{\mathbf{x}}_g^{q} - \boldsymbol{\mu}_0^{q})'\boldsymbol{\Sigma}_{qs}^{-1}\mathbf{v}_b e_{g,b}^{s} + 2\sum_{b=c+1}^{k}(\bar{\mathbf{x}}_g^{q} - \boldsymbol{\mu}_0^{q})'\boldsymbol{\Sigma}_{qs}^{-1}\mathbf{v}_b^{s} e_{g,b}^{s}\right), \qquad (8.9)$$

where $\boldsymbol{\mu}_0^{q}$ is the $p \times 1$ vector of means averaged over all groups for the q^{th} occasion, and $\bar{\mathbf{x}}_g^{q}$ is the $p \times 1$ vector of sample means for the g^{th} group.

The estimating equations for the common variates are considered next. The estimation of variates requires consideration of the constraints. The constraints for the orthogonality of the common variates are incorporated by the method of Lagrangian multipliers. The constraints with Lagrangian multipliers for the unit length of the common variates are as follows (note that here and in subsequent developments the constraints with Lagrangian multipliers are implicitly set to zero):

$$\sum_{a=1}^{c}\frac{\gamma_a}{2}(\mathbf{v}_a'\mathbf{v}_a - 1),$$

where γ_a are c Lagrangian multipliers. The constraints with Lagrangian multipliers for the orthogonality of the common variates are:

$$\sum_{a=1}^{c}\sum_{b=1}^{a-1}\gamma_{ab}\mathbf{v}_a'\mathbf{v}_b,$$

where γ_{ab} are $c(c-1)/2$ Lagrangian multipliers. The constraints with Lagrangian multipliers for the orthogonality of each common variate with all of the unique variates are:

$$\sum_{q=1}^{t}\sum_{a=1}^{c}\sum_{b=c+1}^{u}\gamma_{abq}\mathbf{v}_a'\mathbf{v}_b^{q}, \qquad (8.10)$$

where γ_{abq} are $tc(u-c)$ Lagrangian multipliers. Denote the log-likelihood modified to incorporate the constraints with Lagrangian multipliers by $l^*(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. Take the derivative of $l^*(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $\mathbf{v}_f$:

$$\frac{\delta l^*(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta \mathbf{v}_f} = \sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\left(-\sum_{a=1}^{c} e_{g,a}^{q}\boldsymbol{\Sigma}_{qs}^{-1}\mathbf{v}_a e_{g,f}^{s} - \sum_{b=c+1}^{u} e_{g,f}^{q}\boldsymbol{\Sigma}_{qs}^{-1}\mathbf{v}_b^{s} e_{g,b}^{s} + (\bar{\mathbf{x}}_g^{q} - \boldsymbol{\mu}_0^{q})'\boldsymbol{\Sigma}_{qs}^{-1} e_{g,f}^{s}\right)$$
$$+ \gamma_f\mathbf{v}_f + \sum_{\substack{a=1 \\ a\neq f}}^{c}\gamma_{af}\mathbf{v}_a + \sum_{q=1}^{t}\sum_{b=c+1}^{u}\gamma_{fbq}\mathbf{v}_b^{q}.$$

Setting these derivatives equal to a vector of zeros yields the estimating equations to solve for $\mathbf{v}_f$.

Next I derive the estimating equations for the unique variates. The constraints with Lagrangian multipliers are as follows, starting with those which constrain the unique variates to unit length (note they are implicitly set to zero):

$$\sum_{a=c+1}^{u}\sum_{q=1}^{t}\frac{\gamma_{aq}}{2}\left(\mathbf{v}_a^{q\prime}\mathbf{v}_a^{q}-1\right),$$

where γ_{aq} are $(u-c)t$ Lagrangian multipliers. The constraints for the mutual orthogonality of each unique variate with the other unique variates of the same occasion are:

$$\sum_{q=1}^{t}\sum_{a=c+1}^{u}\sum_{b=c+1}^{a-1}\gamma_{abq}\mathbf{v}_a^{q\prime}\mathbf{v}_b^{q},$$

where γ_{abq} are $t(u-c)(u-c-1)/2$ Lagrangian multipliers. The constraints for the orthogonality of each common variate with all of the unique variates are already given in equation (8.10).

Now take the derivative of $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $\mathbf{v}_f^r$. These derivatives yield the estimating equations for solving for $\mathbf{v}_f^r$ when they are set to zero:

$$\frac{\delta l^*(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta\mathbf{v}_f^r}=\sum_{g=1}^{m}n_g\sum_{q=1}^{t}\left(-\sum_{a=1}^{c}e_{g,a}^{q}\boldsymbol{\Sigma}_{qr}^{-1}\mathbf{v}_a e_{g,w}^{r}-\sum_{b=c+1}^{u}e_{g,b}^{q}\boldsymbol{\Sigma}_{qr}^{-1}\mathbf{v}_w^{r}e_{g,w}^{r}+(\bar{\mathbf{x}}_{gq}-\boldsymbol{\mu}_0^{q})\boldsymbol{\Sigma}_{qr}^{-1}e_{g,f}^{r}\right)$$
$$+\gamma_{rf}\mathbf{v}_f^{r}+\sum_{s=1}^{c}\gamma_{afr}\mathbf{v}_a+\sum_{b=c+1}^{u}\gamma_{\phi\beta\rho}\mathbf{v}_b^{r}.$$

Lastly I derive estimating equations for the group positions, the $e_{g,b}^{s}$ terms, beginning with those corresponding to the common variates. The constraints for these terms are $\sum_{g=1}^{m}n_g e_{g,b}^{s}=0$, for $s=1,\ldots,t$ and $b=1,\ldots,c$. They are handled in the estimation by letting $e_{m,b}^{s}=-\sum_{h=1}^{m-1}e_{h,b}^{s}$ for $s=1,\ldots,t$ and $b=1,\ldots,c$. One takes the derivative of $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $e_{w,f}^{r}$ for $w\neq m$, and sets these derivatives equal to zero to obtain the estimating equations for the $e_{g,b}^{s}$ terms:

$$\frac{\delta l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta e_{w,f}^{r}}=\sum_{q=1}^{t}\left(-\sum_{a=1}^{c}e_{w,a}^{q}\mathbf{v}_f'\boldsymbol{\Sigma}_{qr}^{-1}\mathbf{v}_a-\sum_{b=c+1}^{u}e_{w,b}^{q}\mathbf{v}_f'\boldsymbol{\Sigma}_{qr}^{-1}\mathbf{v}_b^{s}+\left(\bar{\mathbf{x}}_w^{q}-\boldsymbol{\mu}_0^{q}\right)\boldsymbol{\Sigma}_{qr}^{-1}\mathbf{v}_f\right).$$

The estimating equations for the group positions corresponding to the unique variates are handled similarly to those corresponding to the common variates. The constraints are identical to those of the common variates. Take the derivative of $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $e_{w,f}^{r}$ for $h\neq m$:

$$\frac{\delta l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta e_{w,f}^{r}}=\sum_{q=1}^{t}\left(2\sum_{a=1}^{c}e_{w,a}^{q}\mathbf{v}_a'\boldsymbol{\Sigma}_{qr}^{-1}\mathbf{v}_f^{r}+2\sum_{b=c+1}^{u}\mathbf{v}_f'\boldsymbol{\Sigma}_{qr}^{-1}\mathbf{v}_b^{s}e_{w,b}^{s}+2\bar{\mathbf{x}}_w^{q}\boldsymbol{\Sigma}_{qr}^{-1}\mathbf{v}_f^{r}\right).$$

Set these derivatives equal to zero, yielding the estimating equations.

8.3.4 Unchanging Group Positions

Another model of possible interest to the researcher is one that hypothesizes that the scores for the group means on the common variates, $e^{q}_{g,a}$, do not change; i.e., they are equal at different occasions, $e^{q}_{g,a} = e^{s}_{g,a} \;\; \forall\, q \neq s$. Let $e_{g,a}$ denote the unchanging score of the g^{th} group for the a^{th} common variate. The likelihood equation for this model is obtained by substituting $e_{g,a}$ for $e^{q}_{g,a}$ in the likelihood equation for CVA/time (orth.) (8.9). The constraints on the $e_{g,a}$ and the manner of handling them are the same as for the $e^{q}_{g,a}$ terms in Section **8.3.3**. The estimating equations for the $e_{g,a}$ are found by taking the derivative of $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $e_{g,f}$ and setting it equal to zero:

$$\frac{\delta l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta e_{h,w}} = -n_h \sum_{q=1}^{t}\sum_{s=1}^{t}\left(\sum_{a=1}^{c} e_{h,a}\mathbf{v}'_w\boldsymbol{\Sigma}^{-1}_{qr}\mathbf{v}_a + \sum_{b=c+1}^{u} e^{q}_{h,b}\mathbf{v}'_w\boldsymbol{\Sigma}^{-1}_{qr}\mathbf{v}^{s}_b\right) + n_h\left(\sum_{q=1}^{t}\sum_{s=1}^{t}\sum_{a=1}^{c}(\bar{\mathbf{x}}^{s}_h - \boldsymbol{\mu}^{q}_0)'\boldsymbol{\Sigma}^{-1}_{qr}\mathbf{v}_w\right).$$

The estimating equations for $\mathbf{v}_i$ and $\mathbf{v}^{q}_i$ are the same as those in the previous section except that $e_{g,a}$ is substituted for $e^{q}_{g,a}$.

8.3.5 Obtaining Estimates

The estimating equations taken together with the constraints result in a system of non-linear equations which can be solved with a Gauss-Newton algorithm, implementing the Marquardt modification where appropriate. The estimates for all parameters are generally solved for simultaneously, as all the estimating equations must be simultaneously true in order to be at a solution. However, the estimation of $\boldsymbol{\mu}_0$ is an exception to this rule as its estimator is a function only of the data, not of any of the other parameter estimates; see equation (8.7).

The algorithm does not guarentee convergence to a local extremum or saddle point. The convergence of the algorithm to a globally optimal solution depends on good starting values. Reasonable starting values can be obtained for the common variates by performing a common principal components on the group means. Starting values for the unique variates can be obtained by performing separate canonical variate analysis or redundancy analyses at each occasion. For models that include both common and unique variates one has to use both methods to obtain starting values. In some cases it may be necessary to try more than one set of starting values. One can examine the matrix of second order partial derivatives of the likelihood function with to the parameters to determine whether a solution is a local maximum, minimum or saddle point. Unfortunately, one can never be certain one has achieved a global solution.

8.3.6 Statistical Inference

Maximum likelihood estimation has under regularity conditions (Wilks 1962) properties which allow one to perform various forms of statistical inference including hypothesis tests and

confidence intervals for parameter estimates. This section discusses how such inference is obtained.

Firstly, define the composite hypothesis and alternative as follows:

$$H_0 = \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$$
$$H_1 = \boldsymbol{\theta} \in \boldsymbol{\Theta}_1,$$

where $\boldsymbol{\Theta}_0 = \boldsymbol{\Theta} - \boldsymbol{\Theta}_1$, $\boldsymbol{\theta}$ is an *r*-dimensional vector of unknown parameters to be estimated, $\boldsymbol{\Theta}$ is an open region in *r*-dimensional Euclidean space and $\boldsymbol{\Theta}_0$ is *q*-dimensional, $q < r$. Then for such composite hypotheses likelihood ratio tests can be based on the asymptotic chi-square distribution of negative two times the log-likelihood ratio, which is denoted by $-2\log\lambda(\mathbf{X})$, where:

$$\lambda(\mathbf{X}) = \frac{\sup\{L(\mathbf{X},\boldsymbol{\theta}): \boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}}{\sup\{L(\mathbf{X},\boldsymbol{\theta}): \boldsymbol{\theta} \in \boldsymbol{\Theta}_1\}},$$

and $\sup\{L(\mathbf{X},\boldsymbol{\theta}): \boldsymbol{\theta} \in \boldsymbol{\Theta}_1\}$ is the maximum likelihood estimate given $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$. For large sample sizes the following is approximately true:

$$-2\log\lambda(\mathbf{X}) \sim X^2_{r-q}.$$

The form of an α-level test is straightforward: reject H_0 if

$$-2\log\lambda(\mathbf{X}) \geq X^2_{r-q}(1-\alpha).$$

One can perform a likelihood ratio test if the set of parameters of the null hypothesis is nested within the set of parameters of the alternative hypothesis. For example, the set of parameters of the common variate hypothesis is nested within the set of parameters of the unique variate hypothesis. Thus one can test the null hypothesis that a given number of variates are common versus the alternative that they are unique.

Next I point out that maximum likelihood estimates are consistent and asymptotically unbiased. Furthermore, they are asymptotically normal with a covariance matrix equal to the inverse of the information matrix. The information matrix is defined as:

$$\mathbf{I}(\boldsymbol{\theta}_0) = \left[-\frac{\delta^2 l(\mathbf{X},\boldsymbol{\theta})}{\delta\theta_i \delta\theta_\varphi}(\boldsymbol{\theta} = \boldsymbol{\theta}_0) \right],$$

though in practice one evaluates the information matrix at the parameter estimates based on the data. Estimates for variances of the parameter estimates can be obtained from $\mathbf{I}^{-1}(\boldsymbol{\theta}_0)$, enabling one to make confidence intervals and simple hypothesis tests for the parameters. Let θ_i be a parameter in the model, and $\hat{\theta}_i$ be its estimate. Then a $100 \times (1-\alpha)\%$ confidence interval for θ_i would be $\hat{\theta}_i \pm \sqrt{c_{ii}} Z_{(1-\alpha/2)}$, where c_{ii} is the i^{th} diagonal element of $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})$ and $Z_{(1-\alpha/2)}$ is the value of the standard normal variate corresponding to a cumulative probability of $1-\alpha/2$.

An α-level hypothesis test for testing the following hypothesis that a single parameter, θ_i, is zero

$$H_0: \theta_i = 0 \quad \text{vs.} \quad H_1: \theta_i \neq 0,$$

would be to reject H_0 if $\hat{\theta}_i \geq c_{ii} Z_{(1-\alpha/2)}$ or $\hat{\theta}_i \leq -c_{ii} Z_{(1-\alpha/2)}$. One can test more complex hypotheses of the form

$$H_0: \mathbf{C}\boldsymbol{\theta} = 0 \quad \text{vs.} \quad H_1: \mathbf{C}\boldsymbol{\theta} \neq 0$$

where $\boldsymbol{\theta}$ is a $q \times 1$ the vector of parameters and $\mathbf{C}$ is an $r \times q$ matrix of rank r with the Wald (Wald 1945) statistic W, where $W = (\mathbf{C}\hat{\boldsymbol{\theta}})'((\mathbf{C}\hat{\boldsymbol{\theta}})'\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})\mathbf{C}\hat{\boldsymbol{\theta}})^{-1}\mathbf{C}\hat{\boldsymbol{\theta}}$.

8.4 SIMULATIONS

In this section I present a simulated study of the methods developed in Section **8.3**. Simulated data have the advantage that one knows the true structure of the data, hence one can ascertain whether the method is successful in discerning that structure. In particular I attempt to answer the following questions: Is the method obtaining the true parameters? When the null hypothesis is true, is the likelihood ratio test rejecting at the specified alpha-level? Is the test statistic distributed as predicted in the theory? And, does the simulated variance-covariance matrix for the parameter estimates converge to the theoretical asymptotic variance-covariance based on the inverse of the information matrix?

These simulated data were generated using a pseudo-multivariate normal distribution in SAS's Proc IML. The generated multivariate normal data served as residuals which were added to a mean structure specified to be a one common variate model. The actual parameters were chosen arbitrarily. The simulated data consisted of measurements of three variables at three occasions for four groups. The SAS code used to generate the simulation and obtain the estimates is given in Appendix Six. There were three simulations performed. One simulation generated 1,000 datasets of a sample size of 100, or a sample of 25 for each of the four groups. The other two simulations generated 5,000 datasets each of samples of sizes of 400 and 4,000, or 100 and 1,000 for each of the four groups. There were fewer simulations for samples of size 100 (four groups of 25), because the algorithm to solve the estimating equations took prohibitively longer to converge.

The information matrix was calculated using the computer package "Mathematica" (Wolfram 1991). The Mathematica code is given in Appendix Seven. It is clear from equation (8.6) that the theoretical estimate for $\boldsymbol{\mu}_0$ based on the information matrix is independent of the estimates for the rest of the parameters because the derivative of (8.6) with respect to any parameter other than $\boldsymbol{\mu}_0$ will be zero. Further, it is straightforward to show that the derivatives of $\frac{\delta/(\mathbf{X}/\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta\boldsymbol{\Sigma}}$ with respect to $\mathbf{v}$ and $\mathbf{e}_g$ will be zero. Thus $\hat{\boldsymbol{\Sigma}}$ is independent (asymptotically) of the estimates of $\mathbf{v}$ and $\mathbf{e}_g$. Hence **Table 8.2** and Appendix Eight only show the calculated variances of the parameters $\mathbf{v}$ and $\mathbf{e}_g$.

The model for the group means used to simulate the data is shown below. The terms are defined as in Section **8.3**:

$$\boldsymbol{\mu}_g = \boldsymbol{\mu}_0 + \mathbf{e}_g \otimes \mathbf{v},$$

where $g = 1,\ldots,m$ and $\boldsymbol{\mu}_g$ is a 9×1 vector. The parameter values for **v**, the common variate, are:

$$\mathbf{v} = \left[0.5, 0.5, 0.707\right]'.$$

Let **E** be the matrix whose columns are $\mathbf{e}_g$, $g = 1,\ldots,m$, then:

$$\mathbf{E} = \begin{bmatrix} 1 & 0.5 & -.5 & -1 \\ 0 & 1 & 0 & -1 \\ 1 & -.5 & .5 & -1 \end{bmatrix}.$$

The errors have the covariance matrix **C**:

$$\mathbf{C} = \begin{bmatrix} 4.8 & 2.1 & 1.0 & 2.4 & 1.05 & 0.5 & 1.2 & 0.525 & 0.25 \\ 2.1 & 3.3 & 1.4 & 1.05 & 1.65 & 0.7 & 0.525 & 0.825 & 0.35 \\ 1.0 & 1.4 & 2.9 & 0.5 & 0.7 & 1.45 & 0.25 & 0.35 & 0.725 \\ 2.4 & 1.05 & 0.5 & 4.8 & 2.1 & 1.0 & 2.4 & 1.05 & 0..5 \\ 1.05 & 1.65 & 0.7 & 2.1 & 3.3 & 1.4 & 1.05 & 1.65 & 0.7 \\ 0.5 & 0.7 & 1.45 & 1.0 & 1.4 & 2.9 & 0.5 & 0.7 & 1.45 \\ 1.2 & 0.525 & 0.25 & 2.4 & 1.05 & 0.5 & 4.8 & 2.1 & 1.0 \\ 0.525 & 0.825 & 0.35 & 1.05 & 1.65 & 0.7 & 2.1 & 3.3 & 1.4 \\ 0.25 & 0.35 & 0.725 & 0.5 & 0.7 & 1.45 & 1.0 & 1.4 & 2.9 \end{bmatrix}.$$

The one common variate model and the one unique variate model were fit for each dataset. The one unique variate model is as follows:

$$\boldsymbol{\mu}_g = \boldsymbol{\mu}_o + \begin{bmatrix} e_g^1 \mathbf{v}^1 \\ e_g^2 \mathbf{v}^2 \\ e_g^3 \mathbf{v}^3 \end{bmatrix},$$

where by definition $\sum_{g=1}^{m} n_g \mathbf{e}_g = 0$, $\mathbf{v}^{1\prime}\mathbf{v}^1 = 1$, $\mathbf{v}^{2\prime}\mathbf{v}^2 = 1$ and $\mathbf{v}^{3\prime}\mathbf{v}^3 = 1$. Note that for 108 of the 1,000 datasets with sample size of 100 that the algorithm for the unique variates converged to an estimate that was clearly not a global maximum. I determined that an estimate for the unique variates model was not a global maximum if it had a lower likelihood than the estimate of the common variate model. Since the unique variates model has more free parameters than the common variates model the likelihood of its estimate should be greater if it is at the global maximum. The 108 runs were not included in the tables. In only one of the 5,000 runs with the sample size of 400 did this problem occur, and in none of the runs with the sample size of 4,000.

Table 8.1 shows the means of the parameter estimates based on the simulations. At the left are the true parameter values. All three sets of estimates are in the area of the true parameter values. However, the estimates based on the larger samples are clearly closer to the true values.

Table 8.1 Parameter Estimates

Para-meters	True Parameter Values	Estimates for Sample of 4000	Estimates for Sample of 400	Estimates for Sample of 100
v_1	0.5	0.4994	0.4967	0.4911
v_2	0.5	0.4997	0.4981	0.4910
v_3	0.7071	0.7073	0.7058	0.6940
e_1^1	1.0	0.9995	1.0048	1.0368
e_1^2	0.0	0.0006	0.0001	0.0290
e_1^3	1.0	0.9989	1.0018	1.0359
e_2^1	0.5	0.4999	0.4990	0.4892
e_2^2	1.0	0.9989	1.0024	1.0078
e_2^3	-0.5	-0.5000	-0.5057	-0.5181
e_3^1	-0.5	-0.4997	-0.5029	-0.4901
e_3^2	0.0	0.0003	0.0016	0.0039
e_3^3	0.5	0.5007	0.5033	0.5209
e_4^1	-1.0	-0.9997	-1.0009	-1.0360
e_4^2	-1.0	-0.9998	-1.0007	-1.0407
e_4^3	-1.0	-0.9996	-0.9994	-1.0389

Table 8.2 presents the observed variances based on the simulations for comparison with the theoretical variances based on the inverse of the information matrix. The theoretical variances are calculated assuming a sample of size 400. Not presented in **Table 8.2** are the theoretical variances for a sample of size 4,000, which are $1/10^{\text{th}}$ that of the variance for $n = 400$, and the theoretical variances for a sample size of 100, which are four times those of a sample size of 400. One sees that the estimated variances of the parameter estimates are close to the theoretical variances for the sample sizes of 400 and 4,000, but not for a sample size of 100. Examination of the full variance-covariance matrices would reveal the same pattern. The full theoretical variance-covariance matrix is presented in Appendix Eight, while the full variance-covariance matrices based on the estimates from the simulations is presented in Appendix Nine.

Table 8.2 Theoretical and Observed Variances for the Parameter Estimates

Para-meters	Theoretical Values For Sample of 400	Observed for Sample of 4000	Observed for Sample of 400	Observed for Sample of 100
v_1	0.002998	0.000314	0.00326	0.046457
v_2	0.001589	0.000158	0.001802	0.037871
v_3	0.001928	0.0001964	0.00215	0.06856
e_1^1	0.040908	0.00427	0.042776	0.325786
e_1^2	0.039628	0.003878	0.040881	0.185901
e_1^3	0.040908	0.00409	0.042394	0.306227
e_2^1	0.039948	0.004037	0.041292	0.204403
e_2^2	0.040908	0.004177	0.041389	0.324521
e_2^3	0.039948	0.004072	0.041348	0.218552
e_3^1	0.039948	0.004059	0.041273	0.219736
e_3^2	0.039628	0.003866	0.042012	0.197513
e_3^3	0.039948	0.003925	0.041697	0.208513
e_4^1	0.040908	0.004207	0.04335	0.30024
e_4^2	0.040908	0.004036	0.041934	0.308235
e_4^3	0.040908	0.004189	0.04264	0.324294

The next results of interest are the distributions of the likelihood ratio test statistics. **Table 8.3** shows the mean and variance of the test statistics, and the proportion that are greater than the 90th, 95th and 99th percentile of the cumulative distribution of a chi-square with four degrees of freedom. One sees for sample sizes of 4,000 that the mean and variance of the likelihood ratio test statistic are very close to the theoretical values. Also, the proportions of the observed test statistics that are above the $\mathbf{X}^2_{(1-\alpha)}$ are close to what is predicted by the theory. For the datasets with sizes of 400, the variance of the test statistic is larger at about 9, though the proportions are roughly correct. For the data with sample sizes of 100 the distribution of the test statistic deviates more noticeably from the theoretical. This is not surprising as one has only 100 observations with which to estimate a total of 61 parameters (one must also estimate the elements of Σ).

Table 8.3 Theoretical and Observed Values of the Likelihood Ratio Test Statistic

Labels	Mean	Variance	Proportion over 90th Percentile	Proportion over 95th Percentile	Proportion over 99th Percentile
Theoretical	4.0	8.0	0.1	0.05	0.01
Sample of 4000	4.005	8.036	0.099	0.0506	0.0099
Sample of 400	4.212	9.076	0.117	0.0614	0.0146
Sample of 100	3.904	7.000	0.084	0.0426	0.0034

In summary, these simulations confirm the basic methodology. First, they confirm that the algorithm and estimating equations yield correct estimates. The estimates are correct in the sense that they are approaching the true parameter values and that the observed variances of the parameter estimates are close to the theoretical variances, at least for the larger sample sizes. Second, they confirm the correctness of the hypothesis tests for the larger samples. With samples sizes of 400 or 4,000 the test statistic is distributed close to the theoretical distribution under the null hypothesis.

8.5 CVA/TIME - UNCORRELATED VARIATES

In this section I develop an alternative model for analyzing group structure with longitudinal multivariate data, CVA/time with uncorrelated canonical variates. Where CVA/time (orth.) hypothesizes that the group means lie in the space of orthogonal variates, CVA/time (unc.) hypothesizes that the group means lie in the space of uncorrelated canonical variates. CVA/time (unc.), unlike CVA/time (orth.), represents a true generalization of CVA. Furthermore, it is equivalent to Campbell and Tomenson's model under the special circumstance that the covariances of the variables between different occasions are zero.

CVA/time (unc.) follows much of the logic of that for CVA/time (orth.). The CVA/time (unc.) model hypothesizes common variates, unique variates, and group positions. It differs from CVA/time (orth.) in that the model for the means now involves the covariance matrix. Furthermore, as is seen in equation (8.1), a common within-groups covariance structure for each occasion, Σ_w, is required, which will necessitate a certain structure for Σ. The estimation of the structure of Σ will encompass a good part of this section. On the other hand, the development of the estimation for μ_0 is identical to that in Section **8.3.3**, as is the discussion about obtaining estimates in Section **8.3.5** and that of statistical inference in Section **8.3.6**. Hence these topics need not be further addressed with respect to CVA/time (unc.).

It is useful to give a simple example of a common variate model. Assume the positions of two group means can be plotted on a variate in the transformed space which is common over two occasions, and assume the positions of the group means change over time. **Figure 8.3** shows the positions at the first occasion, and **Figure 8.4** shows the positions at the second occasion.

