

THE INSTITUTE  
OF STATISTICS

THE CAROLINA COLLEGE UNIVERSITY  
OF NORTH CAROLINA



RESTRICTED CANONICAL CORRELATIONS  
(DISSERTATION)

by

Shubhabrata Das

April 1993

Mimeo Series #2095

DEPARTMENT OF STATISTICS  
Chapel Hill, North Carolina

MIMEO  
SERIES  
2095

Shubhabrata Das  
RESTRICTED CANONICAL  
CORRELATION

NAME

DATE

Department of Statistics Library

# Restricted Canonical Correlations

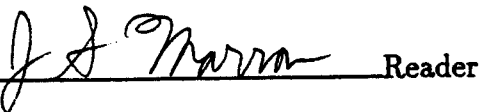
Shubhabrata Das

A dissertation submitted to the faculty of the University of North Carolina  
in partial fulfillment of the requirements for the degree of Doctor of  
Philosophy in the Department of Statistics

Chapel Hill, 1993

Approved by:

  
\_\_\_\_\_  
Advisor

  
\_\_\_\_\_  
Reader

  
\_\_\_\_\_  
Reader

SHUBHABRATA DAS. Restricted Canonical Correlations. (Under the direction of Pranab Kumar Sen.)

### Abstract

Given a  $p$ -dimensional random variable  $Y^{(1)}$  and a  $q$ -dimensional random variable  $Y^{(2)}$ , the first canonical correlation leads to finding  $\alpha^* \in R_p$  and  $\beta^* \in R_q$  which maximizes the correlation between  $\alpha'Y^{(1)}$  and  $\beta'Y^{(2)}$ . However, in many practical situations (e.g. educational testing problems, neural networks), some natural restrictions on the coefficients  $\alpha$  and  $\beta$  may arise which should be incorporated in the maximization procedure. The maximum correlation subject to these constraints is referred to as restricted canonical correlations.

In this work the solution is obtained to the problem under the nonnegativity restriction on  $\alpha$  and  $\beta$ . The analysis is extended to more general form of inequality constraints, and also when the restrictions are present only on some of the coefficients. Also discussed are restricted versions of some other multivariate methods. This includes principal component analysis and different modifications of canonical correlation analysis.

Some properties of restricted canonical correlations including its bounds have been studied. Since the restricted canonical correlation depends only on the covariance matrix, its sample version can naturally be obtained from sample covariance matrix. The asymptotic normality of this sample restricted canonical correlation is proved under reasonably mild conditions.

The study of resampling methods becomes necessary because the asymptotic variance involves usually unknown fourth order moments of the population. The effectiveness of the jackknife method is shown in this case as well as for the usual canonical correlations. Bootstrapping also works out in both these cases. These theoretical results have been supplemented by simulation studies which compare the performances of the two resampling methods in the case of finite samples.

## ACKNOWLEDGEMENTS

I find it difficult to express in few words my deep appreciation for Prof. P.K. Sen; I am grateful to him for being a mentor and introducing me to different areas of statistics. In spite of being extremely busy, he always had time whenever we needed to discuss.

I would also like to thank other members of the committee (Professors E. Carlstein, G. Kallianpur, D.G. Kelly, and J.S. Marron) for being very encouraging and helpful during my studies. I am also grateful to Prof. Stamatis Cambanis for helping me so many times in the last 4 years.

I could not have gone through the graduate program without the support and warmth from my dear friends. Especially, I would like to thank Elisabeti Kira, Antonio Carlos Pedroso de Lima, and Charu Krishnamoorthy who helped me with writing the dissertation and computer programming. It is entirely their credit, that, now I feel less mystified by computer. Finally I would like to express my gratitude for my parents. Their inspiration and complete faith have always acted as primary motivation in me.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Why Restricted Canonical Correlation ? . . . . .	3
1.2.1	Scholastics Studies . . . . .	3
1.2.2	Study of Simultaneous Spike-trains in Neurophysiology . .	4
1.3	Different types of useful Restrictions . . . . .	8
1.4	Restrictions which can be reduced to Nonnegativity type . . . . .	9
1.5	Minimum Restricted Canonical Correlation . . . . .	11
<b>2</b>	<b>Canonical Correlation</b>	<b>12</b>
2.1	Introduction . . . . .	12
2.2	Population Canonical Correlations . . . . .	13
2.3	Sample Canonical Correlation and its Sampling Distribution . . .	15

