Chapter 7

Canonical Correlation Analysis

7.1 Introduction

Let $\boldsymbol{x}$ be the $p \times 1$ vector of predictors, and partition $\boldsymbol{x} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ where $\boldsymbol{w}$ is $m \times 1$ and $\boldsymbol{y}$ is $q \times 1$ where $m = p - q \leq q$ and $m, q \geq 2$. Canonical correlation analysis (CCA) seeks m pairs of linear combinations $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y}), ..., (\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ such that $\text{corr}(\boldsymbol{a}_i^T\boldsymbol{w}, \boldsymbol{b}_i^T\boldsymbol{y})$ is large under some constraints on the $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ where $i = 1, ..., m$. The first pair $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y})$ has the largest correlation. The next pair $(\boldsymbol{a}_2^T\boldsymbol{w}, \boldsymbol{b}_2^T\boldsymbol{y})$ has the largest correlation among all pairs uncorrelated with the first pair and the process continues so that $(\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ is the pair with the largest correlation that is uncorrelated with the first $m - 1$ pairs. The correlations are called *canonical correlations* while the pairs of linear combinations are called *canonical variables*.

Some notation is needed to explain CCA. Let the $p \times p$ positive definite symmetric dispersion matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Let $\boldsymbol{J} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Sigma}_A = \boldsymbol{J}\boldsymbol{J}^T = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$, $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_B = \boldsymbol{J}^T\boldsymbol{J} = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{e}_i$ and $\boldsymbol{g}_i$ be sets of orthonormal eigenvectors, so $\boldsymbol{e}_i^T\boldsymbol{e}_i = 1$, $\boldsymbol{e}_i^T\boldsymbol{e}_j = 0$ for $i \neq j$, $\boldsymbol{g}_i^T\boldsymbol{g}_i = 1$ and $\boldsymbol{g}_i^T\boldsymbol{g}_j = 0$ for $i \neq j$. Let the $\boldsymbol{e}_i$ be $m \times 1$ while the $\boldsymbol{g}_i$ are $q \times 1$.

Let $\boldsymbol{\Sigma}_a$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{a}_1), ..., (\lambda_m, \boldsymbol{a}_m)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Let $\boldsymbol{\Sigma}_A$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$ for $i =$

$1, ..., m$. Let $\boldsymbol{\Sigma}_b$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{b}_1), ..., (\lambda_q, \boldsymbol{b}_q)$. Let $\boldsymbol{\Sigma}_B$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{g}_i)$ for $i = 1, ..., q$. It can be shown that the m largest eigenvalues of the four matrices are the same. Hence $\lambda_i(\boldsymbol{\Sigma}_a) = \lambda_i(\boldsymbol{\Sigma}_A) = \lambda_i(\boldsymbol{\Sigma}_b) = \lambda_i(\boldsymbol{\Sigma}_B) \equiv \lambda_i$ for $i = 1, ..., m$. It can be shown that $\boldsymbol{a}_i = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_i$ and $\boldsymbol{b}_i = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{g}_i$. The eigenvectors $\boldsymbol{a}_i$ are not necessarily orthonormal and the eigenvectors $\boldsymbol{b}_i$ are not necessarily orthonormal.

Theorem 7.1. Assume the $p \times p$ dispersion matrix $\boldsymbol{\Sigma}$ is positive definite. Assume $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_b$ are positive definite and that $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$. Let $\boldsymbol{d}_i$ be an eigenvector of the corresponding matrix. Hence $\boldsymbol{d}_i = \boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{e}_i$ or $\boldsymbol{g}_i$. Let $(\hat{\lambda}_i, \hat{\boldsymbol{d}}_i)$ be the ith eigenvalue eigenvector pair of $\hat{\boldsymbol{\Sigma}}_\gamma$.

a) $\hat{\boldsymbol{\Sigma}}_\gamma \xrightarrow{P} \boldsymbol{\Sigma}_\gamma$ and $\hat{\lambda}_i(\hat{\boldsymbol{\Sigma}}_\gamma) \xrightarrow{P} \lambda_i(\boldsymbol{\Sigma}_\gamma) = \lambda_i$ where $\gamma = A, a, B$ or b.

b) $\boldsymbol{\Sigma}_\gamma \hat{\boldsymbol{d}}_i - \lambda_i \hat{\boldsymbol{d}}_i \xrightarrow{P} \boldsymbol{0}$ and $\hat{\boldsymbol{\Sigma}}_\gamma \boldsymbol{d}_i - \hat{\lambda}_i \boldsymbol{d}_i \xrightarrow{P} \boldsymbol{0}$.

c) If the jth eigenvalue λ_j is unique where $j \leq m$, then the absolute value of the correlation of $\hat{\boldsymbol{d}}_j$ with $\boldsymbol{d}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{d}}_j, \boldsymbol{d}_j)| \xrightarrow{P} 1$.