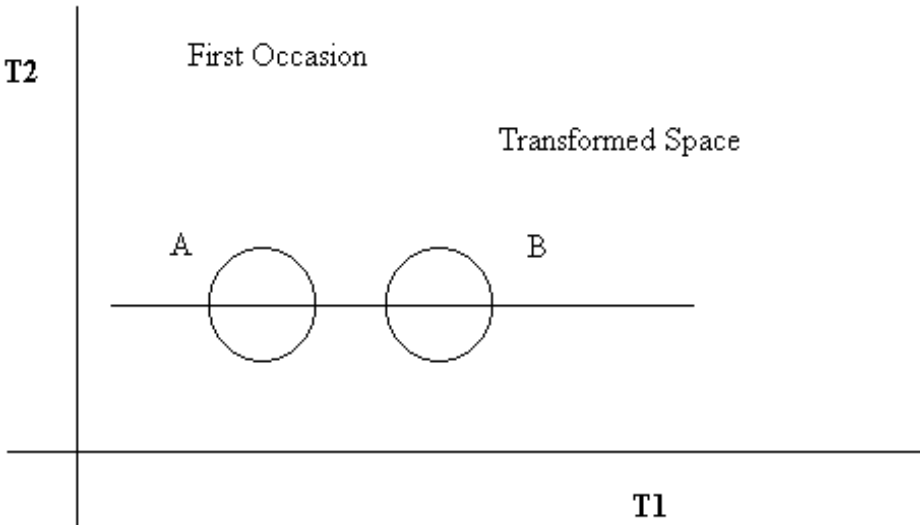

Figure 8.3

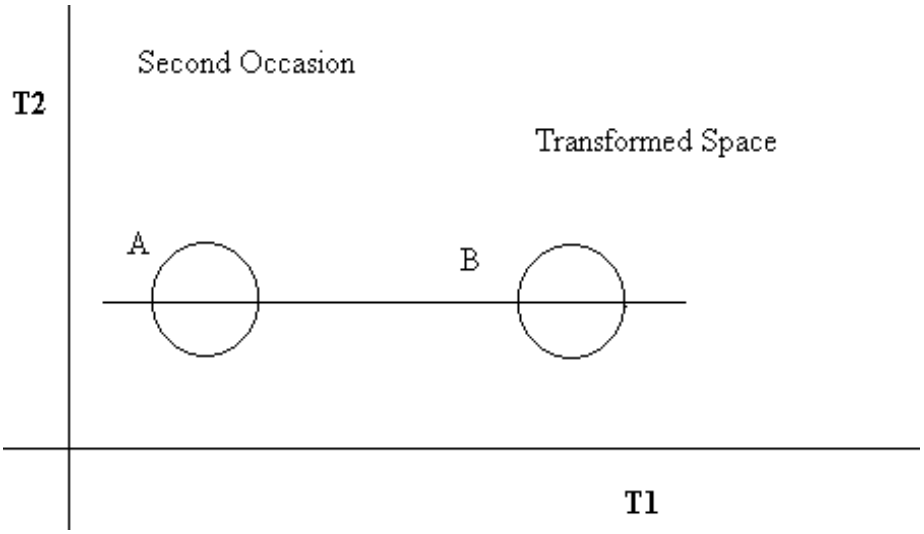

Figure 8.4

8.5.1 The CVA/Time Model with Uncorrelated Variates

CVA/time (unc.) model is as follows:

$$\boldsymbol{\mu}_{\mathrm{g}} = \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_1 \otimes \mathbf{e}_{\mathrm{g},1} + \ldots + \boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_c \otimes \mathbf{e}_{\mathrm{g},c} + \left(\boldsymbol{\Sigma}_{\mathrm{w}} \otimes \mathbf{I}_{t\times t}\right)\begin{bmatrix} \mathrm{e}^1_{\mathrm{g},c+1}\mathbf{v}^1_{c+1} \\ \vdots \\ \mathrm{e}^t_{\mathrm{g},c+1}\mathbf{v}^t_{c+1} \end{bmatrix} + \cdots + \left(\boldsymbol{\Sigma}_{\mathrm{w}} \otimes \mathbf{I}_{t\times t}\right)\begin{bmatrix} \mathrm{e}^1_{\mathrm{g},u}\mathbf{v}^1_{u} \\ \vdots \\ \mathrm{e}^t_{\mathrm{g},u}\mathbf{v}^t_{u} \end{bmatrix}, \quad (8.11)$$

where $\boldsymbol{\mu}_{\mathrm{g}}$ is a $pt \times 1$ vector of means for the g^{th} group, $\boldsymbol{\mu}_0$ is a $pt \times 1$ vector of overall means, $\mathbf{v}_{\mathrm{i}}$ are c $p \times 1$ vectors of common variates, $\mathbf{v}^{\mathrm{k}}_{\mathrm{j}}$ are $t(u - c)$ $p \times 1$ vectors of unique variates, $\mathrm{e}^t_{\mathrm{g,i}}$

is the score for the g^{th} group mean on the i^{th} canonical variate at the q^{th} occasion, and $\mathbf{e}_{g,i}$ is the $t \times 1$ vector whose elements are $e^{q}_{g,i}$.

Note the constraints on the parameters. Those for the positions of the group means are the same as for CVA/time (orth.), $\sum_{g=1}^{m} e^{q}_{g,i} = 0$ for all q,i. This constraint reflects the centering by $\boldsymbol{\mu}_0$. The constraints on the variates differ from those for CVA/time (orth.) as the variates are constrained to be mutually uncorrelated, not orthogonal:

$$\mathbf{V}'_{com}\boldsymbol{\Sigma}_w\mathbf{V}_{com} = \mathbf{I}_{c\times c},$$

where $\mathbf{V}_{com}$ is the matrix whose c columns are the common variates. Furthermore, within each set of unique variates for each occasion the variates are mutually uncorrelated, that is:

$$\mathbf{V}^{q\prime}\boldsymbol{\Sigma}_w\mathbf{V}^{q} = \mathbf{I}_{(u-c)\times(u-c)},$$

where $\mathbf{V}^{q}$ is the matrix whose columns are the unique variates for the q^{th} occasion, $q = 1,\ldots,t$. Finally, each variate in each set of unique variates is orthogonal with each common variate. Thus:

$$\mathbf{V}'_{com}\boldsymbol{\Sigma}_w\mathbf{V}^{q} = [0]_{c\times(u-c)},$$

for $q = 1,\ldots,t$.

To conclude this section I point out that a model that hypothesizes unique variates at each occasion is not equivalent to performing separate canonical variate analyses at each occasion. A unique variates model estimates the variates at each occasion as a part of a larger model. It may be superior to a separate analysis of each occasion because it models the covariances between the measurements made at different occasions.

8.5.2 Estimating the Within-Groups Covariance Matrix

CVA/time (unc.) necessitates a particular structure to $\boldsymbol{\Sigma}$. The estimation of this structure is discussed in this section. Also briefly discussed at the conclusion is how to estimate this same structure for $\boldsymbol{\Sigma}$ if one should hypothesize orthogonal canonical variates.

As stated previously, the logic of CVA/time (unc.) requires within-groups covariance matrices that are equal over time or at minimum proportional over time. However, given this assumption it is reasonable to make the further assumption that the covariances matrices between measurements at different occasions are proportional. Indeed, this additional assumption proves to be necessary to obtain workable estimating equations. Hence the structure given in (8.12) is assumed:

$$\boldsymbol{\Sigma} = \begin{bmatrix} a_{11}\boldsymbol{\Sigma}_w & a_{12}\boldsymbol{\Sigma}_w & \cdots & a_{1t}\boldsymbol{\Sigma}_w \\ a_{21}\boldsymbol{\Sigma}_w & a_{22}\boldsymbol{\Sigma}_w & \cdots & a_{2t}\boldsymbol{\Sigma}_w \\ \vdots & \vdots & \ddots & \vdots \\ a_{t1}\boldsymbol{\Sigma}_w & a_{t2}\boldsymbol{\Sigma}_w & \cdots & a_{tt}\boldsymbol{\Sigma}_w \end{bmatrix} = \mathbf{A}\otimes\boldsymbol{\Sigma}_w, \qquad (8.12)$$

where $\mathbf{A} = [a_{ij}]$, is a $t \times t$ positive definite matrix, and $\boldsymbol{\Sigma}_w$ is a $p \times p$ matrix that is proportional to the within-groups variance-covariance matrix at each occasion. The matrix **A** will be referred to as the matrix of proportionality constants.

The estimating equations for **A** are derived in Section **8.5.3**. I proceed by deriving the estimating equations for $\boldsymbol{\Sigma}_w$. The estimation of $\boldsymbol{\Sigma}_w$ is complicated by the fact that the group means, $\boldsymbol{\mu}_g$, are now functions of $\boldsymbol{\Sigma}_w$, which is seen in (8.11). The log-likelihood for the model is:

$$l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{-np}{2}\log(2\pi) - \frac{nt}{2}\log|\boldsymbol{\Sigma}_w| - \frac{np}{2}\log|\mathbf{A}|$$
$$-\tfrac{1}{2}\left(\sum_{g=1}^{m}\sum_{i=1}^{n_g}\sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s](\mathbf{x}_{iq}-\bar{\mathbf{x}}_{gq})'\boldsymbol{\Sigma}_w^{-1}(\mathbf{x}_{is}-\bar{\mathbf{x}}_{gs}) + \sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s](\bar{\mathbf{x}}_{gq}-\boldsymbol{\Sigma}_w\boldsymbol{\mu}_{gq})'\boldsymbol{\Sigma}_w^{-1}(\bar{\mathbf{x}}_{gs}-\boldsymbol{\Sigma}_w\boldsymbol{\mu}_{gs})\right).$$

Taking the derivative of the log-likelihood with respect to $\boldsymbol{\Sigma}_w$ yields:

$$\frac{\delta l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta \boldsymbol{\Sigma}_w} = -nt\boldsymbol{\Sigma}_w^{-1} + \frac{nt}{2}\operatorname{diag}(\boldsymbol{\Sigma}_w^{-1}) + \sum_{g=1}^{m}\sum_{i=1}^{n_g}\sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\Sigma}_w^{-1}(\mathbf{x}_{is}-\bar{\mathbf{x}}_{gs})(\mathbf{x}_{iq}-\bar{\mathbf{x}}_{gq})'\boldsymbol{\Sigma}_w^{-1}$$
$$-\tfrac{1}{2}\operatorname{diag}\left(\sum_{g=1}^{m}\sum_{i=1}^{n_g}\sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\Sigma}_w^{-1}(\mathbf{x}_{is}-\bar{\mathbf{x}}_{gs})(\mathbf{x}_{iq}-\bar{\mathbf{x}}_{gq})'\boldsymbol{\Sigma}_w^{-1}\right)$$
$$+\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\Sigma}_w^{-1}\bar{\mathbf{x}}_{gs}\bar{\mathbf{x}}_{gq}'\boldsymbol{\Sigma}^{-1} - \tfrac{1}{2}\operatorname{diag}\left(\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{r=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\Sigma}_w^{-1}\bar{\mathbf{x}}_{gs}\bar{\mathbf{x}}_{gq}'\boldsymbol{\Sigma}_w^{-1}\right)$$
$$-\tfrac{1}{2}\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\mu}_{gr}\boldsymbol{\mu}_{gq}'.$$

To obtain the estimating equations one sets these derivatives equal to a matrix of zeros. To put the equations in a simpler form first multiply through by $\boldsymbol{\Sigma}_w$. Then note that the last term above equals the following expression:

$$-\tfrac{1}{2}\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\mu}_{gs}\boldsymbol{\mu}_{gq}' =$$
$$-\tfrac{1}{2}\left(\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\Sigma}_w\boldsymbol{\mu}_{gs}\boldsymbol{\mu}_{gq}'\boldsymbol{\Sigma}_w + \operatorname{diag}\left(\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\Sigma}_w\boldsymbol{\mu}_{gs}\boldsymbol{\mu}_{gq}'\boldsymbol{\Sigma}_w\right)\right)$$
$$+\tfrac{1}{4}\operatorname{diag}\left(\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\Sigma}_w\boldsymbol{\mu}_{gs}\boldsymbol{\mu}_{gq}'\boldsymbol{\Sigma}_w + \operatorname{diag}\left(\sum_{g=1}^{m} n_g \sum_{q=1}^{t}\sum_{s=1}^{t}\mathbf{A}^{-1}[q,s]\boldsymbol{\Sigma}_w\boldsymbol{\mu}_{gs}\boldsymbol{\mu}_{gq}'\boldsymbol{\Sigma}_w\right)\right).$$

Then the estimating equations are solved when equation (8.13) below is solved. Note that $\boldsymbol{\Sigma}_w$ is on both sides of the equations.

$$nt\boldsymbol{\Sigma}_{\mathrm{w}} = \sum_{\mathrm{g}=1}^{m}\sum_{\mathrm{i}=1}^{n_{\mathrm{g}}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}](\mathbf{x}_{\mathrm{is}}-\bar{\mathbf{x}}_{\mathrm{gs}})(\mathbf{x}_{\mathrm{iq}}-\bar{\mathbf{x}}_{\mathrm{gq}})' + \sum_{\mathrm{g}=1}^{m} n_{\mathrm{g}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}]\bar{\mathbf{x}}_{\mathrm{gr}}\bar{\mathbf{x}}'_{\mathrm{gq}}$$

$$-\tfrac{1}{2}\sum_{\mathrm{g}=1}^{m} n_{\mathrm{g}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}]\boldsymbol{\Sigma}_{\mathrm{w}}\boldsymbol{\mu}_{\mathrm{gs}}\boldsymbol{\mu}'_{\mathrm{gq}}\boldsymbol{\Sigma}_{\mathrm{w}} - \tfrac{1}{2}\mathrm{diag}\left(\sum_{\mathrm{g}=1}^{m} n_{\mathrm{g}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}]\boldsymbol{\Sigma}_{\mathrm{w}}\boldsymbol{\mu}_{\mathrm{gs}}\boldsymbol{\mu}'_{\mathrm{gq}}\boldsymbol{\Sigma}_{\mathrm{w}}\right). \quad (8.13)$$

If one wants to estimate the covariance structure assuming orthogonal variates, that is, assuming the CVA/time (orth.) model, then the estimating equations are solved when $\boldsymbol{\Sigma}_{\mathrm{w}}$ equals the expression below. The derivation of this equation involves a modification of the derivation of equation (8.13).

$$\boldsymbol{\Sigma}_{\mathrm{w}} = (nt)^{-1}\sum_{\mathrm{g}=1}^{m}\sum_{\mathrm{i}=1}^{n_{\mathrm{g}}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\mathrm{ai}_{\mathrm{qr}}(\mathbf{x}_{\mathrm{is}}-\bar{\mathbf{x}}_{\mathrm{gs}})(\mathbf{x}_{\mathrm{iq}}-\bar{\mathbf{x}}_{\mathrm{gq}})' + \sum_{\mathrm{g}=1}^{m} n_{\mathrm{g}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\mathrm{ai}_{\mathrm{qs}}(\bar{\mathbf{x}}_{\mathrm{gs}}-\boldsymbol{\mu}_{\mathrm{gs}})(\bar{\mathbf{x}}_{\mathrm{gq}}-\boldsymbol{\mu}_{\mathrm{gq}})'.$$

8.5.3 Estimating the Matrix of Proportionality Constants (A)

To solve for **A** it will be necessary to define some new notation. Let $\mathbf{x}_{\mathrm{i(b)}}$ be the $t \times 1$ vector whose r^{th} element is the b^{th} variable of the i^{th} observation at the r^{th} occasion. In other words, $\mathbf{x}_{\mathrm{i(b)}}$ is composed of the measurements made over the t occasions of the b^{th} variable for the i^{th} subject. Let $\bar{\mathbf{x}}_{\mathrm{g(b)}}$ be the analogous for the g^{th} group mean. Then the log-likelihood becomes as follows:

$$l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{-npt}{2}\log(2\pi) - \frac{nt}{2}\log|\boldsymbol{\Sigma}_{\mathrm{w}}| - \frac{np}{2}\log|\mathbf{A}|$$

$$-\tfrac{1}{2}\left(\sum_{\mathrm{g}=1}^{m}\sum_{\mathrm{i}=1}^{n_{\mathrm{g}}}\sum_{\mathrm{j}=1}^{p}\sum_{\mathrm{v}=1}^{p}\boldsymbol{\Sigma}_{\mathrm{w}}^{-1}[\mathrm{j},\mathrm{v}](\mathbf{x}_{\mathrm{i(j)}}-\bar{\mathbf{x}}_{\mathrm{g(j)}})'\mathbf{A}^{-1}(\mathbf{x}_{\mathrm{i(v)}}-\bar{\mathbf{x}}_{\mathrm{g(v)}})' + \sum_{\mathrm{g}=1}^{m} n_{\mathrm{g}}\sum_{\mathrm{j}=1}^{p}\sum_{\mathrm{v}=1}^{p}\boldsymbol{\Sigma}_{\mathrm{w}}^{-1}[\mathrm{j},\mathrm{v}](\bar{\mathbf{x}}_{\mathrm{g(j)}}-\boldsymbol{\mu}_{\mathrm{g(j)}})'\mathbf{A}^{-1}(\bar{\mathbf{x}}_{\mathrm{g(v)}}-\boldsymbol{\mu}_{\mathrm{g(v)}})\right)$$

The **A** matrix needs to be restrained for scale. The simplest way to do this is to set

$$\mathrm{trace}(\mathbf{A}) - p = 0,$$

which constrains the within-groups covariance matrix to be proportional over time. The constraint with Lagrangian multiplier is $\lambda(\mathrm{trace}(\mathbf{A}) - p)$, where λ is the Lagrangian multiplier. To obtain the estimating equations for **A** take the derivative with respect to **A** of the log-likelihood modified by the constraint with the Lagrangian multipliers yielding (8.14). Set this equal to a $t \times t$ matrix of zeros to get the estimating equations.

$$\frac{\delta l^{*}(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta \mathbf{A}} = -np\mathbf{A}^{-1} + np\,\mathrm{diag}(\mathbf{A}^{-1}) + \sum_{g=1}^{m}\sum_{i=1}^{n_g}\sum_{j=1}^{p}\sum_{v=1}^{p}\boldsymbol{\Sigma}_w^{-1}[j,v]\mathbf{A}^{-1}(\mathbf{x}_{i(j)} - \bar{\mathbf{x}}_{g(j)})(\mathbf{x}_{i(v)} - \bar{\mathbf{x}}_{g(v)})'\mathbf{A}^{-1}$$

$$-\tfrac{1}{2}\,\mathrm{diag}\left(\sum_{g=1}^{m}\sum_{i=1}^{n_g}\sum_{j=1}^{p}\sum_{v=1}^{p}\boldsymbol{\Sigma}_w^{-1}[j,v]\mathbf{A}^{-1}(\mathbf{x}_{i(j)} - \bar{\mathbf{x}}_{g(j)})(\mathbf{x}_{i(v)} - \bar{\mathbf{x}}_{g(v)})'\mathbf{A}^{-1}\right)$$

$$+\sum_{g=1}^{m} n_g \sum_{j=1}^{p}\sum_{v=1}^{p}\boldsymbol{\Sigma}_w^{-1}[j,v]\mathbf{A}^{-1}(\bar{\mathbf{x}}_{g(j)} - \boldsymbol{\mu}_{g(j)})(\bar{\mathbf{x}}_{g(v)} - \boldsymbol{\mu}_{g(v)})'\mathbf{A}^{-1}$$

$$-\tfrac{1}{2}\,\mathrm{diag}\left(\sum_{g=1}^{m} n_g \sum_{j=1}^{p}\sum_{v=1}^{p}\boldsymbol{\Sigma}_w^{-1}[j,v]\mathbf{A}^{-1}(\bar{\mathbf{x}}_{g(j)} - \boldsymbol{\mu}_{g(j)})(\bar{\mathbf{x}}_{g(v)} - \boldsymbol{\mu}_{g(v)})'\mathbf{A}^{-1}\right) - \lambda\mathbf{I}_{p\times p}. \qquad (8.14)$$

If one wants to hypothesize that the within-groups covariance matrices are constant at each occasion, then one would restrict **A** to have ones as its diagonal elements. Let $\mathbf{h}_{(i)}$ denote a vector that has a one as its i^{th} element and zeros as the rest; i.e. $\mathbf{h}_{(i)}[j] = 1$ if $j = i$, else $\mathbf{h}_{(i)}[j] = 0$ if $j \neq i$. Then this restriction is equivalent to requiring:

$$\mathbf{h}'_{(i)}\mathbf{A}\mathbf{h}_{(i)} = 1 \text{ for } i = 1,\ldots,p.$$

The *p* constraints with *p* Lagrangian multipliers are as follows:

$$\sum_{i=1}^{p}\lambda_i\left(\mathbf{h}'_{(i)}\mathbf{A}\mathbf{h}_{(i)} - 1\right) = 0.$$

When differentiated with respect to **A** this expression yields a diagonal matrix **T** whose i^{th} diagonal element is λ_i. Hence the estimating equations for solving for **A** with this method are the same as in (8.14), except that one substitutes **T** for $\lambda\mathbf{I}_{p\times p}$.

8.5.4 Hypothesis Test for the Simple Structure of the Covariance Matrix ($\boldsymbol{\Sigma}$)

One may wish to test the hypothesis of a simple structure for $\boldsymbol{\Sigma}$, that is $H_0: \boldsymbol{\Sigma} = \mathbf{A} \otimes \boldsymbol{\Sigma}_w$, versus the alternative: $H_1: \boldsymbol{\Sigma} \neq \mathbf{A} \otimes \boldsymbol{\Sigma}_w$. Such a test can be performed in two contexts. The first is more consistent with the rest of Section **8.5** but may not be practical. One assumes a specific mean structure, obtains estimates for both the structured and unstructured $\boldsymbol{\Sigma}$, and then performs a likelihood ratio test based on those estimates. The difficulty with this approach is that one usually does not know the means model. This is particularly troublesome since the estimate of $\boldsymbol{\Sigma}_w$ is sensitive to misspecification of the means model, as is seen in equations (8.8) and (8.13). Indeed, a more robust estimate of $\boldsymbol{\Sigma}_w$ may be desirable even if one does not intend to perform the hypothesis test for simple structure.

One can obtain such a robust approach by assuming a saturated model for $\boldsymbol{\mu}_g$; i.e., let $\bar{\mathbf{x}}_g$ be the estimate for $\boldsymbol{\mu}_g$. The log-likelihood for such a model is obtained by substituting the model for the structure of $\boldsymbol{\Sigma}$ (8.12) into the likelihood equation in its general form (8.4), yielding:

$$l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{-np}{2}\log(2\pi) - \frac{nt}{2}\log|\boldsymbol{\Sigma}_{\mathrm{w}}| - \frac{np}{2}\log|\mathbf{A}|$$
$$-\tfrac{1}{2}\left(\sum_{\mathrm{g}=1}^{m}\sum_{\mathrm{i}=1}^{n_{\mathrm{g}}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}](\mathbf{x}_{\mathrm{iq}}-\bar{\mathbf{x}}_{\mathrm{gq}})'\boldsymbol{\Sigma}_{\mathrm{w}}^{-1}(\mathbf{x}_{\mathrm{is}}-\bar{\mathbf{x}}_{\mathrm{gs}}) + \sum_{\mathrm{g}=1}^{m}n_{\mathrm{g}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}](\bar{\mathbf{x}}_{\mathrm{gq}}-\boldsymbol{\mu}_{\mathrm{gq}})'\boldsymbol{\Sigma}_{\mathrm{w}}^{-1}(\bar{\mathbf{x}}_{\mathrm{gs}}-\boldsymbol{\mu}_{\mathrm{gs}})\right).$$

The estimating equations that result are seen to be reasonable. The estimate of the unconstrained variance-covariance reduces to **S**; see (8.8). The derivation of the estimate for the structured variance-covariance matrix is similar to, though simpler than, the derivation in Section **8.5.2**, and yields:

$$\boldsymbol{\Sigma}_{\mathrm{w}} = (nt)^{-1}\sum_{\mathrm{g}=1}^{m}\sum_{\mathrm{i}=1}^{n_{\mathrm{g}}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}](\mathbf{x}_{\mathrm{is}}-\bar{\mathbf{x}}_{\mathrm{gs}})(\mathbf{x}_{\mathrm{iq}}-\bar{\mathbf{x}}_{\mathrm{gq}})',$$

where the matrix **A** is estimated as in Section **8.5.3**. This estimate for $\boldsymbol{\Sigma}_{\mathrm{w}}$ is just the sum of the $p \times p$ submatrices of **S** weighted by the appropriate element of $\mathbf{A}^{-1}$.

A reservation needs to be made about this hypothesis test. It is possible that despite rejecting the null hypothesis of structure for $\boldsymbol{\Sigma}$, a researcher may conclude the deviations from this structure are not of practical significance. Instead, the researcher may prefer to assume a simple structure for $\boldsymbol{\Sigma}$ to obtain a (crude) scale invariance or to reduce the number of parameters that need to be estimated.

8.5.5 Estimating the Canonical Variates and the Group Scores

Next I develop the estimating equations for the canonical variates, $\mathbf{v}_{\mathrm{i}}$ and $\mathbf{v}_{\mathrm{i}}^{\mathrm{q}}$, and the group scores $\mathbf{e}_{\mathrm{g,i}}$. $\boldsymbol{\mu}_{\mathrm{o}}$ is estimated as in equation (8.8). Denote the log-likelihood by $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ and the terms which include neither $\mathbf{v}_{\mathrm{i}}$, $\mathbf{v}_{\mathrm{i}}^{\mathrm{q}}$ nor $\mathbf{e}_{\mathrm{g,i}}$ by C. Then:

$$l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathrm{C} - \tfrac{1}{2}\sum_{\mathrm{g}=1}^{m}n_{\mathrm{g}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\left(\sum_{\mathrm{a}=1}^{c}\sum_{\mathrm{b}=1}^{c}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}]e_{\mathrm{g,a}}^{\mathrm{q}}\mathbf{v}_{\mathrm{a}}'\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{\mathrm{b}}e_{\mathrm{g,b}}^{\mathrm{s}} + 2\sum_{\mathrm{a}=1}^{c}\sum_{\mathrm{b}=c+1}^{u}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}]e_{\mathrm{g,a}}^{\mathrm{q}}\mathbf{v}_{\mathrm{a}}'\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{\mathrm{b}}^{\mathrm{s}}e_{\mathrm{g,b}}^{\mathrm{s}}\right)$$
$$+\sum_{\mathrm{a}=c+1}^{u}\sum_{\mathrm{b}=c+1}^{u}\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}]e_{\mathrm{g,a}}^{\mathrm{q}}\mathbf{v}_{\mathrm{a}}^{\mathrm{q}}\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{\mathrm{b}}^{\mathrm{s}}e_{\mathrm{g,b}}^{\mathrm{s}}$$
$$+\tfrac{1}{2}\sum_{\mathrm{g}=1}^{m}n_{\mathrm{g}}\sum_{\mathrm{q}=1}^{t}\sum_{\mathrm{s}=1}^{t}\left(2\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}]\sum_{\mathrm{b}=1}^{c}(\bar{\mathbf{x}}_{\mathrm{g}}^{\mathrm{q}}-\boldsymbol{\mu}_{0}^{\mathrm{q}})'\mathbf{v}_{\mathrm{b}}e_{\mathrm{g,b}}^{\mathrm{s}} + 2\mathbf{A}^{-1}[\mathrm{q},\mathrm{s}]\sum_{\mathrm{b}=c+1}^{\mathrm{k}}(\bar{\mathbf{x}}_{\mathrm{g}}^{\mathrm{q}}-\boldsymbol{\mu}_{0}^{\mathrm{q}})'\mathbf{v}_{\mathrm{b}}^{\mathrm{s}}e_{\mathrm{g,b}}^{\mathrm{s}}\right), \quad (8.15)$$

where $\boldsymbol{\mu}_{0}^{\mathrm{q}}$ is the $p \times 1$ vector of overall means for the q^{th} occasion and $\bar{\mathbf{x}}_{\mathrm{g}}^{\mathrm{q}}$ is the $p \times 1$ vector of sample means for the g^{th} group.