2.4	Canonical Correlation as a “Smooth” Function and its Asymptotic Linearity . . . . .	17
2.4.1	Canonical Correlation as a Function of Covariance matrix	17
2.4.2	Asymptotic Linearity . . . . .	19
2.5	Interpretation, Usage, Disadvantages and Relation with other Multivariate Methods . . . . .	21
2.5.1	Determining the Effective Number of Common Factors . .	21
2.5.2	Prediction of Original Variables . . . . .	22
2.5.3	Measure of Affine Dependence . . . . .	23
2.5.4	Canonical Loadings, Redundancy Coefficient . . . . .	24
2.5.5	Canonical Correlation Analysis in Relation to other Multivariate Methods . . . . .	26
2.6	Generalized and Constrained Canonical Correlation . . . . .	27
2.6.1	Generalization to More Than Two Sets of Random Variables	27
2.6.2	When the Covariance Matrix is Singular . . . . .	28
2.6.3	Constrained/restricted Canonical Correlations . . . . .	28
<b>3</b>	<b>Analysis of Restricted Canonical Correlation and Related Topics</b>	<b>30</b>
3.1	Introduction . . . . .	30
3.2	Evaluating Restricted Canonical Correlation . . . . .	31

3.2.1	Calculation of Restricted Canonical Correlation using Kuhn-Tucker Theory . . . . .	31
3.2.2	Notation for Subvector and "Proper" Submatrices . . . . .	35
3.2.3	Restricted Canonical Correlation is Canonical Correlation for some "Proper" Submatrix . . . . .	35
3.3	Partially Restricted Canonical Correlation . . . . .	37
3.4	Higher Order Restricted Canonical Correlation . . . . .	39
3.4.1	Approach 1 . . . . .	41
3.4.2	Approach 2 . . . . .	43
3.5	Other Restricted Measures . . . . .	45
3.5.1	Restricted Principal Component . . . . .	45
3.5.2	Restricted and Unrestricted Part, Partial and Bipartial Canonical Correlations . . . . .	46
<b>4</b>	<b>Properties and Sampling Distribution</b>	<b>48</b>
4.1	Introduction . . . . .	48
4.2	Properties of Restricted Canonical Correlation . . . . .	49
4.2.1	Upper and Lower Bounds . . . . .	49
4.2.2	Invariance under Nonsingular Transformation . . . . .	51
4.3	Asymptotic Distribution of Sample Restricted Canonical Correlation	53
4.3.1	Notations and Assumptions . . . . .	53



4.3.2	Representing Sample Restricted Canonical Correlation in terms of Sample Canonical Correlations . . . . .	54
4.3.3	Preliminary Results . . . . .	55
4.3.4	Main Results . . . . .	56
<b>5</b>	<b>Resampling methods in the context of Restricted and usual Canonical Correlations</b>	<b>59</b>
5.1	Introduction . . . . .	59
5.2	Jackknifing Canonical Correlation . . . . .	60
5.3	Jackknifing Restricted Canonical Correlation . . . . .	64
5.4	Bootstrap for Restricted and usual Canonical Correlations . . . . .	68
<b>6</b>	<b>Simulation Study</b>	<b>70</b>
6.1	Introduction . . . . .	70
6.2	Generating Samples from MP and MN Distribution, and choice of simulation parameters . . . . .	71
6.3	Finite Sample Distribution . . . . .	74
6.4	Comparing Performances of the Resampling methods in Estimating Sample Variance . . . . .	85
6.5	Bias Reduction for Canonical Correlation and Restricted Canonical Correlation . . . . .	92
<b>A</b>	<b>Multivariate Poisson Distribution</b>	<b>98</b>

<b>B</b>	<b>Some Results on Jackknifing U-Statistics with a kernel of degree 2</b>	<b>100</b>
<b>C</b>	<b>Computer Programs</b>	<b>103</b>
C.1	Program to Compute Restricted Canonical Correlation . . . . .	103
C.2	Program to Generate sample Restricted Canonical Correlation Replicates to Compute Simulated Bias and Variance . . . . .	106
C.3	Program for Jackknifing Restricted Canonical Correlation . . . . .	109
C.4	Program for Bootstrapping Restricted Canonical Correlation . . . . .	113
C.5	Simulating from Multivariate Normal Distribution . . . . .	116