Proof. a) $\hat{\boldsymbol{\Sigma}}_\gamma \xrightarrow{P} \boldsymbol{\Sigma}_\gamma$ since matrix multiplication is a continuous function of the relevant matrices and matrix inversion is a continuous function of a positive definite matrix. Then $\hat{\lambda}_i(\hat{\boldsymbol{\Sigma}}_\gamma) \xrightarrow{P} \lambda_i$ since an eigenvalue is a continuous function of its associated matrix.

b) Note that $(\boldsymbol{\Sigma}_\gamma - \lambda_i \boldsymbol{I})\hat{\boldsymbol{d}}_i = [(\boldsymbol{\Sigma}_\gamma - \lambda_i \boldsymbol{I}) - (\hat{\boldsymbol{\Sigma}}_\gamma - \hat{\lambda}_i \boldsymbol{I})]\hat{\boldsymbol{d}}_i = o_P(1)O_P(1) \xrightarrow{P} \boldsymbol{0}$, and $\hat{\boldsymbol{\Sigma}}_\gamma \boldsymbol{d}_i - \hat{\lambda}_i \boldsymbol{d}_i \xrightarrow{P} \boldsymbol{\Sigma}_\gamma \boldsymbol{d}_i - \lambda_i \boldsymbol{d}_i = \boldsymbol{0}$.

c) If n is large, then $\hat{\boldsymbol{d}}_i \equiv \hat{\boldsymbol{d}}_{i,n}$ is arbitrarily close to either $\boldsymbol{d}_i$ or $-\boldsymbol{d}_i$, and the result follows.

Rule of thumb 7.1. To use CCA, assume the DD plot and subplots of the scatterplot matrix are linear. Want $n > 10p$ for classical CCA and $n > 20p$ for robust CCA that uses FCH, RFCH or RMVN. Also make the DD plot for the $\boldsymbol{y}$ variables and the DD plot for the $\boldsymbol{z}$ variables.

Definition 7.1. Let the dispersion matrix be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Let $(\lambda_i, \boldsymbol{e}_i)$ and $(\lambda_i, \boldsymbol{g}_i)$ be the eigenvalue eigenvector pairs of $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$. The kth pair of *population canonical variables* is

$$U_k = \boldsymbol{a}_k^T \boldsymbol{w} = \boldsymbol{e}_k^T \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{w} \text{ and } V_k = \boldsymbol{b}_k^T \boldsymbol{y} = \boldsymbol{g}_k^T \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{y}$$

for $k = 1, ..., m$. Then the *population canonical correlations* $\rho_k = corr(U_k, V_k)$

$= \sqrt{\lambda_k}$ for $k = 1, ..., m$. The vectors $\boldsymbol{a}_k = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_k$ and $\boldsymbol{b}_k = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{g}_k$ are the kth *canonical correlation coefficient vectors* for $\boldsymbol{w}$ and $\boldsymbol{y}$.

Theorem 7.2. Johnson and Wichern (1988, p. 440-441): Let the dispersion matrix be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then $V(U_k) = V(V_k) = 1$, $\text{Cov}(C_k, D_j) = corr(C_k, D_j) = 0$ for $k \neq j$ where $C_k = U_k$ or $C_k = V_k$, and $D_j = U_j$ or $D_j = V_j$ and $j, k = 1, ..., m$. That is, U_k is uncorrelated with V_j and U_j for $j \neq k$, and V_k is uncorrelated with V_j and U_j for $j \neq k$. The first pair of canonical variables is the pair of linear combinations (U, V) having unit variances that maximizes $\text{corr}(U, V)$ and this maximum is $\text{corr}(U_1, V_1) = \rho_1$. The ith pair of canonical variables are the linear combinations (U, V) with unit variances that maximize $\text{corr}(U, V)$ among all choices uncorrelated with the previous $k - 1$ canonical variable pairs.