I start by deriving the equations for the common variates. The constraints are incorporated by the method of Lagrangian multipliers. Note that these and subsequent constraints with Lagrangian multipliers are implicitly set to zero. The constraints with Lagrangian multipliers for the unit length of the common variates are as follows below.

$$\sum_{\mathrm{a}=1}^{\mathrm{c}}\frac{\gamma_{\mathrm{a}}}{2}\left(\mathbf{v}_{\mathrm{a}}'\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{\mathrm{a}} - 1\right),$$

where γ_a are c Lagrangian multipliers. The constraints with Lagrangian multipliers for the orthogonality of the common variates are:

$$\sum_{a=1}^{c}\sum_{b=1}^{a-1}\gamma_{ab}\mathbf{v}_a'\Sigma_w\mathbf{v}_b,$$

where γ_{ab} are $c(c-1)/2$ Lagrangian multipliers. The constraints with Lagrangian multipliers for the orthogonality of each common variate with all of the unique variates are:

$$\sum_{q=1}^{t}\sum_{a=1}^{c}\sum_{b=c+1}^{u}\gamma_{abq}\mathbf{v}_a'\Sigma_w\mathbf{v}_b^q, \tag{8.16}$$

where γ_{abq} are $ct(u-c)$ Lagrangian multipliers. The constraints with Lagrangian multipliers for the restriction to unit length of the unique variates are:

$$\sum_{a=c+1}^{u}\sum_{q=1}^{t}\frac{\gamma_{aq}}{2}\left(\mathbf{v}_a^{q\prime}\Sigma_w\mathbf{v}_a^q-1\right),$$

where γ_{aq} are $(u-c)t$ Lagrangian multipliers. The constraints with Lagrangian multipliers for the mutual orthogonality of each unique variate with the other unique variates of the same occasion are:

$$\sum_{q=1}^{t}\sum_{a=c+1}^{u}\sum_{b=c+1}^{a-1}\gamma_{abq}\mathbf{v}_a^{q\prime}\Sigma_w\mathbf{v}_b^q,$$

where γ_{abq} are $t(u-c)(u-c-1)/2$ Lagrangian multipliers.

Denote the log-likelihood modified by constraints with Lagrangian multipliers by $l^*(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. Take the derivative of $l^*(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $\mathbf{v}_f$:

$$\frac{\delta l^*(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta\mathbf{v}_f}=\sum_{g=1}^{m}n_g\sum_{q=1}^{t}\sum_{s=1}^{t}ai_{qs}\left(-\sum_{a=1}^{c}e_{g,a}^q\Sigma_w\mathbf{v}_a e_{g,f}^s-\sum_{b=c+1}^{u}e_{g,f}^q\Sigma_w\mathbf{v}_b^s e_{g,b}^s+(\bar{\mathbf{x}}_g^q-\boldsymbol{\mu}_0^q)e_{g,f}^s\right)$$
$$+\gamma_f\Sigma_w\mathbf{v}_f+\sum_{\substack{a=1\\a\neq f}}^{c}\gamma_{af}\Sigma_w\mathbf{v}_a+\sum_{q=1}^{t}\sum_{b=c+1}^{u}\gamma_{fbq}\Sigma_w\mathbf{v}_b^q.$$

Set this equal to a zero vector to yield the estimating equations for $\mathbf{v}_f$.

Next I derive the estimating equations for the unique variates. Take the derivative of $l^*(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $\mathbf{v}_f^r$:

$$\frac{\delta l^*(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta\mathbf{v}_f^r}=\sum_{g=1}^{m}n_g\sum_{q=1}^{t}\mathbf{A}^{-1}[q,r]\left(-\sum_{a=1}^{c}e_{g,a}^q\Sigma_w\mathbf{v}_a e_{g,f}^r-\sum_{b=c+1}^{u}e_{g,b}^q\Sigma_w\mathbf{v}_f^r e_{g,f}^r+(\bar{\mathbf{x}}_{gq}-\boldsymbol{\mu}_0^q)e_{g,f}^r\right)$$
$$+\gamma_{rf}\Sigma_w\mathbf{v}_f^r+\sum_{s=1}^{c}\gamma_{afr}\Sigma_w\mathbf{v}_a+\sum_{b=c+1}^{u}\gamma_{fbr}\Sigma_w\mathbf{v}_b^r.$$

Set this equal to a vector of zeros to yield the estimating equations for solving for $\mathbf{v}_f^r$.

Lastly I derive estimating equations for the e_{gb}^s terms, beginning with those for the group positions corresponding to the common variates. The constraints here are $\sum_{g=1}^{m}n_g e_{g,b}^s=0$ for

$s = 1,\ldots,t$ and $b = 1,\ldots,c$. They are handled by letting $e^{s}_{m,b} = -\sum_{h=1}^{m-1} e^{s}_{h,b}$ for $s = 1,\ldots,t$, and then taking the derivative of the log-likelihood with respect to $e^{r}_{h,w}$ for $h \neq m$. Then the estimating equations for the e^{s}_{gb} are obtained by setting this derivative equal to zero:

$$\frac{\delta l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta e^{r}_{h,f}} = \sum_{q=1}^{t} \mathbf{A}^{-1}[q,r]\left(-\sum_{a=1}^{c} e^{q}_{h,a}\mathbf{v}'_{f}\boldsymbol{\Sigma}\mathbf{v}_{a} - \sum_{b=c+1}^{u} e^{q}_{h,b}\mathbf{v}'_{f}\boldsymbol{\Sigma}\mathbf{v}^{s}_{b} + (\bar{\mathbf{x}}^{s\prime}_{h} - \boldsymbol{\mu}^{q}_{0})\mathbf{v}_{f} \right).$$

The estimating equations for the group positions corresponding to the unique variates are handled similarly to those corresponding to the common variates. The constraints and the manner of incorporating them into the estimating equations are the same as the for the common variates: $\sum_{g=1}^{m} n_{g} e^{s}_{g,b} = 0$ for $s = 1,\ldots,t$ and $b = c+1,\ldots,u$. Let $e^{s}_{m,b} = -\sum_{h=1}^{m-1} e^{s}_{h,b}$, for $s = 1,\ldots,t$ and $b = c+1,\ldots,u$, and take the derivative of $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $e^{r}_{h,w}$ for $h \neq m$:

$$\frac{\delta l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta e^{r}_{h,f}} = \sum_{q=1}^{t} \mathbf{A}^{-1}[q,r]\left(2\sum_{a=1}^{c} e^{q}_{h,a}\mathbf{v}'_{a}\boldsymbol{\Sigma}_{w}\mathbf{v}^{r}_{f} + 2\sum_{b=c+1}^{u} \mathbf{v}'_{f}\boldsymbol{\Sigma}_{w}\mathbf{v}^{s}_{b}e^{s}_{h,b} + 2\bar{\mathbf{x}}^{q\prime}_{h}\mathbf{v}^{r}_{f} \right).$$

Set these derivatives equal to zero to obtain the estimating equations.

8.5.6 Estimating Unchanging Group Positions

It may be of interest to hypothesize that the scores for the group means on the common variates, $e^{q}_{g,a}$, do not change, i.e., are equal over occasion, and to estimate these stable scores. The unchanging score of the g^{th} group for the b^{th} common variate shall be denoted by $e_{g,b}$. The likelihood equation is obtained by substituting $e_{g,b}$ for $e^{q}_{g,b}$ in the likelihood equation for CVA/time (unc.) (8.15). The constraints and the manner of incorporating them are the same as for the $e^{q}_{g,b}$ terms in Section **8.5.5**. Taking the derivative of $l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $e_{h,w}$ and setting it equal to zero yields the following estimating equation for $e_{h,w}$:

$$\frac{\delta l(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\delta e_{h,w}} = -n_{h}\sum_{q=1}^{t}\sum_{s=1}^{t}\left(\sum_{a=1}^{c} e_{h,a}\mathbf{v}'_{w}\boldsymbol{\Sigma}_{qr}\mathbf{v}_{a} + \sum_{b=c+1}^{u} e^{q}_{h,b}\mathbf{v}'_{w}\boldsymbol{\Sigma}_{qr}\mathbf{v}^{s}_{b} \right)$$
$$+n_{h}\left(\sum_{q=1}^{t}\sum_{s=1}^{t}\sum_{a=1}^{c} (\bar{\mathbf{x}}^{s}_{h} - \boldsymbol{\mu}^{q}_{0})'\mathbf{v}_{w} \right).$$

The estimating equations for $\mathbf{v}_{i}$ and $\mathbf{v}^{s}_{i}$ are the same as those in the previous section except that $e_{g,b}$ is substituted for $e^{q}_{g,b}$.

8.6 EXAMPLE FOR CVA/TIME WITH UNCORRELATED VARIATES - SEX DIFFERENCES IN MATH ANXIETY BEFORE AND AFTER INTRODUCTORY CALCULUS

In this section I present a real data example of modeling the group means over time in the space of uncorrelated canonical variates. The example is a relatively simple one. 423 male college students and 118 female college students enrolled in an introductory calculus course at Virginia Tech were given a questionnaire at the beginning and end of the course. The questionnaire included 19 questions pertaining to "math anxiety". Math anxiety is generally construed to be a particular apprehension some students have about mathematics. The groups of interest are men and women. Thus $p = 19$, $t = 2$ and $m = 2$. The 19 questions are presented in **Table 8.4**. The responses to these questions followed the ordinal scale, as seen in **Table 8.5**. Thus the data are not normal and the inferential techniques used in the analysis are at best approximate.

Table 8.4 The Math Anxiety Questions

1. Generally, I have felt secure about attempting mathematics.
2. The thought of a math test scares me.
3. I usually have been at ease in math classes.
4. It wouldn't bother me at all to take more math classes.
5. It would make me happy to be recognized as an excellent student in math.
6. Figuring out mathematical problems does not appeal to me.
7. Math is enjoyable and stimulating to me.
8. I get a sinking feeling when I think of trying hard math problems.
9. Winning a prize in mathematics would make me feel uncomfortably conspicuous.
10. Even though I study, math seems unusually hard for me.
11. I study mathematics because I know how useful it can be.
12. I wouldn't like people to think I'm smart in math.
13. I like math puzzles.
14. I memorize math formulas and techniques but often don't understand the underlying concepts.
15. I am sure I could do advanced work in math.
16. I'm not the type to do well in math.
17 Mathematics is a worthwhile and necessary subject.
18. I'd be proud to be a top student in math.
19. I would rather have someone give me the solution to a hard math problem than solve it myself.

Table 8.5 Possible Responses

1) Agree 2) Tend to agree 3) Tend to disagree 4)Disagree

The analysis pursues both statistical inference and the interpretation of the results. The first question for statistical inference is, is there a weighted sum of the variables that distinguishes between the sexes? Now, if there is such a weighted sum, is it interpretable as a "math anxiety"

construct? Further inferential questions are, is such a variate stable over time? If yes, are the scores of the group means on the variate, that is, the positions of the group means on the variate, stable over time? How does one interpret the scores or positions of the group means?

It is believed from previous research that there are differences between the sexes in math anxiety. However, it is not known whether such a construct is stable over the course of taking introductory calculus. Since there are two groups the maximum number of canonical variates is one. First the one common canonical variate and one unique variate model will be estimated. Then a hypothesis test will be performed based on the likelihood ratio test statistics with the common variate hypothesis being the null hypothesis.

The one common variate model is:

$$\boldsymbol{\mu}_g = \boldsymbol{\mu}_0 + \begin{bmatrix} e_g^1 \boldsymbol{\Sigma}_w \mathbf{v} \\ e_g^2 \boldsymbol{\Sigma}_w \mathbf{v} \end{bmatrix}, \text{ for } g = 1,2 . \tag{8.17}$$

And the one unique variate model is:

$$\boldsymbol{\mu}_g = \boldsymbol{\mu}_0 + \begin{bmatrix} e_g^1 \boldsymbol{\Sigma}_w \mathbf{v}^1 \\ e_g^2 \boldsymbol{\Sigma}_w \mathbf{v}^2 \end{bmatrix}, \text{ for } g = 1,2 .$$

The test for common versus unique variates is stated as:

$$H_0 : \mathbf{v}^1 = \mathbf{v}^2$$

$$H_1 : \mathbf{v}^1 \neq \mathbf{v}^2 .$$

There are a total of 19 weights to be estimated for each variate, with the constraint that each variate have a variance of one. Thus the difference between the number of parameters to be estimated in the null and in the alternative hypotheses is 18, and the test statistic under the null hypothesis is distributed approximately as a chi-square with 18 degrees of freedom. The parameter estimates were obtained by solving the estimating equations given in Sections **8.5.2**, **8.5.3** and **8.5.5**. The likelihood test statistic was determined as described in Section **8.3.6**. The observed value of the test statistic is 20.8, which is not significant, so one fails to reject H_0.

Given that one has failed to reject the null hypothesis of one common variate, the next question of interest is whether the positions of the group means over time on the common variate are changing. The null hypothesis is that they are unchanging. The test for equality of group positions is stated as:

$$H_0 : e_1^1 = e_1^2, \quad e_2^1 = e_2^2$$

H_1: at least one of the above is untrue.

The estimating equations for the unchanging group positions are given in Section **8.5.6**. The estimates are: $e_1 = 0.1427$ and $e_2 = -0.5117$. The estimates for the changing group positions are obtained as part of the estimation of the common variate model Section **8.5.5**. Those estimates are: $e_1^1 = 0.1909$, $e_1^2 = 0.0804$, $e_2^1 = -0.6844$ and $e_2^2 = -0.2882$. (Note that the group positions for any occasion sum to zero when weighted by sample size).

Under the null hypothesis the test statistic follows a chi-square distribution with one degree of freedom. The resulting p-value is $p<0.002$. Hence one rejects H_0 and concludes $e_g^1 \neq e_g^2$. In other words, one concludes that the differences between the group means change

over the two occasions. In this case they move closer, see **Figure 8.5**. Note that rejecting $H_0: e_g^1 = e_g^2$ obviates any need to test $H_0: e_g^1 = e_g^2 = 0$, which is the test for the existence of treatment effects.

At this point it is appropriate to examine the common canonical variates and ascertain if they arguably comprise a math anxiety construct. The interpretation is clearer when examining the structural coefficients, which are the correlations of the variates with the variables (see Section **2.2.1**). The canonical variate weights and the structural coefficients are presented in **Table 8.6**. From inspection of the structural coefficients it is apparent that the variate is correlated positively with answers that seem to indicate low math anxiety. For example, a high score on Question #4, "It wouldn't bother me at all to take more math classes", is arguably indicative of low math anxiety. The signs of the correlations of all 19 variables with the canonical variate are all arguably consistent with low math anxiety, though these correlations vary in magnitude.

Table 8.6 Canonical Variate Weights and Structural Coefficients

Question #	Canonical Variates	Structural Coefficients
1	-0.2007	0.0459
2	-0.2257	-0.0810
3	0.0632	0.1246
4	0.8437	0.5578
5	-0.2397	0.2101
6	0.1774	-0.2092
7	0.6225	0.4772
8	0.2444	-0.0289
9	-0.6116	-0.4668
10	-0.3278	-0.1353
11	0.0473	0.1242
12	-0.1554	-0.2091
13	0.1338	0.2921
14	-0.6087	-0.2734
15	-0.0988	0.0996
16	0.2919	-0.0620
17	-0.2328	0.0957
18	-0.1069	-0.2529
19	0.1955	-0.0220

Next, consider the estimates of the positions of the group means. Men clearly score higher on this low math anxiety construct, though the difference between the sexes diminishes over time; see **Figure 8.5** below. Note that the axis in **Figure 8.5** is the canonical variate, which by definition has a variance of one, and also that the group means have been centered at zero.

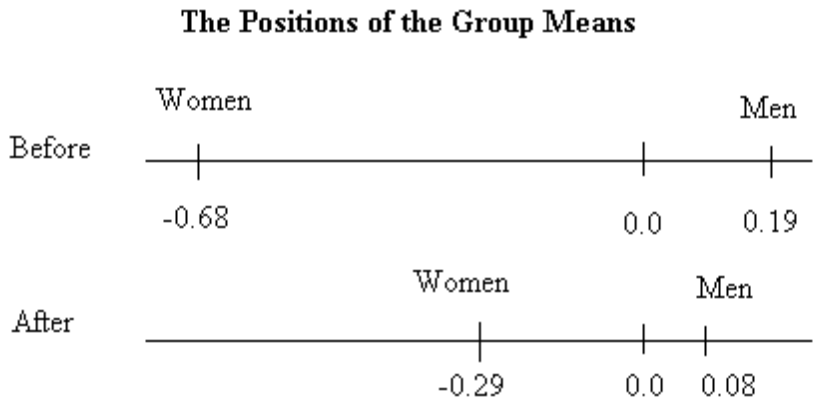

Figure 8.5

Figures 8.6, 8.7, 8.8 and **8.9** below show the group separation more clearly than does **Figure 8.5**. They are histograms of the scores for the 423 men and 118 women at both occasions. These historgrams are based on the uncentered data. As in **Figure 8.5**, the separation is greater for the first occasion.

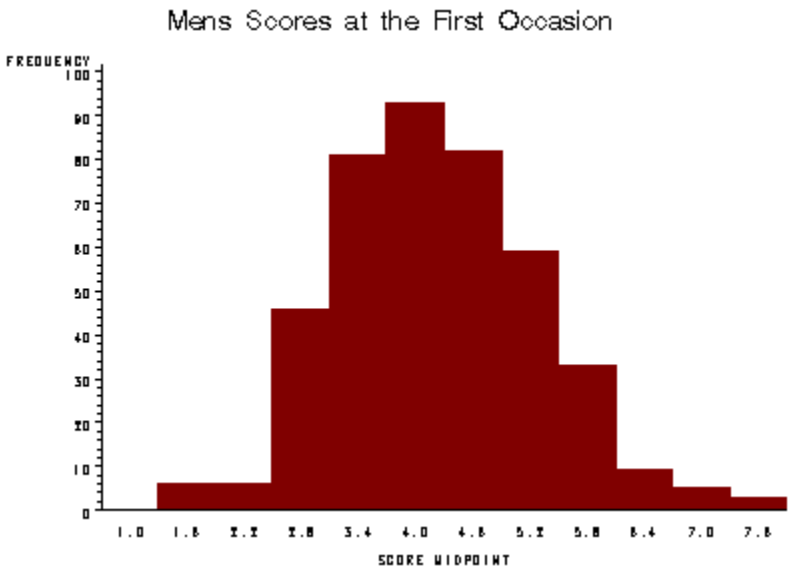

Figure 8.6

Figure 8.7

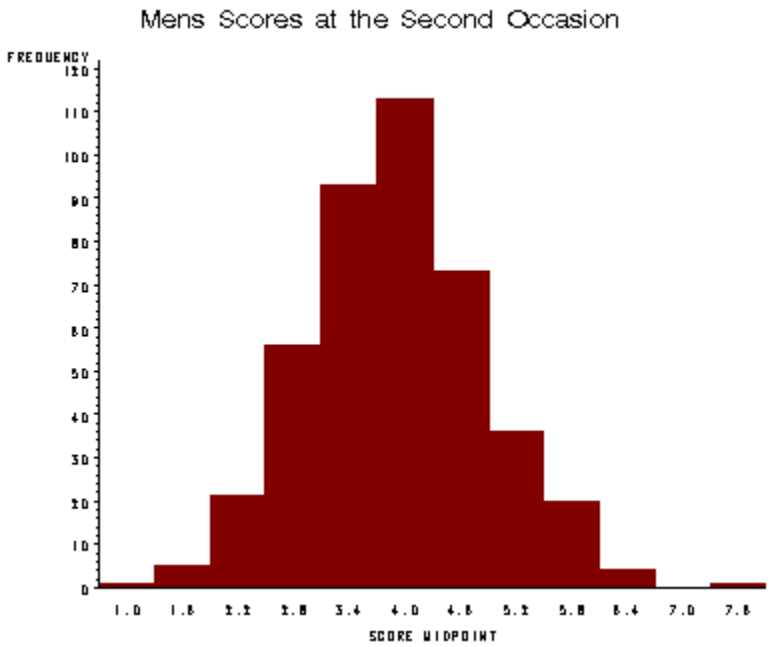

Figure 8.8

Figure 8.9

To summarize the analysis so far, one can conclude that men and women do differ on the construct of math anxiety, and that this construct is stable over time. Further, the women admit to more math anxiety, though the difference in admitted math anxiety between the sexes shrinks at the end of the course.

Next consider what the parameters mean in a geometric sense. The weights for the canonical variate are a direction in the multi-dimensional variable space. The one canonical variate model hypothesizes that the group means, when centered, will line up on this canonical variate. The group positions indicate where on these variates the group means are centered.

Another way to interpret the results is in the spirit of a multivariate regression. That is, one predicts the mean response of each variable for each group. Such an interpretation will allow one to consider the canonical variate and the group positions in conjunction. To obtain the vector of predicted group means for any occasion, apply equation (8.17). **Table 8.7** shows the observed and predicted group means of men and women on the 19 questions at the first occasion. The predicted group means are generally similar to the observed. For example, in Question 18 they are almost the same.

For comparison **Table 8.8** shows the overall group means for each question and the standard deviations. Appendix Ten has the complete sample variance-covariance matrix and the maximum likelihood estimate of the variance-covariance matrix.

Table 8.7 Observed and Predicted Means at the First Occasion

	Men		Women	
Question #	Observed	Predicted	Observed	Predicted
1	1.284	1.281	1.254	1.264
2	2.071	2.073	2.136	2.129
3	1.655	1.652	1.576	1.585
4	1.993	1.988	1.576	1.593
5	1.317	1.317	1.220	1.219
6	1.837	1.837	1.720	1.719
7	2.135	2.133	1.831	1.835
8	1.993	1.993	1.983	1.982
9	1.976	1.976	1.678	1.678
10	1.773	1.772	1.678	1.680
11	1.600	1.597	1.525	1.537
12	1.548	1.549	1.424	1.420
13	2.035	2.036	1.839	1.838
14	2.778	2.777	2.585	2.588
15	1.882	1.877	1.805	1.822
16	1.485	1.484	1.458	1.462
17	1.248	1.247	1.212	1.215
18	1.317	1.317	1.195	1.195
19	1.716	1.712	1.703	1.717

Table 8.8 Group Means and Standard Deviations for each Question

Question #	Overall Mean	Sample Variance	MLE of Variance
1	1.34288	0.591816	0.585489
2	1.98429	0.835527	0.815629
3	1.61553	0.680435	0.660555
4	1.82348	0.837711	0.822719
5	1.34843	0.607680	0.583501
6	1.83087	0.721879	0.700720
7	2.04621	0.749102	0.720939
8	1.97412	0.768150	0.741683
9	1.87061	0.770398	0.744001
10	1.78466	0.726312	0.706184
11	1.64418	0.692900	0.669706
12	1.53789	0.749156	0.751014
13	2.06285	0.859188	0.815019
14	2.78928	0.800275	0.771214
15	1.85120	0.744667	0.731749
16	1.55638	0.624593	0.608705
17	1.31978	0.497792	0.492861
18	1.33272	0.599802	0.589608
19	1.77542	0.694748	0.679944

In concluding this example, **Figure 8.10** presents the matrix of proportionality constants, **A**. What is noteworthy here is that the weights for the within-groups covariance matrices for measurements at the beginning and at the end of the course are nearly equal.

$$\begin{bmatrix} 0.986 & 0.289 \\ 0.289 & 1.014 \end{bmatrix}$$

Figure 8.10 The Matrix of Proportionality Constants (**A)**

8.7 A COMPARISON TO ALTERNATIVE METHODS, INCLUDING DOUBLY MULTIVARIATE REPEATED MEASURES

In this section I make a comparison between CVA/time and alternative methods for longitudinal multivariate data with group structure. These alternative methods attempt to answer the same questions as the common variate hypothesis does; that is to determine what is and is not changing over time. However they attempt this without the clarity and efficacy achieved by explicitly modeling common variates. The interpretations of these models for data which have the common canonical variate structure will be illuminating.

I will begin by briefly considering two simple alternative approaches. Then I will discuss in greater depth two more ambitious approaches. The first of these involves performing a canonical variate analysis with the measurements at different occasions treated as different

variables. The second, which is of particular interest, is an analysis that is loosely called "doubly multivariate repeated measures".

8.7.1 Two Simple Approaches

One simple approach to longitudinal multivariate data with group structure is to perform a separate canonical variate analysis at each occasion. It is easy to see that if common canonical variate structure exists over time that the common variates will be found by each analysis. The limitations of such an approach are that one has no means to test for the appropriateness of a common variate structure, nor for the number of common variates. Furthermore, due to chance variation one does not have a single estimate for a given common variate.

Another simple approach would be to pool the measurements over time. This approach raises the question of whether one centers by an overall mean or separately at each occasion. All of the other approaches discussed in this section either explicitly or implicitly assume that one centers at each occasion. This issue aside, it is clear that pooling the variables will estimate the common variates if they exist. However, inspecting the estimated variates yields neither a hint of which (if any) of the estimated variates are common, nor what is changing over time.

8.7.2 Measurements at Different Occasions Treated as Distinct Variables

An approach one may take is to treat the measurements taken at different occasions as distinct variables in a single canonical variate analysis. A failing of this approach is that if there is an effect, i.e., a statistically significant canonical variate and an associated non-zero canonical correlation, it cannot be attributed specifically to either treatment effects or to time-treatment interaction effects. Further, neither common variates, unique variates nor group positions are estimated. However, the canonical variates one obtains do have a particular structure. The $tp \times 1$ vector of weights of each canonical variate consists of t $p \times 1$ subvectors, each of which is a linear combination of the common and unique variates. Although this point is in itself of minor interest, it is illustrative to show it.

