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 $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$. 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

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

It is apparent from (4.3) that the least squares solution for $\mathbf{\Lambda}_i$ is $\mathbf{\Lambda}_i = \text{Diag}(\mathbf{B}'\mathbf{S}_i\mathbf{B})$. 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

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

From this point the problem is equivalent to maximizing

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

where 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(**B**) with respect to $\boldsymbol{\beta}_r$, set equal to zero, yields

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

where I put $\ell_{rh} = \ell_{hr}$ if r > h. Multiplying (4.5) from the left by $(\tfrac{1}{2})\boldsymbol{\beta}_r'$ 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 $\mathrm{m,r}^{\mathrm{th}}$ 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 $\mathrm{m,r}^{\mathrm{th}}$ 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 $\mathrm{m,r}^{\mathrm{th}}$ 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

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

$1 \le m < r \le p$. 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 $n_i \to \infty$, $\mathbf{S}_i \to \mathbf{\Sigma}_i$ (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 $\mathbf{S}_i = \mathbf{B}_1\mathbf{\Lambda}_{1i}\mathbf{B}_1' + \mathbf{B}_{2i}\mathbf{\Lambda}_{2i}\mathbf{B}_{2i}'$, where $\mathbf{\Lambda}_{1i}$ is a $q \times q$ diagonal matrix and $\mathbf{\Lambda}_{2i}$ 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 $\mathbf{J}_i = \mathbf{B}_{21}'\mathbf{B}_{2i}\mathbf{\Lambda}_{2i}\mathbf{B}_{2i}'\mathbf{B}_{21}$. Then

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

Since $\|\mathbf{a}_d\|^2 = 1$ I can partition $\mathbf{a}_d$ as $\mathbf{a}_d = \begin{pmatrix} c\mathbf{a}_{1d} \\ \dots \\ f\mathbf{a}_{2d} \end{pmatrix}$, 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 $\mathbf{a}_d = \begin{pmatrix} \hat{\mathbf{a}}_{1d} \\ \dots \\ \mathbf{0} \end{pmatrix}$; if $\hat{t} < \hat{u}$ then $f = 1$ and $\mathbf{a}_d = \begin{pmatrix} \mathbf{0} \\ \dots \\ \hat{\mathbf{a}}_{2d} \end{pmatrix}$; 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 $\mathbf{B}'\mathbf{S}_i\mathbf{B}$, $\left(\boldsymbol{\beta}_m'\mathbf{S}_i\boldsymbol{\beta}_r\right)^2$, 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

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

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.