Definition 7.2. Suppose standardized data $\boldsymbol{z} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ is used and the dispersion matrix is the correlation matrix $\boldsymbol{\Sigma} = \boldsymbol{\rho}$. Hence $\boldsymbol{\Sigma}_{ii} = \boldsymbol{\rho}_{ii}$ for $i = 1, 2$. Let $(\lambda_i, \boldsymbol{e}_i)$ and $(\lambda_i, \boldsymbol{g}_i)$ be the eigenvalue eigenvector pairs of $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$. The kth pair of *population canonical variables* is

$$U_k = \boldsymbol{a}_k^T\boldsymbol{w} = \boldsymbol{e}_k^T\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{w} \text{ and } V_k = \boldsymbol{b}_k^T\boldsymbol{y} = \boldsymbol{g}_k^T\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{y}$$

for $k = 1, ..., m$ for $k = 1, ..., m$. Then the *population canonical correlations* $\rho_k = corr(U_k, V_k) = \sqrt{\lambda_k}$ for $k = 1, ..., m$.

Then Theorem 7.2 holds for the standardized data and the canonical correlations are unchanged by the standardization.

Let

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_{22} \end{pmatrix}.$$

Define estimators $\hat{\boldsymbol{\Sigma}}_a$, $\hat{\boldsymbol{\Sigma}}_A$, $\hat{\boldsymbol{\Sigma}}_b$ and $\hat{\boldsymbol{\Sigma}}_B$ in the same manner as their population analogs but using $\hat{\boldsymbol{\Sigma}}$ instead of $\boldsymbol{\Sigma}$. For example, $\hat{\boldsymbol{\Sigma}}_a = \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21}$.

Let $\hat{\boldsymbol{\Sigma}}_a$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{a}}_i)$, and let $\hat{\boldsymbol{\Sigma}}_A$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ for $i = 1, ..., m$. Let $\hat{\boldsymbol{\Sigma}}_b$ have eigenvalue eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{b}}_1)$, and let $\hat{\boldsymbol{\Sigma}}_B$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{g}}_i)$ for $i = 1, ..., q$. For these four matrices $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_m$.

Definition 7.3. Let $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$ if data $\boldsymbol{x} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ is used, and let $\hat{\boldsymbol{\Sigma}} = \boldsymbol{R}$ if standardized data $\boldsymbol{z} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ is used. The kth pair of *sample*

canonical variables is

$$\hat{U}_k = \hat{\boldsymbol{a}}_k^T \boldsymbol{w} = \hat{\boldsymbol{e}}_k^T \hat{\boldsymbol{\Sigma}}_{11}^{-1/2} \boldsymbol{w} \text{ and } \hat{V}_k = \hat{\boldsymbol{b}}_k^T \boldsymbol{y} = \hat{\boldsymbol{g}}_k^T \hat{\boldsymbol{\Sigma}}_{22}^{-1/2} \boldsymbol{y}$$

for $k = 1, ..., m$. Then the *sample canonical correlations* $\hat{\rho}_k = corr(\hat{U}_k, \hat{V}_k)$ $= \sqrt{\hat{\lambda}_k}$ for $k = 1, ..., m$. The vectors $\hat{\boldsymbol{a}}_k = \hat{\boldsymbol{\Sigma}}_{11}^{-1/2} \hat{\boldsymbol{e}}_k$ and $\hat{\boldsymbol{b}}_k = \hat{\boldsymbol{\Sigma}}_{22}^{-1/2} \hat{\boldsymbol{g}}_k$ are the kth *sample canonical correlation vectors* for $\boldsymbol{w}$ and $\boldsymbol{y}$.

Theorem 7.3. Under the conditions of Definition 7.3, the first pair of canonical variables $(\hat{U}_1, \hat{V}_1)$ is the pair of linear combinations $(\hat{U}, \hat{V})$ having unit sample variances that maximizes the sample correlation $\text{corr}(\hat{U}, \hat{V})$ and this maximum is $\text{corr}(\hat{U}_1, \hat{V}_1) = \hat{\rho}_1$. The ith pair of canonical variables are the linear combinations $(\hat{U}, \hat{V})$ with unit sample variances that maximize the sample $\text{corr}(\hat{U}, \hat{V})$ among all choices uncorrelated with the previous $k-1$ canonical variable pairs.