First determine **B**, where **B** is the matrix of between-groups sums of squares and crossproducts, under the assumption that the common variates hypothesis is true. **B** is generally defined as:

$$\mathbf{B} = \sum_{g=1}^{m} n_g (\boldsymbol{\mu}_g - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_g - \boldsymbol{\mu}_0)' .$$

Now consider the common variates model after the data are centered:

$$\boldsymbol{\mu}_g - \boldsymbol{\mu}_0 = \boldsymbol{\Sigma}_w \mathbf{v}_1 \otimes \mathbf{e}_{g,1} + \ldots + \boldsymbol{\Sigma}_w \mathbf{v}_c \otimes \mathbf{e}_{g,c},$$

for $g = 1, \ldots, c$. Then the $p \times p$ submatrix of **B** corresponding to the q^{th}, s^{th} occasions, denoted as $\mathbf{B}^{q,s}$, is as follows:

$$\mathbf{B}^{q,s} = \sum_{g=1}^{m} n_g \left(\boldsymbol{\Sigma}_w \mathbf{v}_1 e_{g,1}^{q} + \ldots + \boldsymbol{\Sigma}_w \mathbf{v}_c e_{g,c}^{q} \right) \left(\boldsymbol{\Sigma}_w \mathbf{v}_1 e_{g,1}^{s} + \ldots + \boldsymbol{\Sigma}_w \mathbf{v}_c e_{g,c}^{s} \right)'$$

$$\mathbf{B}^{q,s} = \sum_{i=1}^{r}\sum_{j=1}^{r} \mathbf{v}_i \mathbf{v}_j' \left(\sum_{g=1}^{m} n_g e_{g,i}^{q} e_{g,j}^{s} \right) = \mathbf{V} \left(\sum_{g=1}^{m} \mathbf{d}_g^{q} \mathbf{d}_g^{s\prime} \right) \mathbf{V}',$$

where $\mathbf{V}$ is the matrix whose i^{th} column is the i^{th} canonical variate, $\mathbf{v}_i$, for $i = 1,\ldots,r$, and $\mathbf{d}_g^{q}$ is the $r \times 1$ vector whose i^{th} element is $n_g^{1/2} e_{g,i}^{q}$, for $i = 1,\ldots,r$ and $q = 1,\ldots,t$. If $r < p$ then complement $\mathbf{V}$ with $p - r$ canonical variates which correspond to canonical correlations of zero, making $\mathbf{V}$ a $p \times p$ matrix, and likewise complement each $\mathbf{d}_g^{q}$ vector with $p - r$ zeros, making them $p \times 1$ vectors. Then the implied between-groups crossproducts matrix, $\mathbf{B}$, is

$$\mathbf{B} = \left(\boldsymbol{\Sigma}_w \otimes \mathbf{I}_{t\times t}\right)\left(\mathbf{V} \otimes \mathbf{I}_{t\times t}\right)\mathbf{D}\left(\mathbf{V}' \otimes \mathbf{I}_{t\times t}\right)\left(\boldsymbol{\Sigma}_w \otimes \mathbf{I}_{t\times t}\right),$$

where $\mathbf{D}$ is an $pt \times pt$ matrix, $\mathbf{D} = \sum_{g=1}^{m} \mathbf{d}_g \mathbf{d}_g'$, with $\mathbf{d}_g = [\mathbf{d}_g^{1\prime}, \mathbf{d}_g^{2\prime}, \ldots, \mathbf{d}_g^{t\prime}]'$.

The canonical variates are obtained as described in Section **2.2.1**, by performing a SVD on $\frac{1}{n\text{-}1}\boldsymbol{\Sigma}^{-1/2\prime}\mathbf{B}\boldsymbol{\Sigma}^{-1/2}$, where $n = \sum_{g=1}^{m} n_g$. The first step in determining this SVD is to express $\frac{1}{n\text{-}1}\boldsymbol{\Sigma}^{-1/2\prime}\mathbf{B}\boldsymbol{\Sigma}^{-1/2}$ as follows:

$$\frac{1}{n}\boldsymbol{\Sigma}^{-1/2\prime}\mathbf{B}\boldsymbol{\Sigma}^{-1/2} = \frac{1}{n}\boldsymbol{\Sigma}^{-1/2\prime}\left(\boldsymbol{\Sigma}_w \otimes \mathbf{I}_{t\times t}\right)\left(\mathbf{V} \otimes \mathbf{I}_{t\times t}\right)\mathbf{D}\left(\mathbf{V}' \otimes \mathbf{I}_{t\times t}\right)\left(\boldsymbol{\Sigma}_w \otimes \mathbf{I}_{t\times t}\right)\boldsymbol{\Sigma}^{-1/2}. \qquad (8.18)$$

Assume $\boldsymbol{\Sigma} = \mathbf{A} \otimes \boldsymbol{\Sigma}_w$ (8.12), so $\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}_w^{-1/2} \otimes \mathbf{A}^{-1/2}$. Recall from Section **2.2.1** that $\mathbf{V} = \boldsymbol{\Sigma}_w^{-1/2}\mathbf{V}^*$, where $\mathbf{V}^*$ is orthogonal. Then $\mathbf{V} \otimes \mathbf{I}_{t\times t} = \left(\boldsymbol{\Sigma}_w^{-1/2} \otimes \mathbf{I}_{t\times t}\right)\left(\mathbf{V}^* \otimes \mathbf{I}_{t\times t}\right)$. Substituting the above expressions back into (8.18) gives:

$$\frac{1}{n}\boldsymbol{\Sigma}^{-1/2\prime}\mathbf{B}\boldsymbol{\Sigma}^{-1/2} = \frac{1}{n}\left(\boldsymbol{\Sigma}_w^{-1/2\prime} \otimes \mathbf{A}^{-1/2\prime}\right)\left(\boldsymbol{\Sigma}_w \otimes \mathbf{I}_{t\times t}\right)\left(\boldsymbol{\Sigma}_w^{-1/2} \otimes \mathbf{I}_{t\times t}\right)\left(\mathbf{V}^* \otimes \mathbf{I}_{t\times t}\right)\mathbf{D}$$
$$\left(\mathbf{V}^{*\prime} \otimes \mathbf{I}_{t\times t}\right)\left(\boldsymbol{\Sigma}_w^{-1/2} \otimes \mathbf{I}_{t\times t}\right)\left(\boldsymbol{\Sigma}_w \otimes \mathbf{I}_{t\times t}\right)\left(\boldsymbol{\Sigma}_w^{-1/2} \otimes \mathbf{A}^{-1/2}\right).$$

The above simplifies to:

$$\frac{1}{n}\boldsymbol{\Sigma}^{-1/2\prime}\mathbf{B}\boldsymbol{\Sigma}^{-1/2} = \left(\mathbf{V}^{*\prime} \otimes \mathbf{I}_{t\times t}\right)\left[\frac{1}{n}\left(\mathbf{I}_{p\times p} \otimes \mathbf{A}^{-1/2\prime}\right)\mathbf{D}\left(\mathbf{I}_{p\times p} \otimes \mathbf{A}^{-1/2}\right)\right]\left(\mathbf{V}^{*\prime} \otimes \mathbf{I}_{t\times t}\right). \qquad (8.19)$$

Note that $\left(\mathbf{V}^{*\prime} \otimes \mathbf{I}_{t\times t}\right)$ is orthogonal. Let the singular value decomposition of the square-bracketed part of (8.19) be $\frac{1}{n}\left(\mathbf{I}_{p\times p} \otimes \mathbf{A}^{-1/2\prime}\right)\mathbf{D}\left(\mathbf{I}_{p\times p} \otimes \mathbf{A}^{-1/2}\right) = \mathbf{D}^*\mathbf{M}\mathbf{D}^{*\prime}$. Then $\left(\mathbf{V}^{*\prime} \otimes \mathbf{I}_{t\times t}\right)\mathbf{D}^*$ is also orthonormal because $\mathbf{D}^*$ is. Hence the matrix of canonical variates obtained, denoted by $\mathbf{U}$, will be:

$$\mathbf{U} = \left(\mathbf{\Sigma}_{\mathrm{w}}^{-1/2} \otimes \mathbf{I}_{t\times t} \right)\left(\mathbf{V}^{*\prime} \otimes \mathbf{I}_{t\times t} \right)\mathbf{D}^{*} = \left(\mathbf{V}^{\prime} \otimes \mathbf{I}_{t\times t} \right)\mathbf{D}^{*}. \tag{8.20}$$

Now one sees that each column of $\mathbf{U}$ is a concatenation of t subvectors which are a linear compound of $\mathbf{V}$: $\mathbf{U}_{\mathrm{i}}^{\mathrm{q}} = \mathbf{V}\mathbf{D}_{\mathrm{i}}^{\mathrm{q}}$, where $\mathbf{U}_{\mathrm{i}}^{\mathrm{q}}$ and $\mathbf{D}_{\mathrm{i}}^{\mathrm{q}}$ are the i^{th} column and q^{th} $p\times 1$ subvector of $\mathbf{U}$ and $\mathbf{D}$.

Next consider that unique variates can be modeled as multiple common variates, a point which is touched on briefly at the end of Section **8.3.1**. The $(u - c)$ unique variates at each occasion can be viewed as up to $t(u - c)$ common variates with the appropriate group positions; if $t(u - c) \geq p$ then the unique variates model is equivalent to a common variates model with a full complement of p common variates. For a simple example, assume that one has one unique variate at each of t occasions and that these unique variates are mutually uncorrelated. Then these unique variates can be viewed as t common variates; a group position on a given common variate is either the position of the original unique variate or zero, depending on whether or not the common variate corresponds to an original unique variate at the given occasion. As a rule, unique variates can always be put into the form of common variates, with the group positions conveying the change over time.

Putting a model with unique variates in the form of a model with only common variates allows one to take advantage of the earlier results for common variates, in particular to generalize (8.20). Hence one can assert that the canonical variates one obtains consist of subvectors which are linear combinations of these "common" variates; that is, of the original common and unique variates.

8.7.3 Doubly Multivariate Repeated Measures

I am aware of only one method other than CVA/time that specifically deals with longitudinal multivariate data with group structure. This is the analysis which goes by the name of doubly multivariate repeated measures, an example of which is given in the SAS-STAT Users Guide (1990). In this section I compare doubly multivariate repeated measures with CVA/time. Doubly multivariate repeated measures is the most sophisticated of the alternatives I discuss. It will be seen that it answers some but not all of the questions CVA/time answers.

The method described in the SAS/STAT User's Guide tests for time effects, treatment effects and time-treatment interactions by performing either a standard or a modified multivariate analyses of variance (MANOVA) on transformed variables. I will review each of these tests in the context of common and unique variate structure and compare their performance to CVA/time (unc.).

First I will discuss the test for time effects. The first step in the SAS approach for testing for simple time effects is to create time profile variates. One transforms the tp original measurements into $(t-1)p$ variates by taking the profiles of the measurements at different occasions with respect to a baseline occasion. If the t^{th} occasion is chosen to be the baseline occasion then one would have $\mathrm{X}_{\mathrm{ij}}^{*} = \mathrm{X}_{\mathrm{ij}} - \mathrm{X}_{\mathrm{i}t}$ for $\mathrm{i} = 1,\ldots,p$ and $\mathrm{j} = 1,\ldots,t-1$. Then,

disregarding group membership (treatment effects), the vector of means of these variates is tested to be zero. In SAS one does this by using the "manova" statement with "H = int ercept" in Proc GLM.

In comparison, the CVA/time model centers the data at each occasion. This centering removes the time effect that the SAS approach tests for. However, if the examination of a simple time effect is of interest, one can perform this same test that SAS does in addition to performing a CVA/time analysis. Note that the doubly multivariate repeated measures analysis test for treatment effects and the test for treatment-time interactions also remove the simple time effect from the analysis by the transformations of the variables that they use.

The first step in the test for treatment effects given by SAS is to create transformed variables by summing up the measurements over occasion. That is, one obtains p transformed variables, X_j^*, where $X_j^* = \sum_{q=1}^{t} X_j^q$ for $i = 1,\ldots,p$ and $j = 1,\ldots,t$. Then one performs a one-way MANOVA on the transformed data.

This test for treatment effects is powerful only in the following circumstances: if the common variate hypothesis is true or approximately true; i.e., the unique variates are nearly collinear; and the scores for the group means at different occasions are equal or similar over time. Indeed, if the common variate hypothesis holds exactly and the positions of the group means are completely stable over time ($e_{g,j}^q = e_{g,k}^q$, $j \neq k$), then the doubly multivariate repeated measures finds the common variates exactly. However, to the extent that variates or group positions change over time the effects will be muddled and the test rendered ineffective. To realize these assertions one can examine the vector of expected values for the transformed variables for each group, denoted by $\boldsymbol{\mu}_g^*$, under the assumption of the common variate structure. That is, for $g = 1,\ldots,m$:

$$E\left(\sum_{q=1}^{t} \mathbf{x}_g^q\right) = \boldsymbol{\mu}_g^* = \sum_{q=1}^{t} \boldsymbol{\mu}_0^q + \boldsymbol{\Sigma}_w \mathbf{v}_1 \left(\sum_{q=1}^{t} e_{g,1}^q\right) + \ldots + \boldsymbol{\Sigma}_w \mathbf{v}_c \left(\sum_{q=1}^{t} e_{g,c}^q\right)$$

$$\boldsymbol{\mu}_g^* = \boldsymbol{\mu}_0^* + \boldsymbol{\Sigma}_w \mathbf{v}_1 e_{g,1}^* + \ldots + \boldsymbol{\Sigma}_w \mathbf{v}_c e_{g,c}^* \qquad (8.21)$$

where $\mathbf{x}_g^q$ is the $p \times 1$ vector of random variables at the q^{th} occasion for the g^{th} group. The form in (8.21) is identical to the form of a canonical variate analysis with the common variates as the canonical variates (see 2.1). (Recall that a one-way MANOVA is equivalent to a canonical variate analysis; to obtain the canonical variates from SAS one requests "canonical" in the "manova" statement). Hence the method estimates the common variates if the e_{gi}^* are not zero. However, if the group positions are not at least similar over time they tend to cancel each other out, resulting in a weaker or non-detectable effect; i.e., $e_{g,i}^*$ terms that are close to zero. Hence the treatment effects may not be detected.

On the other hand, if the variates change over time, i.e., one has unique variates, one can view the analysis as one with extra common variates where the changes in the group positions convey the change over time, as described in Section **8.7.2**. Since group positions on these

common variates change, the treatment effects get muddled in the summation and again the resulting treatment effect will be weaker.

The SAS test for time-treatment interactions analyzes the $(t-1)p$ variables created by taking the profiles of the measurements at different occasions with respect to a baseline occasion, as was done to analyze the time effects. Here, however, one performs a usual one-way MANOVA on the transformed data.

The test for time-treatment interactions does indeed detect change over time in the group structure. But it provides no way to determine if that change is due to changing group positions on common variates or to unique variates at different occasions. Nor does it estimate the common or unique variates. (Recall that a one-way MANOVA is equivalent to a CVA). To see these points, consider the expected values of the transformed variables if the common variate structure exists. Then the expected value of the time profile variates is a function of the common variates and the group positions. That is,

$$\begin{aligned} \mathrm{E}\left(\mathbf{x}_{\mathrm{g}}^{\mathrm{q}}-\mathbf{x}_{\mathrm{g}}^{t}\right)=\boldsymbol{\mu}_{\mathrm{g}}^{\mathrm{q}}&=\boldsymbol{\mu}_{0}^{\mathrm{q}}+\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{1}\mathrm{e}_{\mathrm{g},1}^{\mathrm{q}}+\ldots+\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{c}\mathrm{e}_{\mathrm{g},c}^{\mathrm{q}}-\left(\boldsymbol{\mu}_{0p}+\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{1}\mathrm{e}_{\mathrm{g},1}^{t}+\ldots+\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{c}\mathrm{e}_{\mathrm{g},c}^{t}\right)\\ &=\left(\boldsymbol{\mu}_{0}^{\mathrm{q}}-\boldsymbol{\mu}_{0}^{t}\right)+\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{1}\left(\mathrm{e}_{\mathrm{g},1}^{\mathrm{q}}-\mathrm{e}_{\mathrm{g},1}^{t}\right)+\ldots+\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{c}\left(\mathrm{e}_{\mathrm{g},c}^{\mathrm{q}}-\mathrm{e}_{\mathrm{g},c}^{t}\right), \end{aligned}$$

for $\mathrm{q}=1,\ldots,t-1$. Now let $\boldsymbol{\mu}_{\mathrm{g}}^{*}$ be the $p(t-1)$ vector of expected values of the transformed variables; in other words the concatenation of $\boldsymbol{\mu}_{\mathrm{g}}^{\mathrm{q}}$, $\mathrm{q}=1,\ldots,t-1$. Similarly let $\boldsymbol{\mu}_{0}^{*}$ be the $p(t-1)$ vector of overall means. Then

$$\boldsymbol{\mu}_{\mathrm{g}}^{*}-\boldsymbol{\mu}_{0}^{*}=\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{1}\otimes\mathbf{e}_{\mathrm{g},1}^{*}+\ldots+\boldsymbol{\Sigma}_{\mathrm{w}}\mathbf{v}_{c}\otimes\mathbf{e}_{\mathrm{g},c}^{*}. \tag{8.22}$$

Observe that (8.22) has the form of a common variate model. The first implication of this is that if the common variate structure exists then the test for the time-treatment interaction is equivalent to a test for the equality of the group positions at various occasions, because if the original group positions are equal, the group positions for the transformed data will be zero. A second implication is that when one performs a CVA on the $p(t-1)$ transformed variables one has a situation similar to that in Section **8.7.2**. As argued in Section **8.7.2**, the variates generated will consist of subvectors which are linear compounds of the common variates. Furthermore, the logic of these arguments can be extended to data with unique variates as was done in Section **8.7.2** by replacing each unique variate with up to t common variates. By definition the group positions corresponding to these (extra) common variates cannot be equal over time. Thus this test also detects changes due to changing or unique canonical variates, though one will not be able to determine if the observed effects are due to changes in group positions or changes in unique variates.

In summary, the doubly multivariate approach answers the questions of whether there are differences among the groups, and if these differences change over (interact with) time. But it does not determine what changes over time; i.e., whether it is the variates or the group scores that change.

CHAPTER NINE

SCALING THE VARIABLES

Up to this point the scaling of the data and the (possible) scale invariance of the methods under consideration have been approached on an ad hoc basis in the examples. In this chapter I provide a more thorough discussion. To the best of my knowledge the only systematic treatment on the scale invariance of multivariate data is Jöreskog's (1989) discussion on scale invariance for the analysis of covariance structures. I try to extend his ideas to the methods considered in this dissertation. I show that my methods are often not scale invariant. Hence the issue of which scaling to employ is of obvious importance. I discuss several possible standardizations and when one would want to employ them.

9.1 AN EXAMPLE OF RESCALING THE DATA

Example 9.1

To emphasize the importance of the choice of scale, I begin with an example which shows how sensitive a solution may be to a rescaling of the data. I present a principal components analysis of the covariance matrix (a) before and after rescaling. For this example the rescaling

involves multiplying the first variable by 10 and the second by five. The rescaling could be the result of changing units; for example, when one converts from millimeters to centimeters. The rescaling is equivalent to multiplying the data matrix by the diagonal matrix (b), or by pre-multiplying and post-multiplying the covariance matrix (a) by (b). (c) is the matrix whose columns are the principal components of (a), ordered by the size of their eigenvalues. Compare (c) with (d), the matrix of principal components after the data have been rescaled. There is no nontrivial way to relate the principal components derived from the unscaled data (c) and that of the rescaled data (d).

(a) $\begin{bmatrix} 15 & 9 & -7 \\ 9 & 23 & 0 \\ -7 & 0 & 6 \end{bmatrix}$

(b) $\begin{bmatrix} 10 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

(c) $\begin{bmatrix} 0.578 & -0.598 & 0.554 \\ 0.798 & 0.558 & -0.228 \\ -0.172 & 0.573 & 0.801 \end{bmatrix}$

(d) $\begin{bmatrix} 0.926 & -0.372 & 0.061 \\ 0.375 & 0.926 & -0.048 \\ -0.039 & 0.067 & 0.996 \end{bmatrix}$

9.2 DEFINITIONS OF SCALE INVARIANCE

In point estimation for one location parameter, π, equivariance to scale is defined simply as

$$\hat{\pi}(cy_1, cy_2, \ldots, cy_n) = c\hat{\pi}(y_1, y_2, \ldots, y_n)$$

where $\hat{\pi}$ is the parameter estimate, c is a constant and y_i, $i = 1, \ldots, n$, are the data. Scale invariance needs to be defined more broadly for multivariate methods. For example, consider the relatively simple case of multiple regression. If one multiplies a regressor variable x_i by c, then $\hat{b}_i^* = \hat{b}_i / c$, where $\hat{b}_i$ and $\hat{b}_i^*$ are the estimates of the regression parameter associated with x_i before and after multiplication by c. This relationship of the parameter estimate to scale differs from that of the point estimation for a location parameter as seen above. Nevertheless, researchers consider multiple regression to be scale invariant for two reasons: because there is a simple relationship between each parameter estimate and the scale of its associated regressor variable, and because the parameter estimate for a given b_i is invariant to the scaling of those other regressors not associated with it.

Jöreskog (1989) gives a systematic discussion on scale invariance in covariance structure modeling (see Chapter Seven for a definition of covariance structure modeling). According to

Jöreskog one distinguishes between a model being scale invariant and a fit function being scale invariant. A covariance model $\Sigma(\theta)$ is scale invariant (Browne 1982) if for any diagonal matrix $\mathbf{D}$ of positive scalars and any parameter vector θ there exists another parameter vector θ^* such that

$$\Sigma(\theta^*) = \mathbf{D}\Sigma(\theta)\mathbf{D}.$$

Denote a fit function as $F(\mathbf{S},\Sigma)$, where $\mathbf{S}$ is an observed covariance matrix and Σ is a predicted covariance matrix. Then $F(\mathbf{S},\Sigma)$ is scale invariant (Jöreskog 1989) if for any diagonal matrix $\mathbf{D}$ of positive scalars the following is true:

$$F(\mathbf{DSD},\mathbf{D}\Sigma\mathbf{D}) = F(\mathbf{S},\Sigma). \qquad (9.1)$$

Maximum likelihood and generalized least squares are scale invariant fit functions; least squares is not (Jöreskog 1989). To see that generalized least squares is a scale invariant fit function, note that it minimizes the sum of squares of the deviations weighted by the inverse of the sample covariance matrix, $\mathbf{S}$. Generalized least squares satisfies (8.1) as

$$\mathrm{Tr}\left[\mathbf{D}^{-1}\mathbf{S}^{-1}\mathbf{D}^{-1}(\mathbf{DSD}-\mathbf{D}\Sigma\mathbf{D})\right]^2 = \mathrm{Tr}\left[\mathbf{S}^{-1}(\mathbf{S}-\Sigma)\right]^2.$$

If both the model and the fit function are scale invariant, then the analysis of the same variables in different scales yields results which are properly related; i.e. one can obtain $\hat{\theta}^*$ from $\hat{\theta}$ and $\mathbf{D}$. This is because scale invariance for the fit function implies that the global optima is the same for all scalings. Scale invariance for the model then implies that the parameter estimates under the various scalings can be related.

An additional point is that a parameter estimate, $\hat{\pi}$, is defined to be scale-free if for all $\mathbf{D}$ the following holds:

$$\hat{\pi}(\mathbf{DSD}) = \hat{\pi}(\mathbf{S}).$$

9.3 EXAMPLES OF SCALE INVARIANT METHODS

Multiple regression is an example of a scale invariant method where there is a simple relationship between the estimates of parameters under different choices of scale. However, scale invariance as defined in the previous section does not in itself imply a simple or easily interpretable relationship between the parameter estimates before and after scaling. For example, principal components analysis is scale invariant as defined above, although there is no simple relationship between the parameter estimates before and after rescaling as seen with multiple regression in Example **8.1**. The fit function for the full principal components model is scale invariant for either least squares or maximum likelihood estimation as the fit is always perfect. The principal components model, $\Sigma(\theta) = \mathbf{PLP}'$, where $\mathbf{P}$ is orthogonal and $\mathbf{L}$ diagonal, is scale invariant, as pre-multiplication and post-multiplication by $\mathbf{D}$ yields the following:

$$\Sigma(\theta^*) = \mathbf{DPLP}'\mathbf{D} = \mathbf{P}^*\mathbf{L}^*\mathbf{P}^{*\prime},$$

where $\mathbf{P}^*\mathbf{L}^*\mathbf{P}^{*\prime}$ is the singular value decomposition of $\mathbf{DPLP}'\mathbf{D}$. However, this relationship is trivial and not useful.

One can identify a class of models which will have a simple relationship between parameter estimates based on different choices of scale. If the matrices $\mathbf{F}_i$ of the model $\Sigma(\theta) = \sum_i \mathbf{F}_i \mathbf{M}_i \mathbf{F}_i'$ are unrestricted in their column space, e.g., not constrained to be orthogonal or of unit length, then clearly, $\mathbf{F}_i^* = \mathbf{D}\mathbf{F}_i$. Furthermore, $\mathbf{M}_i$ are scale-free. Factor analysis and multiple regression can both be put into this framework.

Canonical correlation analysis is also scale invariant. Clearly its estimation by maximum likelihood is scale invariant. Further, it can be expressed as a covariance structures model as is shown below, where **W**, **E** and **V** are defined as in Section **2.2.1**:

$$\Sigma(\theta) = \begin{bmatrix} \mathbf{S}_{XX} & \mathbf{S}_{XY} \\ \mathbf{S}_{YX} & \mathbf{S}_{YY} \end{bmatrix} = \begin{bmatrix} \mathbf{W}^{-1\prime}\mathbf{W}^{-1} & \mathbf{W}^{-1\prime}\mathbf{E}\mathbf{V}^{-1} \\ \mathbf{V}^{-1\prime}\mathbf{E}\mathbf{W}^{-1} & \mathbf{V}^{-1\prime}\mathbf{V}^{-1} \end{bmatrix}.$$

Let the diagonal matrix of scale terms be $\mathbf{D} = \begin{bmatrix} \mathbf{K} & \\ & \mathbf{F} \end{bmatrix}$. Then when one pre-multiplies and post-multiplies $\Sigma(\theta)$ by **D** one has the following matrix:

$$\mathbf{D}\Sigma(\theta)\mathbf{D} = \begin{bmatrix} \mathbf{K}'\mathbf{W}^{-1\prime}\mathbf{W}^{-1}\mathbf{K} & \mathbf{K}'\mathbf{W}^{-1\prime}\mathbf{E}\mathbf{V}^{-1}\mathbf{F} \\ \mathbf{F}'\mathbf{V}^{-1\prime}\mathbf{E}\mathbf{W}^{-1}\mathbf{K} & \mathbf{F}'\mathbf{V}^{-1\prime}\mathbf{V}^{-1}\mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{W}^{*-1\prime}\mathbf{W}^{*-1} & \mathbf{W}^{*-1\prime}\mathbf{E}^*\mathbf{V}^{*-1} \\ \mathbf{V}^{*-1\prime}\mathbf{E}^*\mathbf{W}^{*-1} & \mathbf{V}^{*-1\prime}\mathbf{V}^{*-1} \end{bmatrix}.$$

Thus $\mathbf{W}^* = \mathbf{W}\mathbf{K}^{-1}$, $\mathbf{V}^* = \mathbf{V}\mathbf{F}^{-1}$ and $\mathbf{E}^* = \mathbf{E}$. The relationship between the parameter estimates for the canonical variates before and after rescaling is the same as that of multiple regression. Note also that **E** is scale-free.