# List of Tables

6.1	Estimate of Var(RCC) : MP population, 1st covariance . . . . .	87
6.2	Estimate of Var(CC) : MP population, 1st covariance . . . . .	87
6.3	Estimate of Var(RCC) : MN population, 1st covariance . . . . .	88
6.4	Estimate of Var(CC) : MN population, 1st covariance . . . . .	88
6.5	Estimate of Var(RCC) : MP population, 2nd covariance . . . . .	89
6.6	Estimate of Var(CC) : MP population, 2nd covariance . . . . .	89
6.7	Estimate of Var(RCC) : MN population, 2nd covariance . . . . .	90
6.8	Estimate of Var(CC) : MN population, 2nd covariance . . . . .	90
6.9	Estimate of Var(RCC) : MN population, 3rd covariance . . . . .	91
6.10	Estimate of Var(CC) : MN population, 3rd covariance . . . . .	91
6.11	Estimate of RCC : MP population, 1st covariance . . . . .	93
6.12	Estimate of CC : MP population, 1st covariance . . . . .	93
6.13	Estimate of RCC : MN population, 1st covariance . . . . .	94
6.14	Estimate of CC : MN population, 1st covariance . . . . .	94
6.15	Estimate of RCC : MP population, 2nd covariance . . . . .	95
6.16	Estimate of CC : MP population, 2nd covariance . . . . .	95
6.17	Estimate of RCC : MN population, 2nd covariance . . . . .	96
6.18	Estimate of CC : MN population, 2nd covariance . . . . .	96
6.19	Estimate of RCC : MN population, 3rd covariance . . . . .	97
6.20	Estimate of CC : MN population, 3rd covariance . . . . .	97

# List of Figures

1.1	Location of sites in Kuperstein-Whittington's multichannel micro-electrode . . . . .	6
1.2	Another possible arrangement of micro-electrodes . . . . .	7
6.1	Sampling distribution of RCC: MP population, 1st covariance . .	75
6.2	Sampling distribution of CC: MP population, 1st covariance . . .	76
6.3	Sampling distribution of RCC: MN population, 1st covariance . .	77
6.4	Sampling distribution of CC: MN population, 1st covariance . . .	78
6.5	Sampling distribution of RCC: MP population, 2nd covariance . .	79
6.6	Sampling distribution of CC: MP population, 2nd covariance . . .	80
6.7	Sampling distribution of RCC: MN population, 2nd covariance . .	81
6.8	Sampling distribution of CC: MN population, 2nd covariance . . .	82
6.9	Sampling distribution of RCC: MN population, 3rd covariance . .	83
6.10	Sampling distribution of CC: MN population, 3rd covariance . . .	84

## LIST OF ABBREVIATIONS

CC	Canonical Correlation
CCA	Canonical Correlation Analysis
COV	Covariance
CNS	Central Nervous System
i.i.d.	Independent and Identically Distributed
max	Maximum
min	Minimum
MP	Multivariate Poisson
MN	Multivariate Normal
PRCC	Partially Restricted Canonical Correlation
RCC	Restricted Canonical Correlation
VAR	Variance

# Chapter 1

## Introduction

### 1.1 Introduction

The concept of usual correlation originated in the work of Sir Francis Galton (1889). This was mathematically formulated by Pearson (1896,1898). Since then the concept of correlation has been generalized to include part, partial and bi-partial correlations. All these correlations are measures of dependence between two random variables. As opposed to these three forms of correlation, multiple correlation measures dependence between a random variable, termed “dependent variable”, and a random vector (i.e. a set of random variables), termed “independent variables. This naturally extends to a study of dependence between two sets of random variables through canonical correlation (CC). CC, Hotelling’s novel idea, is defined as the maximum correlation between any two *representative* random variables of the two sets, where *representative* of a set of random variables is simply any linear combination of these random variables. The idea of linear combination originated from the linearity of regression in Multivariate Normal (MN) distributions.

However, in many practical situations, there may be some natural restrictions on these linear combinations. In such a case, traditional CC may overemphasize the dependence between the two sets, and it would be more appropriate to consider the maximization problem with the *representatives* being restricted accordingly. Such a measure of dependence can be termed as *restricted canonical correlation* (RCC). The primary focus of this study is to find ways to calculate RCC as well as to explore the statistical properties of its sample version. Similar restricted versions of other multivariate exploratory analysis is also meaningful. For example, Timm and Carlson (1976) considered part and bipartial canonical correlations. Restricted versions of these, as well as that of classical principal component analysis, have been dealt with in section 3.5.