7.2 Robust CCA

The *R* function `cancor` does classical CCA and the *mpack* function `rcancor` does robust CCA by applying `cancor` on the RMVN set: the subset of the data used to compute RMVN.

Some theory is simple: the FCH, RFCH and RMVN methods of RCCA produce consistent estimators of the kth canonical correlation ρ_k on a large class of elliptically contoured distributions.

To see this, suppose $\text{Cov}(\boldsymbol{x}) = c_{\boldsymbol{x}} \boldsymbol{\Sigma}$ and $\boldsymbol{C} \equiv \boldsymbol{C}(\boldsymbol{X}) \xrightarrow{P} c\boldsymbol{\Sigma}$ where $c_x > 0$ and $c > 0$ are some constants. Then $\boldsymbol{C}_{XX}^{-1} \boldsymbol{C}_{XY} \boldsymbol{C}_{YY}^{-1} \boldsymbol{C}_{YX} \xrightarrow{P} \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_{YY}^{-1} \boldsymbol{\Sigma}_{YX}$, and $\boldsymbol{C}_{YY}^{-1} \boldsymbol{C}_{YX} \boldsymbol{C}_{XX}^{-1} \boldsymbol{C}_{XY} \xrightarrow{P} \boldsymbol{\Sigma}_B = \boldsymbol{\Sigma}_{YY}^{-1} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY}$. Note that $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$ only depend on $\boldsymbol{\Sigma}$ and do not depend on the constants c or $c_{\boldsymbol{x}}$.

(If $\boldsymbol{C}$ is also the classical covariance matrix applied to some subset of the data, then the correlation matrix $\boldsymbol{G} \equiv \boldsymbol{R}_{\boldsymbol{C}}$ applied to the same subset satisfies $\boldsymbol{G}_{XX}^{-1} \boldsymbol{G}_{XY} \boldsymbol{G}_{YY}^{-1} \boldsymbol{G}_{YX} \xrightarrow{P} \boldsymbol{R}_A = \boldsymbol{R}_{XX}^{-1} \boldsymbol{R}_{XY} \boldsymbol{R}_{YY}^{-1} \boldsymbol{R}_{YX}$, and $\boldsymbol{G}_{YY}^{-1} \boldsymbol{G}_{YX} \boldsymbol{G}_{XX}^{-1} \boldsymbol{G}_{XY} \xrightarrow{P} \boldsymbol{R}_B = \boldsymbol{R}_{YY}^{-1} \boldsymbol{R}_{YX} \boldsymbol{R}_{XX}^{-1} \boldsymbol{R}_{XY}$.)

Since eigenvalues are continuous functions of the associated matrix, and the FCH, RFCH and RMVN estimators are consistent estimators of $c_1\boldsymbol{\Sigma}, c_2\boldsymbol{\Sigma}$ and $c_3\boldsymbol{\Sigma}$ on a large class of elliptically contoured distributions, Theorem

7.1 holds, so these three RCCA methods and `rcancor` produce consistent estimators the kth canonical correlation ρ_k on that class of distributions.

Example 7.1. Example 2.2 describes the mussel data. Log transformation were taken on *muscle mass* M, *shell width* W and on the *shell mass* S. Then x contained the two log mass measurements while y contains L, H and $\log(W)$. The robust and classical CCAs were similar, but the canonical coefficients were difficult to interpret since $\log(W)$ has different units than L and H. Hence the log transformation were taken on all five variables and output is shown below.

The data set zm contains x and y, and the DD plot showed case 48 was separated from the bulk of the data, but near the identity line. The DD plot for x showed two cases, 8 and 48, were separated from the bulk of the data. Also the plotted points did not cluster tightly about the identity line. The DD plot for y looked fine. The classical CCA produces output \$cor, \$xcoef and \$ycoef. These are the canonical correlations, the $\boldsymbol{a}_i$ and the $\boldsymbol{b}_i$. The labels for the RCCA are \$out\$cor, \$out\$xcoef and \$out\$ycoef.