Estimating the full redundancy analysis (RA) model by least squares yields a perfect fit regardless of multiplication by **D**. Hence it is scale invariant with respect to fit function. RA can be expressed as a covariance structure model as follows:

$$\Sigma(\theta) = \begin{bmatrix} \mathbf{S}_{XX} & \mathbf{S}_{XY} \\ \mathbf{S}_{YX} & \mathbf{S}_{YY} \end{bmatrix} = \begin{bmatrix} \mathbf{W}^{-1\prime}\mathbf{W}^{-1} & \mathbf{W}^{-1\prime}\mathbf{E}\mathbf{V}' \\ \mathbf{V}\mathbf{E}\mathbf{W}^{-1} & \mathbf{V}\mathbf{E}^2\mathbf{V}' + \mathbf{J} \end{bmatrix},$$

where **W**, **R** and **V** are defined as in Section **2.2.2** and $\mathbf{J} = \mathbf{S}_{YY} - \mathbf{V}\mathbf{E}^2\mathbf{V}'$. Pre-multiply and post-multiply $\Sigma(\theta)$ by **D** as defined above. Then one has the following matrix:

$$\mathbf{D}\Sigma(\theta)\mathbf{D} = \begin{bmatrix} \mathbf{K}'\mathbf{W}^{-1\prime}\mathbf{W}^{-1}\mathbf{K} & \mathbf{K}'\mathbf{W}^{-1\prime}\mathbf{E}\mathbf{V}'\mathbf{F} \\ \mathbf{F}'\mathbf{V}\mathbf{E}\mathbf{W}^{-1}\mathbf{K} & \mathbf{F}'\mathbf{V}\mathbf{E}^2\mathbf{V}'\mathbf{F} + \mathbf{F}'\mathbf{J}\mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{W}^{*-1\prime}\mathbf{W}^{*-1} & \mathbf{W}^{*-1\prime}\mathbf{E}^*\mathbf{V}^{*\prime} \\ \mathbf{V}^*\mathbf{E}^*\mathbf{W}^{*-1} & \mathbf{V}^*\mathbf{E}^{*2}\mathbf{V}^{*\prime} + \mathbf{J}^* \end{bmatrix}.$$

One sees RA is model scale-invariant as defined in Section **9.2** as $\mathbf{W}^*$, $\mathbf{V}^*$, $\mathbf{E}^*$ and $\mathbf{J}^*$ can be found from **W**, **E**, **V** and **J**. However, their relationship to **W**, **V**, **E**, **J** is not simple. Nevertheless, if one rescales only the X-variables one has simple relationships between the parameter estimates as $\mathbf{W}^* = \mathbf{W}\mathbf{K}^{-1}$, $\mathbf{E}^* = \mathbf{E}$, $\mathbf{V}^* = \mathbf{V}$ and $\mathbf{J}^* = \mathbf{J}$.

The CVA/time model (7.3) presented in Chapter Seven is not scale invariant. This model can be expressed in the form $\mathbf{V}\mathbf{Q}_{ij}\mathbf{V}' = \mathbf{P}_{ij}$, where $\mathbf{P}_{ij}$ is the submatrix of the between-group covariance matrix corresponding to the covariance between the i^{th} and j^{th} occasions, and $\mathbf{Q}_{ij}$ is

a diagonal matrix. If this model were to have scale invariance then the following must be true: $\mathbf{DVQ}_{ij}\mathbf{V}'\mathbf{D} = \mathbf{V}^*\mathbf{Q}_{ij}^*\mathbf{V}^{*\prime}$ and $\mathbf{DVQ}_{i'j'}\mathbf{V}'\mathbf{D} = \mathbf{V}^*\mathbf{Q}_{i'j'}^*\mathbf{V}^{*\prime}$, where $i, j \neq i', j'$. This implies

$$\mathbf{V}^*\mathbf{Q}_{ij}^*\mathbf{Q}_{i'j'}^*\mathbf{V}^{*\prime} = \mathbf{DVQ}_{ij}\mathbf{V}'\mathbf{DDVQ}_{i'j'}\mathbf{V}'\mathbf{D}. \tag{9.2}$$

But (9.2) can be seen to be generally untrue. The term on the left is symmetric since $\mathbf{Q}_{ij}$ and $\mathbf{Q}_{i'j'}$ are diagonal. However, the term on the right is symmetric only if $\mathbf{Q}_{ij} = \mathbf{Q}_{i'j'}$ or if $\mathbf{V}'\mathbf{D}^2\mathbf{V}$ is diagonal. The former is generally not true and the latter is true only if $\mathbf{V}$ is the diagonal. Hence there is no way to relate the parameters of the model before and after a rescaling.

9.4 SCALE INVARIANCE FOR THREE-MODE PRINCIPAL COMPONENTS ANALYSIS

Kroonenberg (1983) and Harshman and Lundy (1984) have discussed the choice of scale for three-mode data. However, they never discuss scale invariance. Indeed, it seems the researchers who develop and use three-mode methods implicitly assume that scale invariance is unattainable for three-mode data. In this section I show that for certain three-mode models that a type of scale invariance or approximate scale invariance exists. The discussion of how to scale three-mode data when scale invariance does not exist is deferred until Section **9.5**.

I will provide a definition of scale invariance for three-mode PCA that is analogous to Jöreskog's definition of scale invariance for covariance structures. As in Jöreskog's development I will need to define both scale invariance for the fit function and scale invariance for the model. First of all, however, I must define what is meant by rescaling in the context of three-mode PCA. In the analysis of covariance structures there are two modes, a variables mode and an observations mode, and by definition one rescales only the variables mode. In three-mode PCA all three modes are candidates for rescaling. However, (approximate) scale invariance can exist only when one rescales one mode at a time. Hence I restrict myself to considering the rescaling of just one mode. The rescaling of three-way data is defined as follows: if one has g slices of the three-way array, $\mathbf{Z}_i, \ldots, \mathbf{Z}_g$, $i = 1, \ldots, g$, one can post-multiply them by $\mathbf{B}$, a diagonal matrix of positive scalars. Because of symmetry the following arguments generalize to any one of the three modes being rescaled.

Having defined choice of scale in the context of three-mode PCA, I can address the issue of scale invariance for the fit function. The least squares fit function is generally not invariant to scale. However, the Tucker2 and Tucker3 models can decompose a three-mode matrix exactly, which yields scale invariance; i.e., regardless of the scaling, the fit function is zero. If one chooses a solution with less than the full complement of components one can get approximate scale invariance if the fit is good. How good the fit must be remains to be determined.

Next I consider model scale invariance for various three-mode PCA models. The Tucker2 model can be shown to be invariant to column scaling. The Tucker2 is expressed in matrix form as follows, where $\mathbf{G}$, $\mathbf{C}_i$, $\mathbf{H}$ are defined as in Section **2.32**

$$\mathbf{Z}_i = \mathbf{GC}_i\mathbf{H}', \quad \text{for} \quad i = 1, \ldots, g.$$

Post-multiplying the $\mathbf{Z}_i$ terms by **B** yields the following:

$$\mathbf{Z}_i\mathbf{B} = \mathbf{G}\mathbf{C}_i\mathbf{H}'\mathbf{B} = \mathbf{G}^*\mathbf{C}_i^*\mathbf{H}^{*'}, \quad \text{for} \quad i = 1,\ldots,g. \tag{9.3}$$

To get $\mathbf{G}^*$, $\mathbf{C}_i^*$ and $\mathbf{H}^*$ in terms of **G**, $\mathbf{C}_i$, **H** and **B**, perform a singular value decomposition on $\mathbf{H}'\mathbf{B}$, yielding $\mathbf{H}'\mathbf{B} = \mathbf{M}\mathbf{N}\mathbf{P}'$. Then (9.3) becomes

$$\mathbf{G}\mathbf{C}_i\mathbf{M}\mathbf{N}\mathbf{P}' = \mathbf{G}^*\mathbf{C}_i^*\mathbf{H}^{*'}, \quad \text{for} \quad i = 1,\ldots,g. \tag{9.4}$$

Thus $\mathbf{G}^* = \mathbf{G}$, $\mathbf{C}_i^* = \mathbf{C}_i\mathbf{M}\mathbf{N}$ and $\mathbf{H}^{*'} = \mathbf{P}'$. The model is invariant to rescaling the column space of the $\mathbf{Z}_i$ in the sense that $\mathbf{G}^*$, $\mathbf{C}_i^*$ and $\mathbf{H}^*$ can be found for all **G**, $\mathbf{C}_i$, **H** and **B**. Further, **G** is $\mathbf{G}^*$. The core matrices, $\mathbf{C}_i^*$, and the components for the column space, $\mathbf{H}^*$, however, are not related to $\mathbf{C}_i$ and **H** in any simple or useful manner. With similar logic the Tucker3 can be shown to have scale invariance properties.

Now consider the PARAFAC model, first with two sets of components restricted to orthonormality. This is the model one gets when one requires diagonal $\mathbf{C}_i$ in (9.3). Clearly, in order for a solution to exist there must be orthonormal $\mathbf{G}^*$ and $\mathbf{H}^*$ that simultaneously diagonalize $\mathbf{G}\mathbf{C}_i\mathbf{H}'\mathbf{B}$ for $i = 1,\ldots,g$. As they generally do not exist the model is not scale invariant. However, the PARAFAC model without the orthonormality constraints is scale invariant. Again referring to (9.3), one sees that one can relate $\mathbf{G}^*$, $\mathbf{C}_i^*$ and $\mathbf{H}^*$ to **G**, $\mathbf{C}_i$, **H** and **B** by letting $\mathbf{G}^* = \mathbf{G}$, $\mathbf{H}^* = \mathbf{L}\mathbf{H}\mathbf{B}$ and $\mathbf{C}_i^* = \mathbf{C}_i\mathbf{L}^{-1}$, where $\mathbf{L} = \text{Diag}\big((\mathbf{H}\mathbf{B})'\mathbf{H}\mathbf{B}\big)$. Not only is **G** unchanged by the rescaling, but there is an interpretable relationship between the coefficients of $\mathbf{H}^*$ and **H** and between $\mathbf{C}_i^*$ and $\mathbf{C}_i$.

Example 9.2:

This example illustrates the approximate scale invariance of the Tucker2 method. $\mathbf{X}_1$ and $\mathbf{X}_2$ are the data matrices. These data are purposely constructed to fit the one component Tucker2 relatively poorly but to fit the two component Tucker2 excellently, with 59% of the sums of squares being explained by the former model but 99% by the latter. $\mathbf{X}_1$ and $\mathbf{X}_2$ of (a) are the data before rescaling, $\mathbf{X}_1^*$ and $\mathbf{X}_2^*$ of (b) are the data after rescaling. Note the extent to which the scaled and unscaled solutions differ for the one-component solution (c). On the other hand, the scaled and unscaled solutions for two-component matrices are recognizably similar. If the fit were perfect, then the there would exist perfect scale invariance.

(a)
$$\mathbf{X}_1 = \begin{bmatrix} -19.5 & 10.5 & -7 \\ -10.5 & 19.5 & 7 \\ 7 & 7 & 2 \end{bmatrix} \qquad \mathbf{X}_2 = \begin{bmatrix} -27.\overline{6} & 10.\overline{3} & -10.\overline{6} \\ -10.\overline{3} & 27.\overline{6} & 10.\overline{6} \\ 10.\overline{6} & 10.\overline{6} & 6.\overline{6} \end{bmatrix}$$

(b) $$\mathbf{X}_1 = \begin{bmatrix} -195 & 105 & -70 \\ -10.5 & 19.5 & 7 \\ 7 & 7 & 2 \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} -276.\overline{6} & 103.\overline{3} & -106.\overline{6} \\ -10.\overline{3} & 27.\overline{6} & 10.\overline{6} \\ 10.\overline{6} & 10.\overline{6} & 6.\overline{6} \end{bmatrix}$$

(c)

Original	Rescaled
0.7071	0.8680
–0.7071	-0.3760
0.0000	0.3243

(d)

Original		Rescaled	
0.7071	0.5774	0.6688	0.5906
-0.7071	0.5774	-0.7398	0.5887
0.0000	0.5774	0.07335	0.5520

9.5 HOW TO SCALE THE DATA

Since the models I propose are often not scale invariant, the choice of scale of the variables is salient. There are several plausible ways to scale the data. The simplest is to analyze the raw data. This scaling is appropriate only if the measures in their unscaled form are comparable. Some examples of this could be species counts or company sales in dollars for particular types of goods. Because measures are often not comparable, one must choose a scaling that makes them so. In many multivariate applications one "standardizes", or scales the variables to unit length. This standardization effectively gives each variable equal importance in the modeling. Another possibility is to apply the Mahalanobis transformation to the data by post-multiplying $\mathbf{Y}$ by $\mathbf{S}_{YY}^{-1/2}$.

Choosing an appropriate scaling for data over time is complicated by two things. First, the covariances may be changing over time. Hence a scaling that is appropriate for one occasion may not be appropriate for other occasions. Second, the covariance of the Y-variables is assumed to be related to the X-variables either in a causal manner or in a correlational manner. Hence one must decide whether to, and perhaps how to, remove the effect of the X-variables on the Y-variables. This issue is simplified if the X-variables are group indicators, because then one can calculate a within-groups covariance.

I first discuss the issue of standardization for the situation where the X-variables are group indicators. The simplest scenario is that the within-group covariances are assumed not to vary over time. Then one could standardize by post-multiplying $\mathbf{Y}$ by $\mathbf{D} = \text{Diag}^{-1/2}(\mathbf{S}_{YY})$ or apply the Mahalanobis transformation by post-multiplying $\mathbf{Y}$ by $\mathbf{S}_{YY}^{-1/2}$, where $\mathbf{S}_{YY}$ is a pooled estimate over time.

If the within-group covariances are assumed to vary over time the situation is more complex. One possible way to standardize such data is to choose a baseline occasion and use its within-group covariance to standardize. Another possibility is to average the variances over occasions and standardize by the average variance. That is, one can post-multiply **Y** by $\mathbf{D} = \text{Diag}^{-½}((\mathbf{S}_{YY1} + \mathbf{S}_{YY2} + \cdots + \mathbf{S}_{YY3}) / g)$, where g is the number of occasions. Such a standardization is akin to what is done in the factor analysis of multiple groups (Loehlin 1992) and is also recommended for some situations in three-mode PCA (Kroonenberg 1983). It gives each variable the same weight in the overall analysis while allowing the variances to vary over occasion. Applying the Mahalanobis transformation to the data based on a baseline or averaged covariance matrix is problematic as the transformed data will not satisfy $\mathbf{S}_{X^*X^*k} = \mathbf{I}$, for $k = 1, \ldots, g$, where $\mathbf{x}^* = \mathbf{S}_{XX}^{-½}\mathbf{x}$.

If the X-variables are not group indicators but continuous variables the standardization is further complicated because one does not have the elegant partitioning of the variation into between-group effects and within-group effects. One can express the matrix of the total sums of squares of the Y-variables as the sum of a regression sums of squares matrix and a residuals or error matrix. Then one could standardize by the residuals matrix. Such a standardization attempts to make the error terms equal for each observation, which is analogous to what is done when one standardizes by a within-group covariance matrix. For this standardization to be reasonable one should have X-variables that are controlled by the experimenter. If the X-variables are random variables in their own right, then the X-variables and Y-variables are correlated and one may prefer to standardize **Y** based on the total covariance matrix for the Y-variables.

The analysis of covariance structures and three-mode PCA are both methods based on the decomposition of matrices. In contrast, Campbell and Tomenson's model and the CVA/time model of Chapter Eight are means models. These means cannot be put into the framework used for evaluating the scale invariance of the analysis of covariance or three-mode PCA which begins by multiplying a mode by a diagonal matrix of positive scalars. However, there is another useful way to look at these models. Campbell and Tomenson's analysis is equivalent to plotting the group means in the space transformed by the Mahalanobis transformation and then finding a reduced space of common orthogonal variates in which the means approximately lay. This method effectively transforms the data by $\mathbf{S}_{YY}^{-½}$ and thus is scale invariant. Likewise, CVA/time with uncorrelated variates plots the group means in the transformed space and is scale invariant. On the other hand CVA/time with orthogonal variates does not transform the data and is not scale invariant. Hence one must standardize the data by one of the methods described in the previous paragraphs of this section.

This discussion on scaling should make clear that the researcher needs to give thoughtful consideration to the standardization and scaling he uses.

CHAPTER TEN

CONCLUSION AND FURTHER RESEARCH

The problem of greatest interest in this dissertation has been canonical variate analysis with measurements over time. I suggest three approaches: a maximum likelihood approach based on modeling the means (Chapter Eight): a least squares approach based on three-mode principal components (Chapter Five): and an analysis of covariances approach (Chapter Seven). What proves to be a unifying theme in these attempts to model CVA over time, indeed throughout the entire dissertaition, is that of the common variate over time (or over a third mode).

The most ambitious of the methods to model CVA over time is the maximum likelihood approach. In addition to modeling the means, I model the error terms over time. I provide a model for error which is constant over time and changes over time. I also work out estimating equations to solve for the parameter estimates and implement an algorithm to obtain the

estimates. Lastly I show that my approach is superior to doubly multivariate repeated measures in conceptualizing the problem of multivariate grouped data over time.

The least squares method I develop as an exploratory approach (Chapter Five). I show it to have attractive features such as the partitioning of the sums of squares and the nestedness of solutions (Chapter Three). I also develop graphical methods to be used in conjunction with it (Chapter Six).

The covariance structure analysis (COSAN, Chapter Seven) approach puts modeling CVA over time in an larger framework of methods that have an extensive usership, a developed theory, and powerful software. Despite this promise, however, more work needs to be done on the programming and on achieving convergence of the estimating algorithm.

In this dissertation I also approach several problems related to CVA over time. In Chapters Five and Six I propose a schema for modeling a broad variety of data with two sets of variables measured over a third mode. This includes modeling canonical correlation analysis (CCA), canonical variate analysis (CVA), redundancy analysis (RA), Procrustes Rotation (PR) and correspondence analysis with both data over time and multiple datasets.

In Chapter Four I develop a least squares approach to common principal components, which I show is comparable to the maximum likelihood method. I also extend the least squares method to common space analysis.

In Chapters Three and Nine I do not develop any new methods, but rather certain supporting ideas. In Chapter Three I show the partitioning of sums of squares and prove the nestedness of solutions for the PARAFAC (orth.) model. In Chapter Nine I extend the ideas for scale invariance previously developed for the analysis of covariance structures to three-mode principal components models.

I have just summarized the many problems tackled in this dissertation. The work done, however, suggests more problems to be solved. In particular, there is still much that can be done to develop and extend the ideas in the dissertation. What follows is a list of some of the possibilities.

- For the CVA/time model of Chapter Eight work out a method for handling missing values. Use Ware's (1985) approach to longitudinal regression which iteratively estimates the error matrix and the means model.
- Prepare a SAS macro for estimating the CVA/time model.
- Investigate the robustness of CVA/time model to heterogeneous variance over groups, outliers and non-normality.
- Develop a SAS macro for the least squares methods for CVA/third, CCA/third, RA/third and PR/third.
- Develop hypothesis tests and confidence intervals for the three-mode models. A possibility may be to use resampling methods.
- Investigate modeling the relationship between two sets of variables over time when one drops the restriction that both sets of variables be orthogonal.
- Develop a SAS macro for COSAN modeling for canonical variate analysis and redundancy analysis over time.

- Extend the COSAN model to include uncorrelated canonical variates over time and unique variates at each occasion.
- Develop hypothesis tests and confidence intervals for COSAN models.
- Investigate the modeling of error structure over time in the COSAN approach to canonical variate analysis over time.
- Compare the maximum likelihood, least squares and COSAN approaches to CVA over time.
- Investigate how good a Tucker2 model must fit in order to have approximate scale invariance.

BIBLIOGRAPHY

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley & Sons.

Browne, M.W. (1982). Covariance structures. In D.M. Hawkins (Ed.), *Topics in Applied Multivariate Analysis*. Cambridge: Cambridge University Press.

Cailiez, F., & Paigès, J.P. (1976). *Introduction à l`analyse des données* [Introduction to data analysis]. Paris: SMASH

Campbell, N.A., & Atchley, W.R. (1981). The geometry of canonical variate analysis. *Systematic Zoology, 30(3),* 268-280.

Campbell, N.A., & Tomenson, J.A. (1983). Canonical variate analysis for several sets of data. *Biometrics, 39,* 425-435.

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.

Carroll, J.D., Pruzansky, S., & Kruskal, J.B., (1980). CANDELINC: a general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. *Psychometrika, 45*, 3-24.

Diggle, P., Liang, K.Y. & Zeger, S.L., (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*, 211-218.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-184.

Flury, B.N. (1984). Common principal components in k groups. *Journal of the American Statistical Association, 79,* 892-898.

Flury, B.N.,& Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite matrices to nearly diagonal form. *SIAM J. Scientific and Statistical Computing, 7,* 169-184.

Flury, B.N. (1987). Two generalizations of the common principal component model. *Biometrika, 74,* 59-69.

Flury, B. (1988). *Common Principal Components and Related Multivariate Models*. New York: Wiley.

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*, 453-467.

Gittins, R. (1985). *Canonical analysis: a review with applications in ecology*. Berlin: Springer Verlag.

Good, I.J. (1969). Some applications of the singular value decomposition of a matrix. *Technometrics, 11,* 823-831.

Gower, J.C. (1975). Generalized Procrustes analysis. *Psychometrika, 40*, 33-51.

Greenacre, M. (1984). *Theory and applications of correspondence analysis*. Orlando, Florida: Academic Press.

Hand, D., & Crowder, M. (1996). *Practical Longitudinal Data Analysis*. London: Chapman & Hall.

Harshman, R.A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonetics, 16,* 1-84.

Harshman, R.A.,& Lundy, M.E. (1984). The PARAFAC model for three-way factor analysis and multidimensional scaling. In H.G. Law, C.W. Snyder, J.A. Hattie, & R.P. McDonald (Eds.), *Research Methods for Multimode Data Analysis* (pp. 122-215)*,* New York: Praeger,

Harshman, R.A.,& Lundy, M.E. (1994). PARAFAC: parallel factor analysis. *Computational Statistics and Data Analysis, 18, 39-72.*

Hotelling, H. (1935). The most predictable criterion. *Journal of Education Psychology*, *26,* 139-142.

Israëls, A.Z. (1984). Redundancy analysis for qualitative variables. *Psychometrika, 49*, 331-346.

Israëls, A.Z. (1987). *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Johansson, J.K. (1981). An extension of Wollenberg's redundancy analysis. *Psychometrika, 46*, 93-103.

Jöreskog, K.G. (1989). *LISREL 7 A guide to the program and applications, 2nd Edition*. Chicago: SPSS Inc.

Jöreskog, K.G. (1979). Statistical estimation of structural models in longitudinal developmental investigations. *Longitudinal research in in the study of behavior and development.* eds. J.R. Nesselroade & P.B. Baltes, New York: Academic Press.

Kiers, H.A.K. (1991). Hierarchical relations among three-way methods. *Psychometrika, 56(3),* 449-470.

Kroonenberg, P.M. (1983). *Three-mode principal components analysis*. Leiden: DSWO Press.

Kroonenberg, P.M., & De Leeuw, J (1977). TUCKALS2: A principal component analysis of three-mode data. *Res. Bull. RB 001-77*, Department of Data Theory, University of Leiden, Leiden, the Netherlands.

Krzanowski, W.J. (1979). Between-group comparison of principal components. *Journal of the American Statistical Association, 74,* 703-707.

Krzanowski, W.J. (1984). Principal component analysis in the presence of group structure. *Applied Statistics, 33,* 164-168.

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis*, Oxford: Oxford University Press.

Kshirsagar, A.M. (1972). *Multivariate analysis*. New York: Dekker.

Leurgans, S., & Ross, R. T. (1992). Multilinear models: applications in spectroscopy. *Statistical Science, 7,* 289-319.

Liang, K.Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13-22.

Loehlin, J.C. (1992). *Latent Variable Models*. Hillsdale, New Jersey: Lawrence Erlbaum Associtates.

Lynch, D.D., & Dise, N.B. (1985). *Sensitivity of stream basins in Shenandoah National Park to acid deposition.* U.S. Geological Survey, Water-Resources Investigations Report 85-4115.

McDonald, R.P., (1978). A simple comprehensive model of the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 31,* 59-72.

McDonald, R.P., (1980). A simple comprehensive model of the analysis of covariance structures: Some remarks on applications. *British Journal of Mathematical and Statistical Psychology, 33,* 161-183.

Meredith, W. (1964). Canonical correlations with fallible data. *Psychometrika, 29,* 55-65.

Penrose, R. (1955). On the best approximate solutions of linear matrix equations. *Proceedings of the Cambridge Philosophical Society, 51*, 406-413.

Rao, C.R., (1964). The use and interpretation of principal component analysis in applied research. Sankhyã: The Indian Journal of Statistics: Series A.

SAS, (1990). *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume1*. Cary: SAS Institute Inc.

Swaminathan, H., (1984). The factor analysis of longitudinal data. In H.G. Law, C.W. Snyder, J.A. Hattie, & R.P. McDonald (Eds.), *Research Methods for Multimode Data Analysis* (pp. 308-332)*,* New York: Praeger,

Ter Braak, C.J.F. (1990). Interpreting canonical correlation analysis through biplot of structure correlations and weights. *Psychometrika, 55(3),* 519-531.

Tucker, L.R. (1972). Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika, 37*, 3-27.

Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31,* 279-311.

Tyler, D.E. (1982). On the optimality of the simultaneous redundancy transformations. *Psychometrika, 47*, 77-86.

Van der Burg, E & De Leeuw, J. (1983). Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology, 36, 54-80.*

Van de Geer, J.P. (1986). *Introduction to multivariate data analysis, Vol. 1*. Leiden: DSWO Press.

Van de Geer, J.P. (1984). Linear relations among *k* sets of variables. *Psychometrika, 49,* 79-94.

Van den Wollenberg, A.L. (1977). Redundancy analysis: an alternative to canonical correlation analysis. *Psychometrika, 42*, 207-219.

Wald, A. (1945). *Annals of Mathematical Statistics, 16*, 117-186.

Ware, J.H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician, 39(2),* 95-101.

Wilks, S.S., (1962). *Mathematical Statistics*. New York: John Wiley & Sons.

Wolfram, S. (1991). *Mathematica*. Champaign: Wolfram Media.