Section 1.2 is devoted to the need for studying RCC. Although varieties of restrictions are of practical interest, it seems unlikely that a unifying treatment of all such cases is feasible. Some of these different restrictions are described in section 1.3; but for most of the statistical analysis, only the restriction, that the contributing coefficients of the original variables in their *representative* random variables be nonnegative, is considered. This is partly because one would expect the *representative* to be a convex linear combination of the original random variables and since correlation is scale-invariant, this requires only nonnegativity of the coefficients. More importantly, this is an important case, where it was possible to get the exact analytic solution. Also, many other interesting cases may be reduced to this one using suitable transformations. This has been discussed in section 1.4. In general, this class of restrictions is referred to as the *inequality* type. However, there remain many other types of restrictions where it was not possible to find an exact analytic solution. Thus in many circumstances, it may be necessary to settle for approximate solutions. In such a case, numerical analysis and/or a neural network approach may be successful; but this is not pursued in this work. In section 1.5, the notion of minimum RCC has been introduced.

Chapter 2 is mostly a review of canonical correlation analysis (CCA). But,

section 2.4 also includes work which is not readily available in the literature. The analysis and methods to find RCC and other related measures have been discussed in chapter 3. Chapter 4 deals with properties of RCC and the study of the asymptotic distribution of its sample counterpart. In chapter 5, suitable resampling methods for both CC and RCC have been incorporated. Chapter 6 deals with extensive simulations supplementing the theoretical studies done in chapter 4 and 5.

## 1.2 Why Restricted Canonical Correlation ?

If  $\mathbf{Y}^{(1)}$  and  $\mathbf{Y}^{(2)}$  are  $p$  and  $q$ -variate random vectors, then the first CC between them is

$$\max_{\alpha, \beta} \text{Correlation}(\alpha' \mathbf{Y}^{(1)}, \beta' \mathbf{Y}^{(2)}).$$

If  $\begin{bmatrix} Y^{(1)} \\ Y^{(2)} \end{bmatrix}$  has covariance matrix  $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ , then

$$\text{CC} = \max\{\alpha' \Sigma_{12} \beta : \alpha' \Sigma_{11} \alpha = 1 = \beta' \Sigma_{22} \beta\}$$

where  $\alpha = ((\alpha_i))$  and  $\beta = ((\beta_j))$  are the coefficients or weights of the original variables in their *representative* linear combinations. In this section, it will be explained why in many practical situations, it may be necessary to incorporate some restrictions on these coefficients.

### 1.2.1 Scholastics Studies

Suppose from the scores of several students in different tests and homework assignments in two different subjects, one is interested in studying the relationship (correlation) of students' performances in these two subjects. A simple approach is to give "reasonable" weights for the different examinations and assignments to



compute the composite scores in each subject and then compute the correlation between the composite scores. In contrast to having such predetermined weights, a CC approach would look for those weights which *maximize* the correlation between the composite scores. RCC fits somewhere in the middle. It also maximizes the correlation between possible composite scores; but instead of these weights being arbitrary as in the CC case, one may force them to be more reflective of their individual importance. For example, negative weights may be unrealistic and difficult to interpret. So the simple nonnegativity restriction on the weights seems rational. Another potential set of restrictions may be to place an ordering of the weights. For instance, if there are 3 quizzes, a composite homework assignment score, 2 midterms and a final in a subject, one may like to place the restriction

$$0 \leq \alpha_{Q1} = \alpha_{Q2} = \alpha_{Q3} \leq \alpha_H \leq \alpha_{MT1} = \alpha_{MT2} \leq \alpha_F$$

where  $\alpha_{Q_i}$ ,  $\alpha_H$ ,  $\alpha_{MT_j}$  and  $\alpha_F$  represent the weights of the  $i$ -th quiz, the homework assignment, the  $j$ -th midterm and the final respectively. Or, if the course had only several tests each of cumulative nature, then a reasonable restriction will be