Note that the first correlation was about 0.98 while the second correlation was small. The RCCA is the CCA on the RMVN data set, which is contained in a compact ellipsoidal region. The variability of the truncated data set is less than that of the entire data set, hence expect the robust $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ to be larger in magnitude, ignoring sign, than that of the classical $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, since the variance of each canonical variate is equal to one, and RCCA uses the truncated data. Note that $\boldsymbol{a}_1$ was roughly proportional to $\log(S)$ while $\boldsymbol{b}_1$ gave slightly higher weight for $\log(H)$ then $\log(W)$ and then $\log(L)$. Note that the five variables have high pairwise correlations, so $\log(M)$ was not important given that $\log(S)$ was in x. The second pair $(\boldsymbol{a}_2, \boldsymbol{b}_2)$ might be ignored since the second canonical correlation was very low.

```
> cancor(x,y)
$cor
[1] 0.9818605 0.1555381

$xcoef
        [,1]       [,2]
S 0.12650486  0.4077765
M 0.01897332 -0.4872522
```

```
$ycoef
       [,1]       [,2]       [,3]
L 0.1567463  0.7277888  2.1935890
W 0.1605139  0.8650480 -1.0676419
H 0.2143781 -2.0634587 -0.8303862

$xcenter
       S        M
4.563856 2.850187

$ycenter
       L        W        H
5.472944 3.697654 4.723295

> rcancor(x,y)
$out
$out$cor
[1] 0.98596703 0.06797587

$out$xcoef
        [,1]       [,2]
S 0.14966183  0.6460117
M 0.03236328 -0.8543387

$out$ycoef
       [,1]       [,2]       [,3]
L 0.1625452  0.4237524 -2.8492678
W 0.2369692  1.5379681  0.9356495
H 0.2530324 -2.6806462  1.7785931

$out$xcenter
       S        M
4.651941 2.948571

$out$ycenter
       L        W        H
5.496255 3.728292 4.745839
```

7.3 Summary

1) Let $\boldsymbol{x}$ be the $p \times 1$ vector of predictors, and partition $\boldsymbol{x} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ where $\boldsymbol{w}$ is $m \times 1$ and $\boldsymbol{y}$ is $q \times 1$ where $m = p - q \leq q$ and $m, q \geq 2$. Canonical correlation analysis (CCA) seeks m pairs of linear combinations $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y}), ..., (\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ such that $\text{corr}(\boldsymbol{a}_i^T\boldsymbol{w}, \boldsymbol{b}_i^T\boldsymbol{y})$ is large under some constraints on the $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ where $i = 1, ..., m$. The first pair $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y})$ has the largest correlation. The next pair $(\boldsymbol{a}_2^T\boldsymbol{w}, \boldsymbol{b}_2^T\boldsymbol{y})$ has the largest correlation among all pairs uncorrelated with the first pair and the process continues so that $(\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ is the pair with the largest correlation that is uncorrelated with the first $m - 1$ pairs. The correlations are called *canonical correlations* while the pairs of linear combinations are called *canonical variables*.

2) R output is shown in symbols for the following table.

corr					
$\hat{\rho}_1$	$\cdots$	$\hat{\rho}_1$			
wcoef					
$\boldsymbol{w}$	$\hat{\boldsymbol{a}}_1$	$\cdots$	$\hat{\boldsymbol{a}}_m$		
ycoef					
$\boldsymbol{y}$	$\hat{\boldsymbol{b}}_1$	$\cdots$	$\hat{\boldsymbol{b}}_m$	$\cdots$	$\hat{\boldsymbol{b}}_q$

```
64) $out$cor
[1] 0.98596703 0.06797587   $out$ycoef
$out$xcoef                        [,1]       [,2]       [,3]
        [,1]       [,2]    L 0.1625452  0.4237524 -2.8492678
S 0.14966183  0.6460117    W 0.2369692  1.5379681  0.9356495
M 0.03236328 -0.8543387    H 0.2530324 -2.6806462  1.7785931
```

3) Some notation is needed to explain CCA. Let the $p \times p$ positive definite symmetric dispersion matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Let $\boldsymbol{J} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Sigma}_A = \boldsymbol{J}\boldsymbol{J}^T = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$, $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_B = \boldsymbol{J}^T\boldsymbol{J} = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{e}_i$ and $\boldsymbol{g}_i$ be sets of orthonormal eigenvectors, so $\boldsymbol{e}_i^T\boldsymbol{e}_i = 1$, $\boldsymbol{e}_i^T\boldsymbol{e}_j = 0$ for $i \neq j$, $\boldsymbol{g}_i^T\boldsymbol{g}_i = 1$ and $\boldsymbol{g}_i^T\boldsymbol{g}_j = 0$ for $i \neq j$. Let the $\boldsymbol{e}_i$ be $m \times 1$ while the $\boldsymbol{g}_i$ are $q \times 1$.