APPENDIX ONE

THE KRONECKER PRODUCT OF TWO MATRICES

Let **A** be a $p \times m$ matrix and **B** a $q \times n$ matrix, let a_{ij} denote the element in the i^{th} row and j^{th} column of **A** and let b_{rs} denote the element in the i^{th} row and j^{th} column of **B**. The $pq \times mn$ matrix with $a_{ij}b_{rs}$ as the element in the $(iq + r)^{th}$ row and the $(jn + s)^{th}$ column is called the Kronecker or direct product of **A** and **B** is denoted by $\mathbf{A} \otimes \mathbf{B}$; that is,

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1m}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2m}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{p1}\mathbf{B} & a_{p2}\mathbf{B} & \cdots & a_{pm}\mathbf{B} \end{bmatrix}.$$

APPENDIX TWO

THE F-G ALGORITHM

What follows is SAS code for Flury's & Gautschi's F-G algorithm, with adaptation to get leasts squares estimates, applied to Fisher's iris data.

```
proc iml;

*C1, C2 and C3 contain the sample covariance matrices for the three species
 of iris.;

C1={26.6433 8.5184 18.2898 5.5780,
     8.5184 9.8469  8.2653 4.1204,
    18.2898 8.2653 22.0816 7.3102,
     5.5780 4.1204  7.3102 3.9106};

C2={40.4343  9.3763 30.3290 4.9094,
     9.3763 10.4004  7.1380 4.7629,
    30.3290  7.1380 30.4588 4.8824,
     4.9094  4.7629  4.8824 7.5433};

C3={12.4249  9.9216  1.6355 1.0331,
     9.9216 14.3690  1.1698 0.9298,
     1.6355  1.1698  3.0159 0.6069,
     1.0331  0.9298  0.6069 1.1106};



*The F-G algorithm consists of the macro Gstep nested within the macro
 Fstep.  In the Fstep every pair of column vectors of the current
 approximation to B is rotated such that the normal equations are satisfied.
 The Gstep determines these rotations;

*P is the number of variables, K the number of datasets or groups and L is
 the number of iterations.  B is the matrix of common principal components.

%macro Gstep;  M=j(2,2,0);
  %do i=1 %to &K;  H=B[,{&d &e}]; T&i=H`*C&i*H;
    d1&i=Q[,1]`*T&i*Q[,1];   d2&i=Q[,2]`*T&i*Q[,2];
```

```
    *The line of code immediately below is for the least squares solution.
     The line of code following it is for the maximum likelihood algorithm.;

    *M=M + (d1&i-d2&i)*T&i;
    M=M + (d1&i-d2&i)/(d1&i*d2&i)*T&i;
  %end;
%mend;

*The Fstep follows.  Bold stores the matrix B from the previous iteration.;

%macro Fstep;  %let P=4;   %let K=3;   %let L=15;  B=I(&P); Bold=I(&P);
  %do s=1 %to &L;
    %do d=1 %to &P;
      %do e=1 %to &d;    Q=I(2);
        if &e<&d then do;  c=0;
          do until(c=4);
            %Gstep
           normala=Q[,1]`*M*Q[,2];
           Q=eigvec(M);
            B[,{&d &e}]=H*Q;
            c=c+1;
          end;
        If &s=&L then do;
         normal=B[,&d]`*B[,&e];  print &d &e normal;
        end;
        end;
      %end;
    %end;
    Crit=SSQ(B-Bold); print B bold;Bold=B;
    phi=1; SSLF=0;
    %do i=1 %to &K;
      R&i=B`*C&i*B;
      phi=phi*det(diag(R&i))/det(R&i);
      SSLF=SSLF+SSQ(R&i-diag(R&i));
    %end;    print crit phi sslf;
  %end;
%mend; %Fstep

*The following macro determines the eigenvalues;

%macro eigen;   %let P=4;   %let K=3;   %let L=15;
%do i=1 %to &K;
  eig&i=diag(B`*c&i*B)*j(&P,1,1);
  print eig&i;
%end;
%mend;
%eigen
```

APPENDIX THREE

THE PROGRAMS FOR PARAFAC (ORTH.) AND TUCKER2 FOR THE SHENANDOAH EXAMPLE

```
*Program for the PARAFAC with orthogonal variates and the Tucker2;

*First the streamwater variables are standardized;

libname shen 'c:\prg\shen';
%macro first;
data n&L.; set shen.table&L.;
keep site disch cond ph temp ca mg na k alk so4 c1 si no3 nh4;
disch=v3; cond=v4; ph=v5; temp=v6; ca=v7; mg=v8; na=v9; k=v10; alk=v11;
so4=v12; c1=v13; si=v14;
if v15='<1' then no3=0.5; else no3=v15;if v16='<1' then nh4=0.5; else nh4=v16;
if site=2032589 or site=1628910 or site=1630542 then zzz=1;
else output;

proc standard data=n&L. replace mean=0out=nm&L.;   run;
proc corr data=nm&L. noprint nocorr cov out=st&L.(type=cov); run;
%mend first;
%let L=a; %first %let L=b; %first %let L=c; %first
%let L=d; %first %let L=e; %first %let L=f; %first



proc iml;
use sta; read all var {site disch cond ph temp ca mg na k alk so4 c1 si
 no3 nh4} into stna;
use stb; read all var {site disch cond ph temp ca mg na k alk so4 c1 si
 no3 nh4} into stnb;
use stc; read all var {site disch cond ph temp ca mg na k alk so4 c1 si
 no3 nh4} into stnc;
use std; read all var {site disch cond ph temp ca mg na k alk so4 c1 si
 no3 nh4} into stnd;
use ste; read all var {site disch cond ph temp ca mg na k alk so4 c1 si
 no3 nh4} into stne;
use stf; read all var {site disch cond ph temp ca mg na k alk so4 c1 si
 no3 nh4} into stnf;
```

```
*The streamwater variables are standardized such that the total variance
 for each variable over the six occasions is one.;

ssva=(stna+stnb+stnc+stnd+stne+stnf)/6; ssvar=ssva[2:15,2:15];
store ssvar;    quit iml;



%macro mann;

proc sort data=n&L.; by site;proc sort data=shen.table2; by site; run;
data table2n; set shen.table2;
keep site anti hamp wev cat sr pedl or ab2400 dd ew dev;
anti=v4; hamp=v5; wev=v6; cat=v7; sr=v8; pedl=v9; or=v10; ab2400=v11;
dd=v12; ew=v13; dev=v14;
run;



*The geological variables are merged with the streamwater variables.;

data t2&L.; merge table2n(in=tab2) n&L.(in=&L.); by site; drop site sr wev;
if &L.=1 then output t2&L.;




*Covariance matrices are created for each occasion.;

proc corr data=t2&L. nocorr noprint cov out=mad(type=cov); run;
proc iml; use mad; load;
read all var {anti hamp cat pedl or ab2400 dd ew dev
disch cond ph temp ca mg na k alk so4 c1 si no3 nh4} into cov&M.;
ssvart=inv(root(diag(ssvar)));
sxx&L.=cov&M.[1:9,1:9];
syy&L.=ssvart*cov&M.[10:23,10:23]*ssvart;
sxy&L.=cov&M.[1:9,10:23]*ssvart;


*The matrices c1-c6 are what will be modeled;

c&M.=inv(root(sxx&L.)`)*sxy&L.;  store c&M. cov&M.;
%mend mann;

%let L=a; %let M=1; %mann
%let L=b; %let M=2; %mann
%let L=c; %let M=3; %mann
%let L=d; %let M=4; %mann
%let L=e; %let M=5; %mann
%let L=f; %let M=6; %mann

m=0;    load;



*The PARAFAC model withorthogonality constraints follows.  K is the matrix
 of orthogonal variates corresponding to the geological variables.  L is the
 matrix of orthogonalvariates corresponding to the streamwater variates.
 The K and L that immediately follow are matrices of initial values.;
```

```
K={1 0 0 0, 0 1 0 0, 0 0 1 0,
   0 0 0 1, 0 0 0 0, 0 0 0 0,
   0 0 0 0, 0 0 0 0, 0 0 0 0};

L={1 0 0 0, 0 1 0 0, 0 0 1 0,
   0 0 0 1, 0 0 0 0, 0 0 0 0,
   0 0 0 0, 0 0 0 0, 0 0 0 0,
   0 0 0 0, 0 0 0 0, 0 0 0 0,
   0 0 0 0, 0 0 0 0, 0 0 0 0};


*Below follows the alternating least squares algorithm.  K and L are
 at each iteration estimated in a regression that assumes that the rest of
 the parameters are fixed.;

%let m=1; %let n=6;
%macro diag; %do i=&m %to &n; D&i=(diag(K`*C&i*L));
%end; %mend; %diag

ssb=1000000000;

do until (abscrit<0.000001);

%macro sumU;
U=0;
%do i=&m %to &n;
U= U+C&i*L*D&i; %end; %mend;  %sumu

%macro sumV;
V=0;
%do i=&m %to &n;
V= V+C&i`*K*D&i; %end; %mend;  %sumv

K=U*inv(root(U`*U));
L=V*inv(root(V`*V));
          ssa=0;
%macro diagssa; %do i=&m %to &n; D&i=(diag(K`*C&i*L));
ssa=ssa + trace(C&i`*C&i) - 2#trace(C&i`*K*D&i*L`) + trace(D&i**2);
%end; %mend; %diagssa

m=m+1;  crit=ssb-ssa; ssb=ssa;
        abscrit=abs(crit);
print m ssa crit abscrit;
end;

print K L;


*This step just prints the core matrix;

e=j(4,1,1);
D1=d1*e; D2=d2*e; D3=d3*e; D4=d4*e; D5=d5*e; D6=d6*e;
print D1 D2 D3 D4 D5 D6;




*Next follows the Tucker2 model with four geological variable and four
 streamwater variable components.  G is the matrix of orthogonal variates
 corresponding to the geological variables.  H is the matrix of orthogonal
 variates corresponding to the streamwater variates.;

p=0; ssb=0;
```

```
G={1 0 0 0, 0 1 0 0, 0 0 1 0, 0 0 0 1, 0 0 0 0, 0 0 0 0, 0 0 0 0, 0 0 0 0,
   0 0 0 0};
H={1 0 0 0, 0 1 0 0, 0 0 1 0, 0 0 0 1, 0 0 0 0, 0 0 0 0, 0 0 0 0, 0 0 0 0,
   0 0 0 0, 0 0 0 0, 0 0 0 0, 0 0 0 0, 0 0 0 0, 0 0 0 0};


%let m=1; %let n=6; Gold=G; Hold=H; m=&m; n=&n;

do until (abscrit<0.00000001);

%macro sumU;
U=0;
%do i=&m %to &n;
U= U+C&i*H*H`*C&i`; %end; %mend;  &sumu
G=U*G*inv(root(G`*(U**2)*G));

%macro sumV;
V=0;
%do i=&m %to &n;
V= V+C&i`*G*G`*C&i; %end; %mend;  &sumv
H=V*H*inv(root(H`*(V**2)*H));

%macro sumsqu;
ssa=0;
%do i=&m %to &n;     D&i=G`*C&i*H;  Ce&i=G*D&i*H`;
ssa=ssa + trace((C&i-Ce&i)`*(C&i-Ce&i));
%end; %mend; &sumsqu

Gcrit=ssq(G-Gold);Hcrit=ssq(H-Hold);
Gold=G; Hold=H;
p=p+1;  crit=ssb-ssa;  ssb=ssa;
        abscrit=abs(crit);

print p ssa crit Gcrit Hcrit; end;
print G H D1 D2 D3 D4 D5 D6;
```

APPENDIX FOUR

CODE FOR PLOTS AND GRAPHS

```
*Joint plot for the sum of the core matrices.  This sum is dtot.
Gmarker is the matrix of vectors to be plotted for the geological
variables and hmarker is the vector for the streamwater variables.;


dtot=(d1+d2+d3+d4+d5+d6)/6;    ddtot=diag(dtot[1:2]);
ll=9; m=14;
 const=(ll/m)##(1/2);   call svd(u,d,v,ddtot);
 gmarker=const*cvxs[,1:2]*u*sqrt(diag(d));
 hmarker=(1/const)*L[,1:2]*v*sqrt(diag(d));
 ghstar=gmarker//hmarker;  xya=ghstar[1:ll+m, 1:2]; or={0 0};
xy=or//xya;
options ls=67; reset pagesize=39;
*id={'O', 'g1', 'g2', 'g3', 'g4', 'g5', 'g6', 'g7', 'g8', 'g9', 'h1',
  'h2', 'h3', 'h4', 'h5', 'h6', 'h7', 'h8', 'h9', 'h10', 'h11', 'h12',
'h13',
  'h14'};
id={'O', '1', '2', '3', '4', '5', '6', '7', '8', '9', 'a',
  'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'm', 'n',
  'p'};
xlabel='Component 1';
      title='GH Joint Plot for Sum of Core Matrices';  ylabel='Component
2';
reset clip;
call pgraf(xy,id, xlabel, ylabel,title);



*This code creates the plot of the scores of the geological variables on
the streamwater components.  One plot is created for each of the four
streamwater components, these are denoted by scr1, scr2, scr3 and scr4.;


D1=diag(d1); d2=diag(d2); D3=diag(d3); d4=diag(d4); D5=diag(d5);
d6=diag(d6);
scor1=K*(D1[,1]||D2[,1]||D3[,1]||D4[,1]||D5[,1]||D6[,1]);
scr1=(j(9,1,1)||scor1[,1])//(j(9,1,2)||scor1[,2])//(j(9,1,3)||scor1[,3])
 //(j(9,1,4)||scor1[,4])//(j(9,1,5)||scor1[,5])//(j(9,1,6)||scor1[,6]);
```

```
scor2=K*(D1[,2]||D2[,2]||D3[,2]||D4[,2]||D5[,2]||D6[,2]);
scr2=(j(9,1,1)||scor2[,1])//(j(9,1,2)||scor2[,2])//(j(9,1,3)||scor2[,3])
 //(j(9,1,4)||scor2[,4])//(j(9,1,5)||scor2[,5])//(j(9,1,6)||scor2[,6]);


scor3=K*(D1[,3]||D2[,3]||D3[,3]||D4[,3]||D5[,3]||D6[,3]);
scr3=(j(9,1,1)||scor3[,1])//(j(9,1,2)||scor3[,2])//(j(9,1,3)||scor3[,3])
 //(j(9,1,4)||scor3[,4])//(j(9,1,5)||scor3[,5])//(j(9,1,6)||scor3[,6]);

scor4=K*(D1[,4]||D2[,4]||D3[,4]||D4[,4]||D5[,4]||D6[,4]);
scr4=(j(9,1,1)||scor4[,1])//(j(9,1,2)||scor4[,2])//(j(9,1,3)||scor4[,3])
 //(j(9,1,4)||scor4[,4])//(j(9,1,5)||scor4[,5])//(j(9,1,6)||scor4[,6]);

ylabel='score'; xlabel='occasion'; title='scores';

id={'1', '2', '3', '4', '5', '6', '7', '8', '9',
    '1', '2', '3', '4', '5', '6', '7', '8', '9',
    '1', '2', '3', '4', '5', '6', '7', '8', '9',
    '1', '2', '3', '4', '5', '6', '7', '8', '9',
    '1', '2', '3', '4', '5', '6', '7', '8', '9',
    '1', '2', '3', '4', '5', '6', '7', '8', '9'};
options ls=72;  reset pagesize=50;
call pgraf(scr1,id, xlabel, ylabel,title);
options ls=72;  reset pagesize=35;
call pgraf(scr2,id, xlabel, ylabel,title);
options ls=72;  reset pagesize=28;
call pgraf(scr3,id, xlabel, ylabel,title);
options ls=72;  reset pagesize=40;
call pgraf(scr4,id, xlabel, ylabel,title);



*What follows is the code for the residual plots for the Tucker2.  For
the
 the PARAFAC model with orthogonality constraints set G=K and H=L;


*G=K; *H=L;
%macro sqres;
ssqres=j(9,14,0); sst=j(9,14,0);  ssfit=j(9,14,0);
%do i=1 %to 6;
  D&i=diag(d&i);  Ce&i=G*D&i*H`;
  sqres&i=(C&i-Ce&i)##2;    sst&i=(C&i)##2;  ssfit&i=sst&i-sqres&i;
  ssqres=sqres&i+ssqres;    sst=sst&i+sst;   ssfit=ssfit&i+ssfit;
%end;
ssqrsvar=j(1,14,0);  ssqftvar=j(1,14,0);
%do j=1 %to 9;
  ssqrsvar=ssqres[&j,] + ssqrsvar;
  ssqftvar=ssfit[&j,] + ssqftvar;
%end;

ssqrssb=j(9,1,0);  ssqftsb=j(9,1,0);
%do j=1 %to 14;
  ssqrssb=ssqres[,&j] + ssqrssb;
  ssqftsb=ssfit[,&j] + ssqftsb;
%end;

%mend;
%sqres
```

```
print ssqrsvar ssqftvar; print  ssqrssb ssqftsb;

seev=sum(ssqrsvar); seeb=sum(ssqrssb); print seev seeb;
xy=ssqftvar`||ssqrsvar`;
options ls=70; reset pagesize=25;
id={ 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'm', 'n',
'p'};
xlabel='Sums of Squares Explained';
      title='Residual Plots';  ylabel='Sums of Squares Lack of Fit';
reset clip;
call pgraf(xy,id, xlabel, ylabel,title);
```

APPENDIX FIVE

THE PROGRAM CODE FOR THE COSAN MODEL

The method outlined fixes certain values of **V**, the matrix of orthogonal variates, to be zero, depending on how many canonical variates are to be modeled. Those variates to be estimated in the model are unrestricted. The columns of **V** not modeled are dummy variates for the purpose of making **V** orthogonal. Suppose there are *t* of these dummy variates. If $t = 1$, then the first *p*-1 columns of **V** determine the p^{th} column and it is not necessary to fix any elements of **V** to be zero. If $t = 2$, then the first *p*-2 columns are unrestricted. Set an arbitrary element of the $(p\text{-}1)^{\text{th}}$ vector to be zero to fix it, and the and the p^{th} vector follows. If $t = 3$, then set two elements of the $(p\text{-}2)^{\text{th}}$ column to be zero, and one element of the $(p\text{-}1)^{\text{th}}$ vector. One continues in this manner to constrain the desired number of variates.

Now in the COSAN model these elements of **V** cannot be directly set equal to zero because **V** is not directly estimated. Instead, **H** is directly estimated, where **V** is a function of **H**, $\mathbf{V} = (\mathbf{I} - \mathbf{H}')^{-1}(\mathbf{I} - \mathbf{H})$. However, since specific elements of **H** do not correspond to specific elements of **V**, the restrictions must be put on $(\mathbf{I} - \mathbf{H}')^{-1}(\mathbf{I} - \mathbf{H})$ indirectly. This is done by pre-multiplying and post-multiplying $(\mathbf{I} - \mathbf{H}')^{-1}(\mathbf{I} - \mathbf{H})$ by specific matrices to select for the appropriate elements. These are then set equal to zero by placing zeros in the data matrix.

Pre-multiply **V** $(= (\mathbf{I}_{(p)} + \mathbf{H})^{-1}(\mathbf{I}_{(p)} - \mathbf{H}))$ by a diagonal matrix with ones on the last *t*-1 diagonal elements and call this matrix **J**. Post-multiply **V** by a matrix with a 1 on the t^{th} diagonal element, and zeros elsewhere and call this matrix **K**. This yields a matrix with zeros in all columns except the t^{th} column, which has the last *t*-1 elements of the t^{th} column of **V**. Call this **U**. Then in the COSAN model set $\mathbf{U}\mathbf{U}'$ equal to a $p \times p$ matrix of zeros.

For example, to set three elements of the $(p-3)^{\text{th}}$ column of **V** to zero, choose **J** and **K** such that $\mathbf{U} = \mathbf{JVK}$, where:

$$\mathbf{U}=\begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & 0 & & \\ & & & & 1 & \\ & & & & & 1 \end{bmatrix}\mathbf{V}\begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & 1 & & \\ & & & & 0 & \\ & & & & & 0 \end{bmatrix}=\begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & 0 & 0 & 0 \\ & & & v_{p-4,p-3} & 0 & 0 \\ & & & v_{p-5,p-3} & 0 & 0 \\ & & & v_{p-6,p-3} & 0 & 0 \end{bmatrix}$$

Now in the COSAN model one places $\mathbf{U}\mathbf{U}'=\mathbf{Z}$, where $\mathbf{Z}$ is a $p\times p$ matrix of zeros. Thus one has set the three elements of $\mathbf{V}$ to zero, $v_{p-4,p-3}$, $v_{p-5,p-3}$ and $v_{p-6,p-3}$. Such a statement must be set up for each dummy variate in the model except the first.

To the data matrix, that is, the between groups covariance matrix $\mathbf{S}$, append p rows and p columns of zeros, so

$$\mathbf{S}^{*}=\begin{bmatrix} \mathbf{S} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

In the code note the following. The fit function is least squares. There are 42 variables, 28 for the streamwater variables, 14 for the dummy variables. The first line of the COSAN statement is

```
t(42,gen,inv)*v(42,gen)*r(42,sym)*p1(42,sym).
```

The block diagonal matrices for $\mathbf{V}$ are `t(42,gen,inv)*v(42,gen).` `r(42,sym)` is a matrix whose diagonal elements have the root variances explained. The next lines in the COSAN statement, starting with

```
s1a(42,sym)*ta(42,gen,inv)*va(42,gen)*s1b(42,sym)
```

restrict the selected elements of $\mathbf{V}$ to be zero as described earlier; `Ta(42,gen,inv)*va(42,gen)` is a matrix with $\mathbf{V}$ in its bottom block diagonal; `S1a(42,sym)` and `s1b(42,sym)` correspond to the matrices $\mathbf{J}$ and $\mathbf{K}$ from above.