$$0 \leq \alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_p$$

### 1.2.2 Study of Simultaneous Spike-trains in Neurophysiology

It is widely accepted that higher brain functions are performed by large neural networks. Also, there is a degree of uncertainty inherent in the behavior of such a system. This uncertainty is largely reflected by the stochastic variability, in the form of brief impulse-like electrical events which are known as action potentials or spikes. Because of the established evidence that the action potentials are nearly identical in shape, it is believed that they contain a limited amount of information in their shape of wave forms. Therefore, it is generally believed that the temporal patterns of the spike trains are the true information carriers in most of the vital

areas of the central nervous system (CNS). Based on intensive research during the past three decades, it has now become quite acceptable (viz., Kandel and Schwartz 1981) that the CNS consists of a number of regions (sectors), each one of which has a special portfolio of activities, and that there are mechanisms to convey inter-regional information in a conceivable manner. It is the second factor that is relevant in the present context. As a matter of fact, there are a large number of such sectors, and within each sector, there are millions of neurons. Because of this complexity, it is of natural interest to identify the nature of association or dependence (in a statistical sense) between the nerve cells belonging to the same sector or between them in different sectors which may or may not be adjacent in a physiological sense. For two or more nerve cells, the temporal patterns of the (simultaneous) spike-trains should cast light on such assignable factors, and statistically such information should be extractable. Temporal spike-train data relate to times of occurrences of the spikes with which is associated a counting process, and in view of the multitude of neurons, one ends up with a multivariate counting process. The sheer dimensionality of such a counting process poses challenging problems in formulating the statistical models and in carrying out the statistical analysis in a convenient way. CCA, possibly extended to more than two sets of variables, may play a vital role in this kind of study.

However, it is quite plausible that the picture of association may depend very much on the physiological topography of the sectors and other factors. There are millions of nerve cells in the CNS. While nerve cells in close proximities may have similar behavior, those from different sectors may have a less direct effect. Also, in practice, the use of micro-electrodes may incorporate only a finite number of such nerve cells for simultaneous recording of the related spike-trains. For this reason, the physiology of the CNS needs to be incorporated in the actual selection of such a handful of neural cells; and even so, there may be uncontrollable variation of the actual sites, even within small neighborhoods of target points.

Consider, for example, the figure 1.1. Kuperstein and Whittington (1981) developed a multi-channel microelectrode for neural spike-train recordings. Using this, it is possible to obtain simultaneous spike-trains from twenty-four sites which are arranged as indicated in the figure. Now, if one is interested in studying the dependence pattern between location A and location B, then it is sensible to perform CCA, with spike trains from each row constituting the two sets. But instead of allowing arbitrary weights to maximize correlation, it may be reasonable to place an ordering of weights according to the distances of the sites from A and B respectively:

$$\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_{12}$$

$$\beta_1 \geq \beta_2 \geq \dots \geq \beta_{12}$$

where  $\alpha_i$  and  $\beta_j$  denote the weight of the  $i$ -th site in the first row and  $j$ -th site in the second row respectively.

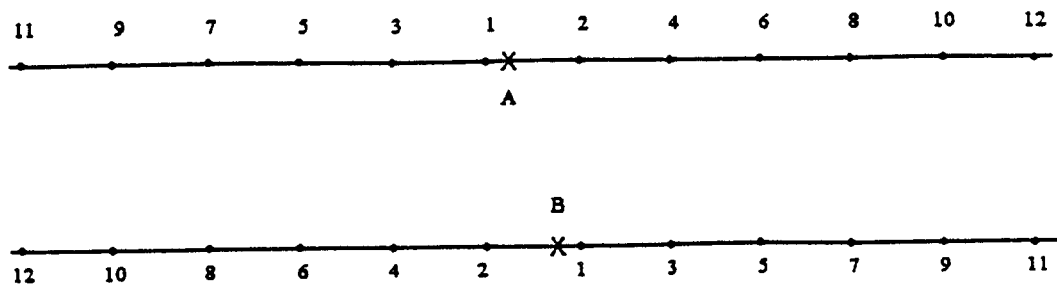


Figure 1.1: Location of sites in Kuperstein-Whittington's multichannel microelectrode

Or, if the micro-electrodes are located as shown in figure 1.2, then a pair of sensible sets of restrictions may be:

$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 \geq \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 \geq 0$$

$$\beta_1 = \beta_2 \geq \beta_3 = \beta_4 \geq \beta_5 = \beta_6 = \beta_7 \geq 0$$

Of course, these are very simplistic treatments. Physiological closeness may have little to do with the closeness in location in some cases, and it may be necessary to modify these restrictions. But this can be a good starting point for this kind of problems.