Let $\boldsymbol{\Sigma}_a$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{a}_1), ..., (\lambda_m, \boldsymbol{a}_m)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Let $\boldsymbol{\Sigma}_A$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$ for $i =$

$1, ..., m$. Let $\boldsymbol{\Sigma}_b$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{b}_1), ..., (\lambda_q, \boldsymbol{b}_q)$. Let $\boldsymbol{\Sigma}_B$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{g}_i)$ for $i = 1, ..., q$. It can be shown that the m largest eigenvalues of the four matrices are the same. Hence $\lambda_i(\boldsymbol{\Sigma}_a) = \lambda_i(\boldsymbol{\Sigma}_A) = \lambda_i(\boldsymbol{\Sigma}_b) = \lambda_i(\boldsymbol{\Sigma}_B) \equiv \lambda_i$ for $i = 1, ..., m$. It can be shown that $\boldsymbol{a}_i = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_i$ and $\boldsymbol{b}_i = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{g}_i$. The eigenvectors $\boldsymbol{a}_i$ are not necessarily orthonormal and the eigenvectors $\boldsymbol{b}_i$ are not necessarily orthonormal.

Theorem 7.1. Assume the $p \times p$ dispersion matrix $\boldsymbol{\Sigma}$ is positive definite. Assume $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_b$ are positive definite and that $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$. Let $\boldsymbol{d}_i$ be an eigenvector of the corresponding matrix. Hence $\boldsymbol{d}_i = \boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{e}_i$ or $\boldsymbol{g}_i$. Let $(\hat{\lambda}_i, \hat{\boldsymbol{d}}_i)$ be the ith eigenvalue eigenvector pair of $\hat{\boldsymbol{\Sigma}}_\gamma$.

a) $\hat{\boldsymbol{\Sigma}}_\gamma \xrightarrow{P} \boldsymbol{\Sigma}_\gamma$ and $\hat{\lambda}_i(\hat{\boldsymbol{\Sigma}}_\gamma) \xrightarrow{P} \lambda_i(\boldsymbol{\Sigma}_\gamma) = \lambda_i$ where $\gamma = A, a, B$ or b.

b) $\boldsymbol{\Sigma}_\gamma \hat{\boldsymbol{d}}_i - \lambda_i \hat{\boldsymbol{d}}_i \xrightarrow{P} \boldsymbol{0}$ and $\hat{\boldsymbol{\Sigma}}_\gamma \boldsymbol{d}_i - \hat{\lambda}_i \boldsymbol{d}_i \xrightarrow{P} \boldsymbol{0}$.

c) If the jth eigenvalue λ_j is unique where $j \leq m$, then the absolute value of the correlation of $\hat{\boldsymbol{d}}_j$ with $\boldsymbol{d}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{d}}_j, \boldsymbol{d}_j)| \xrightarrow{P} 1$.

7.4 Complements

Muirhead and Waternaux (1980) shows that if the population canonical correlations ρ_k are distinct and if the underlying population distribution has a finite fourth moments, then the limiting joint distribution of $\sqrt{n}(\hat{\rho}_k^2 - \rho_k^2)$ is multivariate normal where the $\hat{\rho}_k$ are the classical sample canonical correlations and $k = 1, ..., p$. If the data are iid from an elliptically contoured distribution with kurtosis 3κ, then the limiting joint distribution of

$$\sqrt{n}\ \frac{\hat{\rho}_k^2 - \rho_k^2}{2\rho_k(1 - \rho_k^2)}$$

for $k = 1, ..., p$ is $N_p(\boldsymbol{0}, (\kappa + 1)\boldsymbol{I}_p)$. Note that $\kappa = 0$ for multivariate normal data.