The full code follows below.

```
libname shen 'c:\prg\shen';
proc calis cov data=shen.z11 method=U outest=cal
omethod=cg maxiter=1000 maxfunc=2000 nostderr privec; title ' ';


*The variables statement defines the variables.  Variables with an "a" as a
 suffix indicate the first occasion, those a "b" indicate the second
 occasion, and those with a "z" indicate the the dummies which are set to
 zero.;



var discha conda pha tempa caa mga naa ka alka so4a c1a sia no3a nh4a
    dischb condb phb tempb cab mgb nab kb alkb so4b c1b sib no3b nh4b
    dischz condz phz tempz caz mgz naz kz alkz so4z c1z siz no3z nh4z;
```

```
*The "cosan" statement defines the model.  t, v, r and p are matrices.  In
 the model the inverse of t is modeled.;

cosan t(42,gen,inv)*v(42,gen)*r(42,sym)*p1(42,sym)+
      s1a(42,sym)*ta(42,gen,inv)*va(42,gen)*s1b(42,sym)+
      s2a(42,sym)*ta(42,gen,inv)*va(42,gen)*s2b(42,sym)+
      s3a(42,sym)*ta(42,gen,inv)*va(42,gen)*s3b(42,sym)+
      s4a(42,sym)*ta(42,gen,inv)*va(42,gen)*s4b(42,sym)+
      s5a(42,sym)*ta(42,gen,inv)*va(42,gen)*s5b(42,sym)+
      s6a(42,sym)*ta(42,gen,inv)*va(42,gen)*s6b(42,sym)+
      s7a(42,sym)*ta(42,gen,inv)*va(42,gen)*s7b(42,sym)+
      s8a(42,sym)*ta(42,gen,inv)*va(42,gen)*s8b(42,sym)+
      s9a(42,sym)*ta(42,gen,inv)*va(42,gen)*s9b(42,sym)+
      s10a(42,sym)*ta(42,gen,inv)*va(42,gen)*s10b(42,sym)+
      s11a(42,sym)*ta(42,gen,inv)*va(42,gen)*s11b(42,sym)+
      s12a(42,sym)*ta(42,gen,inv)*va(42,gen)*s12b(42,sym);


*The "matrix" statement sets the matrix elements equal to parameters or
 constants.  "{1,2}=v12"  sets the element in the first row and second
 column equal to the parameter v12.  Matrix elements not specified default
 to zero values.;

matrix t
 {1,1}=-1, {2,2}=-1, {3,3}=-1, {4,4}=-1, {5,5}=-1, {6,6}=-1, {7,7}=-1,
 {8,8}=-1, {9,9}=-1, {10,10}=-1, {11,11}=-1,{12,12}=-1,{13,13}=-1,{14,14}=-1,
 {15,15}=-1,{16,16}=-1,{17,17}=-1,{18,18}=-1,{19,19}=-1,{20,20}=-1,{21,21}=-1,
 {22,22}=-1,{23,23}=-1,{24,24}=-1,{25,25}=-1,{26,26}=-1,{27,27}=-1,{28,28}=-1,
 {29,29}=1,{30,30}=1,{31,31}=1,{32,32}=1,{33,33}=1,{34,34}=1,{35,35}=1,
 {36,36}=1,{37,37}=1,{38,38}=1,{39,39}=1,{40,40}=1,{41,41}=1,{42,42}=1,

 {1,2}=v1, {1,3}=v2, {1,4}=v3, {1,5}=v4, {1,6}=v5, {1,7}=v6, {1,8}=v7,
 {1,9}=v8,{1,10}=v9,{1,11}=v10,{1,12}=v11,{1,13}=v12,{1,14}=v13,
 {2,3}=v14, {2,4}=v15, {2,5}=v16, {2,6}=v17, {2,7}=v18, {2,8}=v19,{2,9}=v20,
 {2,10}=v21,{2,11}=v22,{2,12}=v23,{2,13}=v24,{2,14}=v25,
 {3,4}=v26, {3,5}=v27, {3,6}=v28, {3,7}=v29, {3,8}=v30, {3,9}=v31,
 {3,10}=v32,{3,11}=v33,{3,12}=v34,{3,13}=v35,{3,14}=v36,
 {4,5}=v37, {4,6}=v38, {4,7}=v39, {4,8}=v40, {4,9}=v41,
 {4,10}=v42,{4,11}=v43,{4,12}=v44,{4,13}=v45,{4,14}=v46,
 {5,6}=v47, {5,7}=v48, {5,8}=v49, {5,9}=v50,
 {5,10}=v51,{5,11}=v52,{5,12}=v53,{5,13}=v54,{5,14}=v55,
 {6,7}=v56, {6,8}=v57, {6,9}=v58,
 {6,10}=v59,{6,11}=v60,{6,12}=v61,{6,13}=v62,{6,14}=v63,
 {7,8}=v64,{7,9}=v65,{7,10}=v66,{7,11}=v67,{7,12}=v68,{7,13}=v69,{7,14}=v70,
 {8,9}=v71,{8,10}=v72,{8,11}=v73,{8,12}=v74,{8,13}=v75,{8,14}=v76,
 {9,10}=v77,{9,11}=v78,{9,12}=v79,{9,13}=v80,{9,14}=v81,
 {10,11}=v82,{10,12}=v83,{10,13}=v84,{10,14}=v85,
 {11,12}=v86,{11,13}=v87,{11,14}=v88,{12,13}=v89,{12,14}=v90,{13,14}=v91,

{15,16}=v1,{15,17}=v2,{15,18}=v3,{15,19}=v4,{15,20}=v5,{15,21}=v6,{15,22}=v7,
{15,23}=v8,{15,24}=v9,{15,25}=v10,{15,26}=v11,{15,27}=v12,{15,28}=v13,
{16,17}=v14,{16,18}=v15,{16,19}=v16,{16,20}=v17,{16,21}=v18,{16,22}=v19,
{16,23}=v20,{16,24}=v21,{16,25}=v22,{16,26}=v23,{16,27}=v24,{16,28}=v25,
{17,18}=v26, {17,19}=v27, {17,20}=v28, {17,21}=v29, {17,22}=v30, {17,23}=v31,
{17,24}=v32,{17,25}=v33,{17,26}=v34,{17,27}=v35,{17,28}=v36,
{18,19}=v37, {18,20}=v38, {18,21}=v39, {18,22}=v40, {18,23}=v41,
{18,24}=v42,{18,25}=v43,{18,26}=v44,{18,27}=v45,{18,28}=v46,
{19,20}=v47, {19,21}=v48, {19,22}=v49, {19,23}=v50,
{19,24}=v51,{19,25}=v52,{19,26}=v53,{19,27}=v54,{19,28}=v55,
```

```
{20,21}=v56, {20,22}=v57, {20,23}=v58,
{20,24}=v59,{20,25}=v60,{20,26}=v61,{20,27}=v62,{20,28}=v63,
{21,22}=v64,{21,23}=v65,{21,24}=v66,{21,25}=v67,{21,26}=v68,{21,27}=v69,
{21,28}=v70,{22,23}=v71,{22,24}=v72,{22,25}=v73,{22,26}=v74,{22,27}=v75,
{22,28}=v76,{23,24}=v77,{23,25}=v78,{23,26}=v79,{23,27}=v80,{23,28}=v81,
{24,25}=v82,{24,26}=v83,{24,27}=v84,{24,28}=v85,
{25,26}=v86,{25,27}=v87,{25,28}=v88,{26,27}=v89,{26,28}=v90,{27,28}=v91,


 {2,1}=t1, {3,1}=t2, {4,1}=t3, {5,1}=t4, {6,1}=t5, {7,1}=t6, {8,1}=t7,
 {9,1}=t8,{10,1}=t9,{11,1}=t10,{12,1}=t11,{13,1}=t12,{14,1}=t13,
 {3,2}=t14, {4,2}=t15, {5,2}=t16, {6,2}=t17, {7,2}=t18, {8,2}=t19,{9,2}=t20,
 {10,2}=t21,{11,2}=t22,{12,2}=t23,{13,2}=t24,{14,2}=t25,
 {4,3}=t26, {5,3}=t27, {6,3}=t28, {7,3}=t29, {8,3}=t30, {9,3}=t31,
 {10,3}=t32,{11,3}=t33,{12,3}=t34,{13,3}=t35,{14,3}=t36,
 {5,4}=t37, {6,4}=t38, {7,4}=t39, {8,4}=t40, {9,4}=t41,
 {10,4}=t42,{11,4}=t43,{12,4}=t44,{13,4}=t45,{14,4}=t46,
 {6,5}=t47, {7,5}=t48, {8,5}=t49, {9,5}=t50,
 {10,5}=t51,{11,5}=t52,{12,5}=t53,{13,5}=t54,{14,5}=t55,
 {7,6}=t56, {8,6}=t57, {9,6}=t58,
 {10,6}=t59,{11,6}=t60,{12,6}=t61,{13,6}=t62,{14,6}=t63,
 {8,7}=t64,{9,7}=t65,{10,7}=t66,{11,7}=t67,{12,7}=t68,{13,7}=t69,{14,7}=t70,
 {9,8}=t71,{10,8}=t72,{11,8}=t73,{12,8}=t74,{13,8}=t75,{14,8}=t76,
 {10,9}=t77,{11,9}=t78,{12,9}=t79,{13,9}=t80,{14,9}=t81,
 {11,10}=t82,{12,10}=t83,{13,10}=t84,{14,10}=t85,
 {12,11}=t86,{13,11}=t87,{14,11}=t88,{13,12}=t89,{14,12}=t90,{14,13}=t91,

{16,15}=t1, {17,15}=t2, {18,15}=t3, {19,15}=t4, {20,15}=t5, {21,15}=t6,
{22,15}=t7,
{23,15}=t8,{24,15}=t9,{25,15}=t10,{26,15}=t11,{27,15}=t12,{28,15}=t13,
{17,16}=t14, {18,16}=t15, {19,16}=t16, {20,16}=t17, {21,16}=t18,
{22,16}=t19,{23,16}=t20,
{24,16}=t21,{25,16}=t22,{26,16}=t23,{27,16}=t24,{28,16}=t25,
{18,17}=t26, {19,17}=t27, {20,17}=t28, {21,17}=t29, {22,17}=t30, {23,17}=t31,
{24,17}=t32,{25,17}=t33,{26,17}=t34,{27,17}=t35,{28,17}=t36,
{19,18}=t37, {20,18}=t38, {21,18}=t39, {22,18}=t40, {23,18}=t41,
{24,18}=t42,{25,18}=t43,{26,18}=t44,{27,18}=t45,{28,18}=t46,
{20,19}=t47, {21,19}=t48, {22,19}=t49, {23,19}=t50,
{24,19}=t51,{25,19}=t52,{26,19}=t53,{27,19}=t54,{28,19}=t55,
{21,20}=t56, {22,20}=t57, {23,20}=t58,
{24,20}=t59,{25,20}=t60,{26,20}=t61,{27,20}=t62,{28,20}=t63,
{22,21}=t64,{23,21}=t65,{24,21}=t66,{25,21}=t67,{26,21}=t68,{27,21}=t69,
{28,21}=t70,{23,22}=t71,{24,22}=t72,{25,22}=t73,{26,22}=t74,{27,22}=t75,
{28,22}=t76,{24,23}=t77,{25,23}=t78,{26,23}=t79,{27,23}=t80,{28,23}=t81,
{25,24}=t82,{26,24}=t83,{27,24}=t84,{28,24}=t85,
{26,25}=t86,{27,25}=t87,{28,25}=t88,{27,26}=t89,{28,26}=t90,{28,27}=t91;


Matrix v
 {1,1}=-1, {2,2}=-1, {3,3}=-1, {4,4}=-1, {5,5}=-1, {6,6}=-1, {7,7}=-1,
 {8,8}=-1, {9,9}=-1, {10,10}=-1, {11,11}=-1,{12,12}=-1,{13,13}=-1,{14,14}=-1,
 {15,15}=-1,{16,16}=-1,{17,17}=-1,{18,18}=-1,{19,19}=-1,{20,20}=-1,{21,21}=-1,
 {22,22}=-1,{23,23}=-1,{24,24}=-1,{25,25}=-1,{26,26}=-1,{27,27}=-1,{28,28}=-1,
 {29,29}=1,{30,30}=1,{31,31}=1,{32,32}=1,{33,33}=1,{34,34}=1,{35,35}=1,
 {36,36}=1,{37,37}=1,{38,38}=1,{39,39}=1,{40,40}=1,{41,41}=1,{42,42}=1,

 {2,1}=v1, {3,1}=v2, {4,1}=v3, {5,1}=v4, {6,1}=v5, {7,1}=v6, {8,1}=v7,
 {9,1}=v8,{10,1}=v9,{11,1}=v10,{12,1}=v11,{13,1}=v12,{14,1}=v13,
 {3,2}=v14, {4,2}=v15, {5,2}=v16, {6,2}=v17, {7,2}=v18, {8,2}=v19,{9,2}=v20,
 {10,2}=v21,{11,2}=v22,{12,2}=v23,{13,2}=v24,{14,2}=v25,
```

```
 {4,3}=v26, {5,3}=v27, {6,3}=v28, {7,3}=v29, {8,3}=v30, {9,3}=v31,
 {10,3}=v32,{11,3}=v33,{12,3}=v34,{13,3}=v35,{14,3}=v36,
 {5,4}=v37, {6,4}=v38, {7,4}=v39, {8,4}=v40, {9,4}=v41,
 {10,4}=v42,{11,4}=v43,{12,4}=v44,{13,4}=v45,{14,4}=v46,
 {6,5}=v47, {7,5}=v48, {8,5}=v49, {9,5}=v50,
 {10,5}=v51,{11,5}=v52,{12,5}=v53,{13,5}=v54,{14,5}=v55,
 {7,6}=v56, {8,6}=v57, {9,6}=v58,
 {10,6}=v59,{11,6}=v60,{12,6}=v61,{13,6}=v62,{14,6}=v63,
 {8,7}=v64,{9,7}=v65,{10,7}=v66,{11,7}=v67,{12,7}=v68,{13,7}=v69,{14,7}=v70,
 {9,8}=v71,{10,8}=v72,{11,8}=v73,{12,8}=v74,{13,8}=v75,{14,8}=v76,
 {10,9}=v77,{11,9}=v78,{12,9}=v79,{13,9}=v80,{14,9}=v81,
 {11,10}=v82,{12,10}=v83,{13,10}=v84,{14,10}=v85,
 {12,11}=v86,{13,11}=v87,{14,11}=v88,{13,12}=v89,{14,12}=v90,{14,13}=v91,

{16,15}=v1, {17,15}=v2, {18,15}=v3, {19,15}=v4, {20,15}=v5, {21,15}=v6,
{22,15}=v7,
{23,15}=v8,{24,15}=v9,{25,15}=v10,{26,15}=v11,{27,15}=v12,{28,15}=v13,
{17,16}=v14, {18,16}=v15, {19,16}=v16, {20,16}=v17, {21,16}=v18,
{22,16}=v19,{23,16}=v20,
{24,16}=v21,{25,16}=v22,{26,16}=v23,{27,16}=v24,{28,16}=v25,
{18,17}=v26, {19,17}=v27, {20,17}=v28, {21,17}=v29, {22,17}=v30, {23,17}=v31,
{24,17}=v32,{25,17}=v33,{26,17}=v34,{27,17}=v35,{28,17}=v36,
{19,18}=v37, {20,18}=v38, {21,18}=v39, {22,18}=v40, {23,18}=v41,
{24,18}=v42,{25,18}=v43,{26,18}=v44,{27,18}=v45,{28,18}=v46,
{20,19}=v47, {21,19}=v48, {22,19}=v49, {23,19}=v50,
{24,19}=v51,{25,19}=v52,{26,19}=v53,{27,19}=v54,{28,19}=v55,
{21,20}=v56, {22,20}=v57, {23,20}=v58,
{24,20}=v59,{25,20}=v60,{26,20}=v61,{27,20}=v62,{28,20}=v63,
{22,21}=v64,{23,21}=v65,{24,21}=v66,{25,21}=v67,{26,21}=v68,{27,21}=v69,
{28,21}=v70,{23,22}=v71,{24,22}=v72,{25,22}=v73,{26,22}=v74,{27,22}=v75,
{28,22}=v76,{24,23}=v77,{25,23}=v78,{26,23}=v79,{27,23}=v80,{28,23}=v81,
{25,24}=v82,{26,24}=v83,{27,24}=v84,{28,24}=v85,
{26,25}=v86,{27,25}=v87,{28,25}=v88,{27,26}=v89,{28,26}=v90,{28,27}=v91,


{1,2}=t1, {1,3}=t2, {1,4}=t3, {1,5}=t4, {1,6}=t5, {1,7}=t6, {1,8}=t7,
 {1,9}=t8,{1,10}=t9,{1,11}=t10,{1,12}=t11,{1,13}=t12,{1,14}=t13,
 {2,3}=t14, {2,4}=t15, {2,5}=t16, {2,6}=t17, {2,7}=t18, {2,8}=t19,{2,9}=t20,
 {2,10}=t21,{2,11}=t22,{2,12}=t23,{2,13}=t24,{2,14}=t25,
 {3,4}=t26, {3,5}=t27, {3,6}=t28, {3,7}=t29, {3,8}=t30, {3,9}=t31,
 {3,10}=t32,{3,11}=t33,{3,12}=t34,{3,13}=t35,{3,14}=t36,
 {4,5}=t37, {4,6}=t38, {4,7}=t39, {4,8}=t40, {4,9}=t41,
 {4,10}=t42,{4,11}=t43,{4,12}=t44,{4,13}=t45,{4,14}=t46,
 {5,6}=t47, {5,7}=t48, {5,8}=t49, {5,9}=t50,
 {5,10}=t51,{5,11}=t52,{5,12}=t53,{5,13}=t54,{5,14}=t55,
 {6,7}=t56, {6,8}=t57, {6,9}=t58,
 {6,10}=t59,{6,11}=t60,{6,12}=t61,{6,13}=t62,{6,14}=t63,
 {7,8}=t64,{7,9}=t65,{7,10}=t66,{7,11}=t67,{7,12}=t68,{7,13}=t69,{7,14}=t70,
 {8,9}=t71,{8,10}=t72,{8,11}=t73,{8,12}=t74,{8,13}=t75,{8,14}=t76,
 {9,10}=t77,{9,11}=t78,{9,12}=t79,{9,13}=t80,{9,14}=t81,
 {10,11}=t82,{10,12}=t83,{10,13}=t84,{10,14}=t85,
 {11,12}=t86,{11,13}=t87,{11,14}=t88,{12,13}=t89,{12,14}=t90,{13,14}=t91,

{15,16}=t1,{15,17}=t2,{15,18}=t3,{15,19}=t4,{15,20}=t5,{15,21}=t6,{15,22}=t7,
{15,23}=t8,{15,24}=t9,{15,25}=t10,{15,26}=t11,{15,27}=t12,{15,28}=t13,
{16,17}=t14,{16,18}=t15,{16,19}=t16,{16,20}=t17,{16,21}=t18,{16,22}=t19,
{16,23}=t20,{16,24}=t21,{16,25}=t22,{16,26}=t23,{16,27}=t24,{16,28}=t25,
{17,18}=t26, {17,19}=t27, {17,20}=t28, {17,21}=t29, {17,22}=t30, {17,23}=t31,
{17,24}=t32,{17,25}=t33,{17,26}=t34,{17,27}=t35,{17,28}=t36,
{18,19}=t37, {18,20}=t38, {18,21}=t39, {18,22}=t40, {18,23}=t41,
```

```
{18,24}=t42,{18,25}=t43,{18,26}=t44,{18,27}=t45,{18,28}=t46,
{19,20}=t47, {19,21}=t48, {19,22}=t49, {19,23}=t50,
{19,24}=t51,{19,25}=t52,{19,26}=t53,{19,27}=t54,{19,28}=t55,
{20,21}=t56, {20,22}=t57, {20,23}=t58,
{20,24}=t59,{20,25}=t60,{20,26}=t61,{20,27}=t62,{20,28}=t63,
{21,22}=t64,{21,23}=t65,{21,24}=t66,{21,25}=t67,{21,26}=t68,{21,27}=t69,
{21,28}=t70,{22,23}=t71,{22,24}=t72,{22,25}=t73,{22,26}=t74,{22,27}=t75,
{22,28}=t76,{23,24}=t77,{23,25}=t78,{23,26}=t79,{23,27}=t80,{23,28}=t81,
{24,25}=t82,{24,26}=t83,{24,27}=t84,{24,28}=t85,
{25,26}=t86,{25,27}=t87,{25,28}=t88,{26,27}=t89,{26,28}=t90,{27,28}=t91;

matrix p1
{1,1}=1, {2,2}=1, {3,3}=1, {4,4}=1, {5,5}=1, {6,6}=1, {7,7}=1,{8,8}=1,{9,9}=1,
{10,10}=1,{11,11}=1,{12,12}=1,{13,13}=1,{14,14}=1,

{1,15}=1, {2,16}=1, {3,17}=1, {4,18}=1, {5,19}=1, {6,20}=1, {7,21}=1,{8,22}=1,
{9,23}=1,{10,24}=1,{11,25}=1,{12,26}=1,{13,27}=1,{14,28}=1,

{1,29}=1, {2,30}=1, {3,31}=1, {4,32}=1, {5,33}=1, {6,34}=1, {7,35}=1,{8,36}=1,
{9,37}=1,{10,38}=1,{11,39}=1,{12,40}=1,{13,41}=1,{14,42}=1,


{15,1}=1, {16,2}=1, {17,3}=1, {18,4}=1, {19,5}=1, {20,6}=1, {21,7}=1,{22,8}=1,
{23,9}=1,{24,10}=1,{25,11}=1,{26,12}=1,{27,13}=1,{28,14}=1,

{15,15}=1, {16,16}=1, {17,17}=1, {18,18}=1, {19,19}=1, {20,20}=1, {21,21}=1,
{22,22}=1,{23,23}=1,{24,24}=1,{25,25}=1,{26,26}=1,{27,27}=1,{28,28}=1,

{15,29}=1, {16,30}=1, {17,31}=1, {18,32}=1, {19,33}=1, {20,34}=1, {21,35}=1,
{22,36}=1,{23,37}=1,{24,38}=1,{25,39}=1,{26,40}=1,{27,41}=1,{28,42}=1,

{29,1}=1, {30,2}=1, {31,3}=1, {32,4}=1, {33,5}=1, {34,6}=1, {35,7}=1,{36,8}=1,
{37,9}=1,{38,10}=1,{39,11}=1,{40,12}=1,{41,13}=1,{42,14}=1,

{29,15}=1, {30,16}=1, {31,17}=1, {32,18}=1, {33,19}=1, {34,20}=1, {35,21}=1,
{36,22}=1,{37,23}=1,{38,24}=1,{39,25}=1,{40,26}=1,{41,27}=1,{42,28}=1,

{29,29}=1, {30,30}=1, {31,31}=1, {32,32}=1, {33,33}=1, {34,34}=1, {35,35}=1,
{36,36}=1,{37,37}=1,{38,38}=1,{39,39}=1,{40,40}=1,{41,41}=1,{42,42}=1;

matrix r
{1,1}=d1d1,{2,2}=d1d2,{3,3}=d1d3,{4,4}=d1d4,{5,5}=d1d5,{6,6}=d1d6,{7,7}=d1d7,
{8,8}=d1d8,{9,9}=d1d9,{10,10}=d1d10,

{15,15}=d2d1,{16,16}=d2d2,{17,17}=d2d3,{18,18}=d2d4,{19,19}=d2d5,
{20,20}=d2d6,{21,21}=d2d7,{22,22}=d2d8,{23,23}=d2d9,{24,24}=d2d10;

matrix ta
 {1,1}=1, {2,2}=1, {3,3}=1, {4,4}=1, {5,5}=1, {6,6}=1, {7,7}=1,
 {8,8}=1, {9,9}=1, {10,10}=1, {11,11}=1,{12,12}=1,{13,13}=1,{14,14}=1,
 {15,15}=1,{16,16}=1,{17,17}=1,{18,18}=1,{19,19}=1,{20,20}=1,{21,21}=1,
 {22,22}=1,{23,23}=1,{24,24}=1,{25,25}=1,{26,26}=1,{27,27}=1,{28,28}=1,

{29,30}=v1,{29,31}=v2,{29,32}=v3,{29,33}=v4,{29,34}=v5,{29,35}=v6,{29,36}=v7,
{29,37}=v8,{29,38}=v9,{29,39}=v10,{29,40}=v11,{29,41}=v12,{29,42}=v13,
{30,31}=v14,{30,32}=v15,{30,33}=v16,{30,34}=v17,{30,35}=v18,{30,36}=v19,
{30,37}=v20,{30,38}=v21,{30,39}=v22,{30,40}=v23,{30,41}=v24,{30,42}=v25,
{31,32}=v26, {31,33}=v27, {31,34}=v28, {31,35}=v29, {31,36}=v30, {31,37}=v31,
{31,38}=v32,{31,39}=v33,{31,40}=v34,{31,41}=v35,{31,42}=v36,
{32,33}=v37, {32,34}=v38, {32,35}=v39, {32,36}=v40, {32,37}=v41,
```

```
{32,38}=v42,{32,39}=v43,{32,40}=v44,{32,41}=v45,{32,42}=v46,
{33,34}=v47, {33,35}=v48, {33,36}=v49, {33,37}=v50,
{33,38}=v51,{33,39}=v52,{33,40}=v53,{33,41}=v54,{33,42}=v55,
{34,35}=v56, {34,36}=v57, {34,37}=v58,
{34,38}=v59,{34,39}=v60,{34,40}=v61,{34,41}=v62,{34,42}=v63,
{35,36}=v64,{35,37}=v65,{35,38}=v66,{35,39}=v67,{35,40}=v68,{35,41}=v69,
{35,42}=v70,{36,37}=v71,{36,38}=v72,{36,39}=v73,{36,40}=v74,{36,41}=v75,
{36,42}=v76,{37,38}=v77,{37,39}=v78,{37,40}=v79,{37,41}=v80,{37,42}=v81,
{38,39}=v82,{38,40}=v83,{38,41}=v84,{38,42}=v85,
{39,40}=v86,{39,41}=v87,{39,42}=v88,{40,41}=v89,{40,42}=v90,{41,42}=v91,

{30,29}=t1, {31,29}=t2, {32,29}=t3, {33,29}=t4, {34,29}=t5, {35,29}=t6,
{36,29}=t7,
{37,29}=t8,{38,29}=t9,{39,29}=t10,{40,29}=t11,{41,29}=t12,{42,29}=t13,
{31,30}=t14, {32,30}=t15, {33,30}=t16, {34,30}=t17, {35,30}=t18,
{36,30}=t19,{37,30}=t20,
{38,30}=t21,{39,30}=t22,{40,30}=t23,{41,30}=t24,{42,30}=t25,
{32,31}=t26, {33,31}=t27, {34,31}=t28, {35,31}=t29, {36,31}=t30, {37,31}=t31,
{38,31}=t32,{39,31}=t33,{40,31}=t34,{41,31}=t35,{42,31}=t36,
{33,32}=t37, {34,32}=t38, {35,32}=t39, {36,32}=t40, {37,32}=t41,
{38,32}=t42,{39,32}=t43,{40,32}=t44,{41,32}=t45,{42,32}=t46,
{34,33}=t47, {35,33}=t48, {36,33}=t49, {37,33}=t50,
{38,33}=t51,{39,33}=t52,{40,33}=t53,{41,33}=t54,{42,33}=t55,
{35,34}=t56, {36,34}=t57, {37,34}=t58,
{38,34}=t59,{39,34}=t60,{40,34}=t61,{41,34}=t62,{42,34}=t63,
{36,35}=t64,{37,35}=t65,{38,35}=t66,{39,35}=t67,{40,35}=t68,{41,35}=t69,
{42,35}=t70,{37,36}=t71,{38,36}=t72,{39,36}=t73,{40,36}=t74,{41,36}=t75,
{42,36}=t76,{38,37}=t77,{39,37}=t78,{40,37}=t79,{41,37}=t80,{42,37}=t81,
{39,38}=t82,{40,38}=t83,{41,38}=t84,{42,38}=t85,
{40,39}=t86,{41,39}=t87,{42,39}=t88,{41,40}=t89,{42,40}=t90,{42,41}=t91;

matrix va
 {1,1}=1, {2,2}=1, {3,3}=1, {4,4}=1, {5,5}=1, {6,6}=1, {7,7}=1,
 {8,8}=1, {9,9}=1, {10,10}=1, {11,11}=1,{12,12}=1,{13,13}=1,{14,14}=1,
 {15,15}=1,{16,16}=1,{17,17}=1,{18,18}=1,{19,19}=1,{20,20}=1,{21,21}=1,
 {22,22}=1,{23,23}=1,{24,24}=1,{25,25}=1,{26,26}=1,{27,27}=1,{28,28}=1,

{30,29}=v1, {31,29}=v2, {32,29}=v3, {33,29}=v4, {34,29}=v5, {35,29}=v6,
{36,29}=v7,
{37,29}=v8,{38,29}=v9,{39,29}=v10,{40,29}=v11,{41,29}=v12,{42,29}=v13,
{31,30}=v14, {32,30}=v15, {33,30}=v16, {34,30}=v17, {35,30}=v18,
{36,30}=v19,{37,30}=v20,
{38,30}=v21,{39,30}=v22,{40,30}=v23,{41,30}=v24,{42,30}=v25,
{32,31}=v26, {33,31}=v27, {34,31}=v28, {35,31}=v29, {36,31}=v30, {37,31}=v31,
{38,31}=v32,{39,31}=v33,{40,31}=v34,{41,31}=v35,{42,31}=v36,
{33,32}=v37, {34,32}=v38, {35,32}=v39, {36,32}=v40, {37,32}=v41,
{38,32}=v42,{39,32}=v43,{40,32}=v44,{41,32}=v45,{42,32}=v46,
{34,33}=v47, {35,33}=v48, {36,33}=v49, {37,33}=v50,
{38,33}=v51,{39,33}=v52,{40,33}=v53,{41,33}=v54,{42,33}=v55,
{35,34}=v56, {36,34}=v57, {37,34}=v58,
{38,34}=v59,{39,34}=v60,{40,34}=v61,{41,34}=v62,{42,34}=v63,
{36,35}=v64,{37,35}=v65,{38,35}=v66,{39,35}=v67,{40,35}=v68,{41,35}=v69,
{42,35}=v70,{37,36}=v71,{38,36}=v72,{39,36}=v73,{40,36}=v74,{41,36}=v75,
{42,36}=v76,{38,37}=v77,{39,37}=v78,{40,37}=v79,{41,37}=v80,{42,37}=v81,
{39,38}=v82,{40,38}=v83,{41,38}=v84,{42,38}=v85,
{40,39}=v86,{41,39}=v87,{42,39}=v88,{41,40}=v89,{42,40}=v90,{42,41}=v91,

{29,30}=t1,{29,31}=t2,{29,32}=t3,{29,33}=t4,{29,34}=t5,{29,35}=t6,{29,36}=t7,
{29,37}=t8,{29,38}=t9,{29,39}=t10,{29,40}=t11,{29,41}=t12,{29,42}=t13,
{30,31}=t14,{30,32}=t15,{30,33}=t16,{30,34}=t17,{30,35}=t18,{30,36}=t19,
```

```
{30,37}=t20,{30,38}=t21,{30,39}=t22,{30,40}=t23,{30,41}=t24,{30,42}=t25,
{31,32}=t26, {31,33}=t27, {31,34}=t28, {31,35}=t29, {31,36}=t30, {31,37}=t31,
{31,38}=t32,{31,39}=t33,{31,40}=t34,{31,41}=t35,{31,42}=t36,
{32,33}=t37, {32,34}=t38, {32,35}=t39, {32,36}=t40, {32,37}=t41,
{32,38}=t42,{32,39}=t43,{32,40}=t44,{32,41}=t45,{32,42}=t46,
{33,34}=t47, {33,35}=t48, {33,36}=t49, {33,37}=t50,
{33,38}=t51,{33,39}=t52,{33,40}=t53,{33,41}=t54,{33,42}=t55,
{34,35}=t56, {34,36}=t57, {34,37}=t58,
{34,38}=t59,{34,39}=t60,{34,40}=t61,{34,41}=t62,{34,42}=t63,
{35,36}=t64,{35,37}=t65,{35,38}=t66,{35,39}=t67,{35,40}=t68,{35,41}=t69,
{35,42}=t70,{36,37}=t71,{36,38}=t72,{36,39}=t73,{36,40}=t74,{36,41}=t75,
{36,42}=t76,{37,38}=t77,{37,39}=t78,{37,40}=t79,{37,41}=t80,{37,42}=t81,
{38,39}=t82,{38,40}=t83,{38,41}=t84,{38,42}=t85,
{39,40}=t86,{39,41}=t87,{39,42}=t88,{40,41}=t89,{40,42}=t90,{41,42}=t91;



matrix s1a
{42,42}=1;
matrix s1b
{41,41}=1;
matrix s2a
{42,42}=1, {41,41}=1;
matrix s2b
{40,40}=1;
matrix s3a
{42,42}=1,{41,41}=1,{40,40}=1;
matrix s3b
{39,39}=1;
matrix s4a
{42,42}=1,{41,41}=1,{40,40}=1,{39,39}=1;
matrix s4b
{38,38}=1;
matrix s5a
{42,42}=1,{41,41}=1,{40,40}=1,{39,39}=1,{38,38}=1;
matrix s5b
{37,37}=1;
matrix s6a
{42,42}=1,{41,41}=1,{40,40}=1,{39,39}=1,{38,38}=1,{37,37}=1;
matrix s6b
{36,36}=1;
matrix s7a
{42,42}=1,{41,41}=1,{40,40}=1,{39,39}=1,{38,38}=1,{37,37}=1,{36,36}=1;
matrix s7b
{35,35}=1;
matrix s8a
{42,42}=1,{41,41}=1,{40,40}=1,{39,39}=1,{38,38}=1,{37,37}=1,{36,36}=1,
{35,35}=1;
matrix s8b
{34,34}=1;
matrix s9a
{42,42}=1,{41,41}=1,{40,40}=1,{39,39}=1,{38,38}=1,{37,37}=1,{36,36}=1,
{35,35}=1,{34,34}=1;
matrix s9b
{33,33}=1;
matrix s10a
{42,42}=1,{41,41}=1,{40,40}=1,{39,39}=1,{38,38}=1,{37,37}=1,{36,36}=1,
{35,35}=1,{34,34}=1,{33,33}=1;
matrix s10b
{32,32}=1;
```

```
matrix s11a
{42,42}=1,{41,41}=1,{40,40}=1,{39,39}=1,{38,38}=1,{37,37}=1,{36,36}=1,
{35,35}=1,{34,34}=1,{33,33}=1,{32,32}=1;
matrix s11b
{31,31}=1;
matrix s12a
{42,42}=1,{41,41}=1,{40,40}=1,{39,39}=1,{38,38}=1,{37,37}=1,{36,36}=1,
{35,35}=1,{34,34}=1,{33,33}=1,{32,32}=1,{31,31}=1;
matrix s12b
{30,30}=1;


*What follows are SAS programming statements that constrain the
 parameters.  t1=-v1 constrains t1 to be the additive inverse of v1.;

t1=-v1;t2=-v2;t3=-v3;t4=-v4;t5=-v5;t6=-v6;t7=-v7;t8=-v8;t9=-v9;t10=-v10;
t11=-v11;t12=-v12;t13=-v13;t14=-v14;t15=-v15;t16=-v16;t17=-v17;t18=-v18;
t19=-v19;t20=-v20;t21=-v21;t22=-v22;t23=-v23;t24=-v24;t25=-v25;t26=-v26;
t27=-v27;t28=-v28;t29=-v29;t30=-v30;t31=-v31;t32=-v32;t33=-v33;t34=-v34;
t35=-v35;t36=-v36;t37=-v37;t38=-v38;t39=-v39;t40=-v40;t41=-v41;t42=-v42;
t43=-v43;t44=-v44;t45=-v45;t46=-v46;t47=-v47;t48=-v48;t49=-v49;t50=-v50;
t51=-v51;t52=-v52;t53=-v53;t54=-v54;t55=-v55;t56=-v56;t57=-v57;t58=-v58;
t59=-v59;t60=-v60;t61=-v60;t62=-v62;t63=-v63;t64=-v64;t65=-v65;t66=-v66;
t67=-v67;t68=-v68;t69=-v69;t70=-v70;t71=-v71;t72=-v72;t73=-v73;t74=-v74;
t75=-v75;t76=-v76;t77=-v77;t78=-v78;t79=-v79;t80=-v80;t81=-v81;t82=-v82;
t83=-v83;t84=-v84;t85=-v85;t86=-v86;t87=-v87;t88=-v88;t89=-v89;t90=-v90;
t91=-v91;


d1d2=0;d1d3=0;d1d4=0;d1d5=0;d1d6=0;d1d7=0;d1d8=0;d1d9=0;d1d10=0;
d2d2=0;d2d3=0;d2d4=0;d2d5=0;d2d6=0;d2d7=0;d2d8=0;d2d9=0;d2d10=0;

run;



*The subsequent code calculates V.;

 proc iml; use cal; read point 4
 var {v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v14 v15
 v16 v17 v18 v19 v20 v21 v22 v23 v24 v25 v26 v27 v28 v29 v30 v31 v32 v33 v34
 v35 v36 v37 v38 v39 v40
 v41 v42 v43 v44 v45 v46 v47 v48 v49 v50 v51 v52 v53 v54 v55 v56 v57 v58 v59
 v60 v61 v62 v63 v64 v65
 v66 v67 v68 v69 v70 v71 v72 v73 v74 v75 v76 v77 v78 v79 v80 v81 v82 v83 v84
 v85 v86 v87 v88 v89 v90 v91} into K;




 v=J(14,14,0);
 v[1,2]=K[,1]; v[1,3]=K[,2]; v[1,4]=K[,3];    v[1,5]=K[,4];  v[1,6]=K[,5];
 v[1,7]=K[6]; v[1,8]=K[7];
 v[1,9]=K[,8]; v[1,10]=K[,9];v[1,11]=K[,10];  v[1,12]=K[,11];v[1,13]=K[,12];
 v[1,14]=K[,13];
 v[2,3]=K[,14]; v[2,4]=K[,15]; v[2,5]=K[,16]; v[2,6]=K[,17]; v[2,7]=K[,18];
 v[2,8]=K[,19];v[2,9]=K[,20];
 v[2,10]=K[,21];v[2,11]=K[,22];v[2,12]=K[,23];v[2,13]=K[,24];v[2,14]=K[,25];
 v[3,4]=K[,26]; v[3,5]=K[,27]; v[3,6]=K[,28]; v[3,7]=K[,29]; v[3,8]=K[,30];
```

```
v[3,9]=K[,31];
v[3,10]=K[,32];v[3,11]=K[,33];v[3,12]=K[,34];v[3,13]=K[,35];v[3,14]=K[,36];
v[4,5]=K[,37]; v[4,6]=K[,38]; v[4,7]=K[,39]; v[4,8]=K[,40]; v[4,9]=K[,41];
v[4,10]=K[,42];v[4,11]=K[,43];v[4,12]=K[,44];v[4,13]=K[,45];v[4,14]=K[,46];
v[5,6]=K[,47]; v[5,7]=K[,48]; v[5,8]=K[,49]; v[5,9]=K[,50];
v[5,10]=K[,51];v[5,11]=K[,52];v[5,12]=K[,53];v[5,13]=K[,54];v[5,14]=K[,55];
v[6,7]=K[,56]; v[6,8]=K[,57]; v[6,9]=K[,58];
v[6,10]=K[,59];v[6,11]=K[,60];v[6,12]=K[,61];v[6,13]=K[,62];v[6,14]=K[,63];
v[7,8]=K[,64]; v[7,9]=K[,65]; v[7,10]=K[,66];v[7,11]=K[,67];v[7,12]=K[,68];
v[7,13]=K[,69];v[7,14]=K[,70];
v[8,9]=K[,71]; v[8,10]=K[,72];v[8,11]=K[,73];v[8,12]=K[,74];v[8,13]=K[,75];
v[8,14]=K[,76];
v[9,10]=K[,77];v[9,11]=K[,78];v[9,12]=K[,79];v[9,13]=K[,80];v[9,14]=K[,81];
v[10,11]=K[,82];v[10,12]=K[,83];v[10,13]=K[,84];v[10,14]=K[,85];
v[11,12]=K[,86];v[11,13]=K[,87];v[11,14]=K[,88];v[12,13]=K[,89];
v[12,14]=K[,90];v[13,14]=K[,91];
Vo=v-v`-I(14); To=-v+v`-I(14);
Orth=inv(Vo)*To; print orth;
```