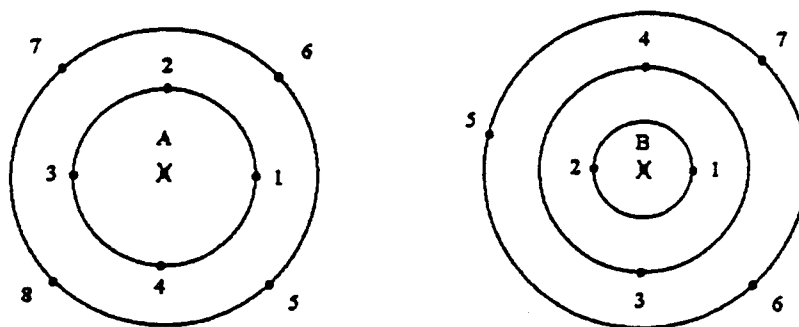


Figure 1.2: Another possible arrangement of micro-electrodes

The number of spikes occurring during a certain length of time, can reasonably be represented by a Poisson random variable. This indicates that a Multivariate Poisson (MP) model may be appropriate when several spike-trains are observed simultaneously or when a single spike-train, initiated by external stimulation, is converted into a histogram process by proper binning of the time interval. Although the MP model is not very rich, it often does reasonably satisfactory job of modeling simultaneous spike-trains. Moreover, there is no better simple parametric model for this purpose. This has been the principal motivation (besides having a case-study for multivariate discrete distribution) for simulating from MP distributions in chapter 6.

Perkel et al (1967), Habib and Sen (1985), Brillinger (1988), and Tuckwell

(1989) provide excellent review work of statistical principles in neurophysiology.

### 1.3 Different types of useful Restrictions

In the previous section some of the useful restrictions, arising in CCA, have been introduced. These are

(i)  $\alpha_i \geq 0, \beta_j \geq 0, i = 1, \dots, p, j = 1, \dots, q$ , (*nonnegativity restriction*);

(ii)  $0 \leq \alpha_1 \leq \dots \leq \alpha_p, 0 \leq \beta_1 \leq \dots \leq \beta_q$ , (*monotone restriction*);

(iii)  $0 \leq \alpha_1 = \dots = \alpha_{p_1} \leq \alpha_{p_1+1} = \dots = \alpha_{p_1+p_2} \leq \dots \leq \alpha_{p_1+\dots+p_{g-1}+1} = \dots = \alpha_{p_1+\dots+p_g}$ ,  
 $0 \leq \beta_1 = \dots = \beta_{q_1} \leq \beta_{q_1+1} = \dots = \beta_{q_1+q_2} \leq \dots \leq \beta_{q_1+\dots+q_{h-1}+1} = \dots = \beta_{q_1+\dots+q_h}$ .

(This can be called the *monotone layer restriction*, because coefficients corresponding to variables which are in successive layers, are monotone, while those corresponding to same layers are equal.)

(iv) In some situations, one may have partial restriction in the sense that additional restrictions are present only on some of the coefficients. For example, in the case of monotone restriction, if the smallest coefficients ( $\alpha_1$  and  $\beta_1$ ) are not restricted to be nonnegative, this would come in handy after using the transformation which is described in the following section. The term *partially restricted canonical correlation* may be used to denote the maximal correlation between two linear combinations of original variables, under a set of restrictions on some of the coefficients while the remaining ones may vary freely. The difference between this and the *restricted partial canonical correlation* may be noted here; the latter is the usual RCC obtained from the residual covariance matrix of two sets of variables after the linear effect of the third set has been removed (discussed in section 3.5.2). The calculation of PRCC is discussed in section 3.3.

(v) Prof. Ley of INRA-Laboratoire de Biometrie, France has encountered problems

in genetical population dynamics, where the first few coefficients in two groups should be the same; i.e.

$$\alpha_r = \beta_r \text{ for } 1 \leq r \leq p_0; \text{ where } p_0 < \min(p, q)$$

It is conceivable that there can be other types of restrictions. For example:

(vi) some coefficients are pre-assigned and are not flexible, i.e.

$$\alpha_r = \alpha_r^0; \text{ for } 1 \leq r \leq p_1 < p;$$

$$\beta_r = \beta_r^0; \text{ for } 1 \leq r \leq q_1 < q;$$

where  $\alpha_r^0$ 's and  $\beta_r^0$ 's are fixed numbers.