Alkenani and Yu (2012), Zhang (2011) and Zhang, Olive and Ye (2012) develop robust CCA based on FCH, RFCH and RMVN.

7.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USE-

FUL.

7.1*. Examine the R output in Example 7.1. a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

7.2. The R output below is for a canonical correlation analysis on Venables and Ripley (2003) CPU data. The variables were syct = log(cycle time + 1),
mmin = log(minimum main memory + 1),
chmin = log(minimum number of channels + 1),
chmax = log(maximum number of channels + 1),
perf = log(published performance + 1) and
estperf = $20/\sqrt{(}$estimated performance$+1)$. These six variables had a linear scatterplot matrix and DD plot and similar variances. Want to compare the two performance variables with the four remaining variables.

a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

d) Interpret the second canonical variable $U_2 = \hat{\boldsymbol{a}}_2^T \boldsymbol{w}$.

```
> cancor(w,y)
$cor
[1] 0.8769433 0.2278554

$xcoef
```

```
              [,1]      [,2]
perf     0.02536432 0.1558717
estperf -0.04121870 0.1431100

$ycoef
              [,1]        [,2]         [,3]         [,4]
syct  -0.013613254  0.05700360  0.089757416 -0.011423664
mmin   0.037485282 -0.01874858  0.084442460  0.005859654
chmin  0.006932264  0.09843612 -0.021782624  0.090756713
chmax  0.019998948  0.01159728  0.007855559 -0.094198608
```

7.3. Edited SAS output for SAS Institute (1985, p. 146) Fitness Club Data is given below for CCA. Three physiological and three exercise variables measured on 20 middle aged men at a fitness club.

a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

```
 Canonical
Correlation
0.7956
0.2006
0.0726
```

```
Raw Canonical Coefficients for the Physiological Variables
        PHYS1    PHYS2    PHYS3
weight -0.0314  -0.0763  -0.0077
waist   0.0493   0.3687   0.1580
pulse  -0.0082  -0.0321   0.1457

Raw Canonical Coefficients for the Exercise Variables
         Exer1    Exer2   Exer3
chinups -0.0661  -0.0714  -0.2428
situps  -0.0168   0.0020   0.0198
jumps    0.0140   0.0207  -0.0082
```

7.4. The output below is for a canonical correlations analysis on the R Seatbelts data set where $y_1 = drivers$ = number of drivers killed or seriously injured, $y_2 = front$ = number of front seat passengers killed or seriously injured, and $y_3 = rear$ = number of back seat passengers killed or seriously injured, $x_1 = kms$ = distance driven, $x_2 = PetrolPrice$ = petrol price and $x_3 = VanKilled$ = number of van drivers killed. The data consists of 192 monthly totals in Great Britain from January 1969 to December 1984.

a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

d) Let $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. The from the DD plot, the $\boldsymbol{z}_i$ appeared to follow a multivariate normal distribution. Sketch the DD plot.

```
> rcancor(x,y)
$out
$out$cor
```

```
[1] 0.8116953 0.5064619 0.1376399

$out$xcoef
                       [,1]          [,2]          [,3]
x.kms         -2.080206e-05 -0.0000233873 -2.259723e-06
x.PetrolPrice -1.847967e+00  3.7173715818  5.292041e+00
x.VanKilled    1.597620e-03 -0.0168450843  1.673662e-02

$out$ycoef
                    [,1]          [,2]          [,3]
y.drivers  1.678751e-06 -2.487259e-05  0.0004717902
y.front    5.594715e-04 -7.797027e-05 -0.0008157585
y.rear    -9.964980e-04 -7.521578e-04  0.0005045756
```

7.5. The R output below is for a canonical correlation analysis on some iris data. An iris is a flower, and there were 50 observations with 4 variables sepal length, sepal width, petal length and petal width.

a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

```
w<-iris3[,,3]
x <- w[,1:2]
y <- w[,3:4]
cancor(x,y)

$cor
[1] 0.8642869 0.4836991

$xcoef
                 [,1]       [,2]
Sepal L. -0.223034210 -0.1186117
```

```
Sepal W. -0.006920448  0.4980378

$ycoef
                 [,1]        [,2]
Petal L. -0.257853414 -0.09094352
Petal W. -0.006108292  0.54939125
```