*The matrix of canonical variates, **V**, is called “orth”.;

APPENDIX SIX

SAS CODE FOR THE SIMULATION OF THE COMMON VARIATES MODEL

```
*Programming code for simulating 5000 datasets of size 4000 and then
 estimating the CVA/time model.  P=3 is the number of variables, t=3 is the
 number of occasions, and m=4 is the number of groups.  tn=4000 is the total
 sample size and gn=1000 is the sample size in each group.;

%let P=3; %let t=3; %let m=4;
%let count=5000;  %let tn=4000;   %let gn=1000;


*The code generates a dataset, then estimates both the common variate model
 and the unique variate model.  The global do loop invokes several macros
 repeatedly.  Hence the macros are discussed first.;


*fititu and fititc are macros that calculate the fit of the estimates for
 the unique variates and the common variates models, respectively.  These
 fits are used in the calculation of the maximum likelihood test statistic.;

%macro fititu;
fit=j(9,9,0);v1=b[1:3,];v2=b[4:6,];v3=b[7:9];
  %do g=1 %to &m;
    %do s=1 %to &t;
      res&g&s=x&g&s-e[&g,&s]*v&s;
    %end;
    res&g=res&g.1//res&g.2//res&g.3;
    res&g.res=res&g*res&g`;
    fit=&gn*res&g.res+fit;
  %end;
%mend;

%macro fititc;
fit=j(9,9,0);     v=b[1:3,];
  %do g=1 %to &m;
    %do s=1 %to &t;
      res&g&s=x&g&s-e[&g,&s]*v;
    %end;
```

```
    res&g=res&g.1//res&g.2//res&g.3;
    res&g.res=res&g*res&g`;
    fit=&gn*res&g.res+fit;
  %end;
%mend;


*The macros makmat and xbar set up the data matrices that are used in the
 algorithm. Makmat inverts S (the sample covariance) and then breaks it
 into smaller submatrices.  xbar sets up the group means.;

%macro makmat;
  Si=inv(S);
  %do i=1 %to &t;
    %do j=1 %to &t;
      a=&i*&t-(&t-1); k=&i*&t;   g=&j*&t-(&t-1); d=&j*&t;
      Si&i&j=Si[a:k,g:d];
    %end;
  %end;
%mend;

%macro xbar;
%do g=1 %to 4;
  %do s=1 %to 3;
     j1=&s*3-2; j2=&s*3;
     x&g&s= fava[&g,j1:j2]`;
  %end;
%end;
%mend;



*The macros normalF and normalu evaluate the normal equations for the
 common variates and the unique variates models respectively.;

%macro normalF;  v=B[1:&p,];  lm=B[L,1];  e=j(&m,&t,0);
  %do g=1 %to &m;
    j1=&P+(&g-1)*&t+1;  j2=&P+&g*&t;
    e[&g,]=B[j1:j2,]`;
  %end;
  F&w=j(L,1,0); wr=j(&P,1,0);
  %do g=1 %to &m;
    %do q=1 %to &t;
      %do s=1 %to &t;
      wn=e[&g,&q]*e[&g,&s]*Si&q&s*v - e[&g,&q]*Si&q&s*x&g&s +lm*v;
      wr=wr+wn;
      %end;
    %end;
  %end;
  D=j(&m,&t,0);
  %do g=1 %to &m;
    %do q=1 %to &t;
       %do s=1 %to &t;
         D[&g,&q]=D[&g,&q]+(0.5)*e[&g,&s]*v`*Si&q&s*v+
                  (0.5)*e[&g,&s]*v`*Si&s&q*v - v`*Si&q&s*x&g&s;
       %end;
    %end;
  %end;
  norm=ssq(v)-1;
  F&w=wr//D[1,]`//D[2,]`//D[3,]`//D[4,]`//norm;
```

```
%mend;

%macro normalu; nv=&p*&t; v=B[1:nv,];  e=j(&m,&t,0);
  %do r=1 %to &t;
    j1=&p*&r-&p+1;      j2=&p*&r;
    v&r=v[j1:j2,];      j3=&p*&t + &t*&m + &r;
    lm&r=B[j3,1];
  %end;
  %do g=1 %to &m;
    j1=(&p-1)*&t + &g*&t + 1;  j2=(&p-1)*&t + &g*&t + &t;
    e[&g,]=B[j1:j2,]`;
  %end;
  F&w=j(L,1,0);
  %do r=1 %to &t;      w&r=j(&P,1,0);
    %do g=1 %to &m;
      %do s=1 %to &t;
      wn=-e[&g,&r]*e[&g,&s]*Si&r&s*v&s + e[&g,&r]*Si&r&s*x&g&s +
        e[&g,&r]*e[&g,&s]*v&r`*Si&r&s*v&s*v&r-e[&g,&r]*v&r`*Si&r&s*x&g&s*v&r;
      w&r=w&r+wn;
      %end;
    %end;
    w&r=w&r+lm&r*v&r;
  %end;
  D=j(&m,&t,0);
  %do g=1 %to &m;
    %do r=1 %to &t;
       %do s=1 %to &P;
         D[&g,&r]=D[&g,&r] + e[&g,&s]*v&r`*Si&r&s*v&s - v&r`*Si&r&s*x&g&s;
       %end;
    %end;
  %end;
  norm=j(&t,1,0);
  %do s=1 %to &t;
    norm[&s,1]=ssq(v&s)-1;
  %end;
  F&w=w1//w2//w3//D[1,]`//D[2,]`//D[3,]`//D[4,]`//norm;
%mend;




proc iml;
BigBc=j(&count,15,0);BigBu=j(&count,23,0);
devmatc=j(&count,1,0);  devmatu=j(&count,1,0);   llcm=j(&count,5,0);
llum=j(&count,5,0); testy=j(&count,2,0);       llc=j(1,5,0); llu=j(1,5,0);
test=j(1,2,0);

do ttt=1 to &count;



*First a data set is created.  W is the covariance matrix of the data, v is
 the common canonical variate, and e is the matrix of group scores.  seed is
 is a 4000 by 9 matrix of zeros which serve as seeds for the random number
 generator.  The random numbers follow the normal distribution.  The matrix
 of random numbers, data, is treated as a matrix of residuals.;

  seed=j(&tn,9,0);  data=normal(seed);

W={  4.8   2.1   1.0   2.4  1.05 0.5  1.2   0.525 0.25,
```

```
     2.1   3.3   1.4   1.05 1.65 0.7  0.525 0.825 0.35,
     1.0   1.4   2.9   0.5  0.7  1.45 0.25  0.35  0.725,
     2.4   1.05  0.5   4.8  2.1  1.0  2.4   1.05  0.5,
     1.05  1.65  0.7   2.1  3.3  1.4  1.05  1.65  0.7,
     0.5   0.7   1.45  1.0  1.4  2.9  0.5   0.7   1.45,
     1.2   0.525 0.25  2.4  1.05 0.5  4.8   2.1   1.0,
     0.525 0.825 0.35  1.05 1.65 0.7  2.1   3.3   1.4,
     0.25  0.35  0.725 0.5  0.7  1.45 1.0   1.4   2.9};

  call svd(a,b,c,W);
  tdata=data*(diag(b)##0.5)*c`;
rhalf=(0.5)##(0.5);
v={0.5, 0.5, 0};    v[3,1]=rhalf;

                     e={1.0 0.0 1.0,
                        0.5 1.0 -.5,
                        -.5 0.0 0.5,
                        -1  -1  -1 };


*The subsequent statements center the data and create a matrix of group
 means and a within-groups error matrix.;

u1=e[1,]`@v;
u2=e[2,]`@v;
u3=e[3,]`@v;
u4=e[4,]`@v;
    u=u1||u2||u3||u4;
d1=j(&gn,1,1)@u1`;
d2=j(&gn,1,1)@u2`;
d3=j(&gn,1,1)@u3`;
d4=j(&gn,1,1)@u4`;
d=d1//d2//d3//d4;   bt=d`*d;
f=d+tdata;

ftot=j(1,9,0);
do i=1 to &tn;
  ftot=ftot+f[i,];
end;
fmean=ftot/&tn;
m=j(&tn,1,1);
 fmeans=m@fmean;
fad=f-fmeans;
sstot=fad`*fad;

f1tot=j(1,9,0);  f2tot=j(1,9,0);  f3tot=j(1,9,0);  f4tot=j(1,9,0);
f1=f[1:1000,]; f2=f[1001:2000,]; f3=f[2001:3000,]; f4=f[3001:&tn,];
do i=1 to &gn;
  f1tot=f1tot+f1[i,];
  f2tot=f2tot+f2[i,];
  f3tot=f3tot+f3[i,];
  f4tot=f4tot+f4[i,];
end;
f1ave=f1tot/&gn;
f2ave=f2tot/&gn;
f3ave=f3tot/&gn;
f4ave=f4tot/&gn;
ftot=(f1ave+f2ave+f3ave+f4ave)/4;
f1ava=f1ave-ftot; f2ava=f2ave-ftot; f3ava=f3ave-ftot; f4ava=f4ave-ftot;
fava=f1ava//f2ava//f3ava//f4ava;      *print fava;
```

```
bhat=(fava`*fava)*&gn;

f1ad=j(&gn,9,0); f2ad=j(&gn,9,0); f3ad=j(&gn,9,0); f4ad=j(&gn,9,0);
do i=1 to &gn;
  f1ad[i,]=f1[i,]-f1ave;
  f2ad[i,]=f2[i,]-f2ave;
  f3ad[i,]=f3[i,]-f3ave;
  f4ad[i,]=f4[i,]-f4ave;
end;
ss1=f1ad`*f1ad; ss2=f2ad`*f2ad; ss3=f3ad`*f3ad; ss4=f4ad`*f4ad;
sst=ss1 + ss2 + ss3 + ss4;
sstotal=bhat + sst;




%xbar

within=sst/&tn; S=within;

%let P=3; %let t=3; %let m=4;  L=&P+&t*&m+1;

%makmat


*The following set of lines is the code that finds the maximum likelihood
 estimate for the common variate model by solving the normal equations.
 The algorithm is a Gauss-Newton.;

  delta=0.000001; epsilon=1;   a=0;   r=10;   h=-2; con=0.01;   dev=0;
  B={.5,0.5,.707,1.0,0.0,1.0,0.5,1.0,-.5,-.5, 0.0, 0.5, -1, -1, -1,0 };
  bold=B;
  do until(alpha<0.000001);
    a=a+1;
    Hessian=j(L,L,0);
    rold=r;
    %let w=1;
    %NormalF
    do co=1 to 16;
      B=bold;     devold=dev;
      B[co,1]=bold[co,1]+delta;
      %let w=2;
      %NormalF
      Hessian[,co]=(F2-F1)/delta;
    end;
    B=bold-epsilon*inv(Hessian+I(16)*con)*F1;
    %let w=3;
    %NormalF
    bdiff=b-bold;  dev=sum(abs(f3));
    *print  a h epsilon f3 f1 dev B;
    sumabsd=sum(abs(f3)-abs(f1));
    ralpha=0;
    if sumabsd<0 then do; epsilon=1;  bold=b; h=-2; con=0.01; devold=dev;end;
    else do;
      b=bold;         dev=devold;
      if epsilon=1 then epsilon=0.1;
      else do;
        h=h+1;
```

```
        con=r**h; epsilon=1;
      end;
    end;
    alpha=abs(sum(f3));            if h=3 then alpha=0;
  end;


*The following lines of code generate the maximum likelihood test
 statistics.;

%fititc
n={&tn};
Sh=within + fit*(1/&tn);  iw=inv(within);
invSh=inv(Sh);

lrscapx=trace(iw*fit);
lrs=n*log(det(Sh))+n*trace(invSh*within)+trace(invSh*fit);
lrs1=n*log(det(Sh)); lrs2=trace(invSh*within); lrs3=trace(invSh*fit);
llc[,1]=lrs1;llc[,2]=lrs2;llc[,3]=lrs3;llc[,4]=lrs;llc[,5]=lrscapx;

BigBc[ttt,]=B[1:15,]`;   devmatc[ttt,]=dev;

Lu=&t*(&P+&m+1);



*The following set of lines is the code that finds the maximum likelihood
 estimate for the unique variate model by solving the normal equations.;

  delta=0.000001; epsilon=1;   a=0;   r=10;   h=-2; con=0.01;   dev=0;
  balt=b; b=j(24,1,0);
  B[1:3,]=Balt[1:3,]; B[4:6,]=Balt[1:3,]; B[7:9,]=Balt[1:3,];
  B[10:21,]=Balt[4:15,]; B[22:24,]={0, 0, 0};
  bold=B;
  do until(alpha<0.000001);
    a=a+1;
    Hessian=j(Lu,Lu,0);
    rold=r;
    %let w=1;
    %Normalu
    do co=1 to 24;
      B=bold;     devold=dev;
      B[co,1]=bold[co,1]+delta;
      %let w=2;
      %Normalu
      Hessian[,co]=(F2-F1)/delta;
    end;
    B=bold-epsilon*inv(Hessian+I(24)*con)*F1;
    %let w=3;
    %Normalu
    bdiff=b-bold;  dev=sum(abs(f3));
    sumabsd=sum(abs(f3)-abs(f1));
    ralpha=0;
    if sumabsd<0 then do; epsilon=1;  bold=b; h=-2; con=0.01; devold=dev;end;
    else do;
      b=bold;         dev=devold;
      if epsilon=1 then epsilon=0.1;
      else do;
        h=h+1;
        con=r**h; epsilon=1;
```

```
      end;
    end;
    alpha=abs(sum(f3));            if h=3 then alpha=0;
  end;


*The following lines of code generate the maximum likelihood test
 statistics.;

%fititu
n={&tn};
Sh=within + fit*(1/&tn);  iw=inv(within);
invSh=inv(Sh);

lrsuapx=trace(iw*fit);
lrus=n*log(det(Sh))+n*trace(invSh*within)+trace(invSh*fit);
lrsu1=n*log(det(Sh)); lrsu2=trace(invSh*within); lrsu3=trace(invSh*fit);
llu[,1]=lrsu1;llu[,2]=lrsu2;llu[,3]=lrsu3;llu[,4]=lrus;llu[,5]=lrsuapx;
test[,1]=lrs-lrus;
test[,2]=lrscapx-lrsuapx;

BigBu[ttt,]=B[1:23,]`; devmatu[ttt,]=dev;
   llum[ttt,1:5]=llu; llcm[ttt,1:5]=llc;  testy[ttt,1:2]=test;
end;

libname wat 'c:\prg'; reset storage=wat.simk5k;
store bigbc bigbu llum llcm testy devmatc devmatu;
quit iml;
```

APPENDIX SEVEN

MATHEMATICA CODE FOR CALCULATING THE ASYMPTOTIC COVARIANCE MATRIX OF THE ESTIMATES

*pm is a matrix of parameters.

```
bb=Table[0,{11},{11}];
pm={{v1},{v2},{e[1,1]},{e[1,2]},{e[1,3]},{e[2,1]},
{e[2,2]},{e[2,3]},{e[3,1]},{e[3,2]},{e[3,3]}};
e[4,1]=-e[3,1]-e[2,1]-e[1,1];e[4,2]=-e[3,2]-e[2,2]-e[1,2];
e[4,3]=-e[3,3]-e[2,3]-e[1,3];
```

*Next the covariance matrix w is created.

```
 h={{4.8, 2.1, 1.0},
    {2.1, 3.3, 1.4},
    {1.0, 1.4, 2.9}};
hw=h*0.5;
qw=h*0.25;
w=Join[Transpose[Join[h,hw,qw]],Transpose[Join[hw,h,hw]],
  Transpose[Join[qw,hw,h]]];
ssi=Inverse[w];
Do[si[i,j]=ssi[[Range[3i-2,3i],Range[3j-2,3j] ]],{i,3},{j,3}];
```

*The group means are set up.

```
v={{v1},{v2},{(1-v1^2-v2^2)^0.5}};
eg={{1,0,1},{0.5,1,-0.5},{-0.5,0,0.5},{-1,-1,-1}};
vf={0.5,0.5,(0.5^0.5)};
Do[x[i,j]=eg[[i,j]]vf,{i,4},{j,3}];
```

*The maximum likelihood equations are set up.

```
bb=Table[0,{11},{11}];
f=e[g,q]e[g,s]Transpose[v].si[q,s].v-
  2e[g,q]Transpose[v].si[q,s].x[g,s];
kk=Sum[f,{g,1,4},{q,1,3},{s,1,3}];
```

*The matrix of second order derivatives is obtained, then evaluated.

```
Do[bb[[i,j]]=D[kk,pm[[i,1]],pm[[j,1]]],{i,11},{j,11}];
tt=bb /. {v1->0.5, v2->0.5, e[1,1]->1, e[1,2]->0, e[1,3]->1,
    e[2,1]->0.5, e[2,2]->1, e[2,3]->-0.5, e[3,1]->-0.5,
    e[3,2]->0, e[3,3]->0.5};
```

*In the remaining code the information matrix is inverted and the asymptotic covariances are printed.

```
bbbb=tt;
bflat=Flatten[bbbb];
cd=Table[0,{11},{11}];
Do[cd[[i,j]]=bflat[[11(i-1)+j]], {i,11},{j,11}];
cdinv=Inverse[50cd];
Do[Print[cdinv[[i,i]]],{i,11}];
```

APPENDIX EIGHT

ASYMPTOTIC COVARIANCE MATRIX FOR THE PARAMETER ESTIMATES

Below is the matrix of asymptotic covariances between the parameter estimates based on the inverse of the information matrix for the model simulated in Section **8.4**. Since v_3 is a function of v_1 and v_2 it is not a free parameter and was not included. The same is true for e_{41}, e_{42} and e_{43}.

	v_1	v_2	e_{11}	e_{12}	e_{13}	e_{21}	e_{22}	e_{23}	e_{31}	e_{32}	e_{33}
v_1	0.002998	-0.00037	0.001426	0	0.001426	0.000713	0.001426	-0.00071	-0.00071	0	0.000713
v_2	-0.00037	0.001589	0.000789	0	0.000789	0.000395	0.000789	-0.00040	-0.00040	0	0.000395
e_{11}	0.001426	0.000789	0.040908	0.019814	0.011187	-0.01257	-0.00533	-0.00394	-0.01385	-0.00661	-0.00266
e_{12}	0	0	0.019814	0.039628	0.019814	-0.00661	-0.01321	-0.00661	-0.00661	-0.01321	-0.00661
e_{13}	0.001426	0.000789	0.011187	0.019814	0.040908	-0.00266	-0.00533	-0.01385	-0.00394	-0.00661	-0.01257
e_{21}	0.000713	0.000395	-0.01257	-0.00661	-0.00266	0.039948	0.020454	0.009587	-0.01353	-0.00661	-0.00298
e_{22}	0.001426	0.000789	-0.00533	-0.01321	-0.00533	0.020454	0.040908	0.019174	-0.00724	-0.01321	-0.00597
e_{23}	-0.00071	-0.00040	-0.00394	-0.00661	-0.01385	0.009587	0.019174	0.039948	-0.00298	-0.00661	-0.01353
e_{31}	-0.00071	-0.00040	-0.01385	-0.00661	-0.00394	-0.01353	-0.00724	-0.00298	0.039948	0.019814	0.009587
e_{32}	0	0	-0.00661	-0.01321	-0.00661	-0.00661	-0.01321	-0.00661	0.019814	0.039628	0.019814
e_{33}	0.000713	0.000395	-0.00266	-0.00661	-0.01257	-0.00298	-0.00597	-0.01353	0.009587	0.019814	0.039948