(vii) the ratios of the first few coefficients are equal in the two groups; i.e.

$$\alpha_1 : \alpha_2 : \dots : \alpha_{p_0} = \beta_1 : \beta_2 : \dots : \beta_{p_0}$$

where  $p_0 < \min(p, q)$ . A similar type of restriction may be:

$$\alpha_i : \beta_i = \text{constant, for } 1 \leq r \leq p_0.$$

## 1.4 Restrictions which can be reduced to Non-negativity type

Restrictions (ii) and (iii), introduced in section 1.3, can be reduced to the problem with nonnegativity restrictions by using a transformation, which is discussed in this section. In general, these type of restrictions may be characterized as *inequality restrictions*.

First consider the monotone restriction (ii). Suppose

$$\rho^* = \sup\{\alpha' \Sigma_{12} \beta : \alpha' \Sigma_{11} \alpha = 1 = \beta' \Sigma_{22} \beta; 0 \leq \alpha_1 \leq \dots \leq \alpha_p, 0 \leq \beta_1 \leq \dots \leq \beta_q\}.$$

Define

$$\begin{array}{ll}
\tilde{\alpha}_1 = \alpha_1 & \tilde{\beta}_1 = \beta_1 \\
\tilde{\alpha}_2 = \alpha_2 - \alpha_1 & \tilde{\beta}_2 = \beta_2 - \beta_1 \\
\vdots & \vdots \\
\tilde{\alpha}_p = \alpha_p - \alpha_{p-1} & \tilde{\beta}_q = \beta_q - \beta_{q-1}.
\end{array}$$

Or, equivalently,

$$\begin{array}{ll}
\alpha_1 = \tilde{\alpha}_1 & \beta_1 = \tilde{\beta}_1 \\
\alpha_2 = \tilde{\alpha}_2 + \tilde{\alpha}_1 & \beta_2 = \tilde{\beta}_2 + \tilde{\beta}_1 \\
\vdots & \vdots \\
\alpha_p = \tilde{\alpha}_p + \tilde{\alpha}_{p-1} & \beta_q = \tilde{\beta}_q - \tilde{\beta}_{q-1}.
\end{array}$$

So  $\alpha = L_p \tilde{\alpha}$  and  $\beta = L_q \tilde{\beta}$ , where  $L_r$  is a  $r \times r$  lower-triangular matrix with  $(i, j)$ -th entries being 1's for  $i \geq j$ . Thus

$$\rho^* = \sup\{\tilde{\alpha}' \tilde{\Sigma}_{12} \tilde{\beta} : \tilde{\alpha}' \tilde{\Sigma}_{11} \tilde{\alpha} = 1 = \tilde{\beta}' \tilde{\Sigma}_{22} \tilde{\beta}; \tilde{\alpha}_i \geq 0; \tilde{\beta}_j \geq 0\}$$

where  $\tilde{\Sigma}_{12} = L'_p \Sigma_{12} L_q$ ,  $\tilde{\Sigma}_{11} = L'_p \Sigma_{11} L_p$ ,  $\tilde{\Sigma}_{22} = L'_q \Sigma_{22} L_q$ . This reduces the problem to one with nonnegativity restriction. Note that, the  $(i, j)$ -th element of  $\tilde{\Sigma}_{kl}$  is nothing but the sum of all elements of  $\Sigma_{kl}$  except for those belonging to first  $(i - 1)$  rows and  $(j - 1)$  columns; i.e.

$$\tilde{\sigma}_{kl}(i, j) = \sum_{m=i}^p \sum_{n=j}^q \sigma_{kl}(m, n) \quad k, l = 1, 2.$$

Next consider the monotone layer restriction (iii), which is in fact a more general form of (ii). The transformation used in (ii), reduces this problem also to the one with nonnegativity restriction. Now one needs to work with the  $g \times h$  dimensional covariance matrix  $\check{\Sigma} = \begin{pmatrix} \check{\Sigma}_{11} & \check{\Sigma}_{12} \\ \check{\Sigma}_{21} & \check{\Sigma}_{11} \end{pmatrix}$  where the  $(i, j)$ -th element of  $\check{\Sigma}_{kl}$  is

$$\check{\sigma}_{kl}(i, j) = \tilde{\sigma}_{kl} \left( 1 + \sum_{t=1}^{i-1} p_t, 1 + \sum_{t=1}^{j-1} q_t \right).$$

## 1.5 Minimum Restricted Canonical Correlation

Traditional CC is always nonnegative. This is because if  $\text{correlation}(\alpha Y, \beta Z) = \rho < 0$  then  $\text{correlation}(-\alpha Y, \beta Z) = -\rho > 0$ . However, since  $-\alpha$  may not satisfy the restriction in a given problem, this is not true for RCC. In fact, in a restricted study, it may be more sensible to consider the corresponding minimization problem rather than the maximization one. Consider, for example, a quasi-intraclass correlation model where the covariance matrices are given by

$$\Sigma_{12} = \rho J_{p \times q}, \quad \Sigma_{11} = \rho_1 J_{p \times p} (1 - \rho_1) I_p, \quad \Sigma_{22} = \rho_2 J_{q \times q} + (1 - \rho_2) I_q.$$

where  $I_p$  denote the  $p \times p$  identity matrix and  $J_{p \times q}$  denote  $p \times q$  dimensional matrix consisting of all ones. Let  $\rho_1, \rho_2 \geq 0$ , and  $\rho < 0$ . Then in restricted CCA with  $\alpha, \beta$  nonnegative, the minimum RCC and the usual CC are related functionally.

More generally, such a situation arises if all elements of  $\Sigma_{12}$  are negative. However, the theory and calculation of finding the minimum RCC (the solution to the minimization problem), as well as its properties (sampling distribution), are very similar to those of maximum RCC. Hence for most part of this work, only the maximum RCC will be considered.



# Chapter 2

## Canonical Correlation

### 2.1 Introduction

In this chapter CCA is reviewed in some details. Emphasis has been given to parts which have direct relevance to the main study. Hence no attempts have been made to include any testing methods based on CC's. Section 2.2 deals with calculation of CC's given a population covariance matrix. In section 2.3, sample CC and its sampling distributions are considered. The topic of section 2.4 is the smoothness of the function which transforms a covariance matrix into its first CC. The discussion can easily be generalized for other CC's. Intuitively, this smoothness is what makes all relevant statistical methods work for CC. The issues associated with interpretation, usage and drawbacks of CC, as well as a comparison of CCA with other multivariate methods are discussed in section 2.5. The different generalizations of CC, including reviews of works in constrained CC's, are the topics of discussion of the last section, 2.6.

## 2.2 Population Canonical Correlations

In multivariate analysis, when one has two very large sets of variates and is interested in studying the inter-correlations, it is practical to consider only a few linear combinations from each set. Clearly, one should consider those linear combinations which contain as much dependence structure as possible. Such variables are called canonical variables and the correlations between them are called CC. The theory was first developed by Hotelling (1935, 1936).

More precisely, let  $\mathbf{Y}^{(1)}$  and  $\mathbf{Y}^{(2)}$  be  $p$  and  $q$ -variate random vectors with  $p \leq q$ . Then the  $r$ -th canonical variables are the pair of linear combinations  $U_r = (\boldsymbol{\alpha}_r)' \mathbf{Y}^{(1)}$  and  $V_r = (\boldsymbol{\beta}_r)' \mathbf{Y}^{(2)}$  each of unit variance and uncorrelated with previous  $(r - 1)$  pairs of canonical variables, where  $(U_r, V_r)$  must have maximum correlation among all linear combinations satisfying the above conditions.  $\rho_r$ , the correlation between  $U_r$  and  $V_r$ , is called the  $r$ -th CC. Suppose  $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{bmatrix}$

has covariance matrix  $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$  with  $\boldsymbol{\Sigma}_{11}$  and  $\boldsymbol{\Sigma}_{22}$  nonsingular. Let  $\text{rank}(\boldsymbol{\Sigma}_{12}) = m$ . Then one can have  $m$  non-zero (and hence positive) CC's; but often for large dimensional data the higher order C.C.'s are negligible and hence the interrelations are explained mostly by the first few canonical variables. There exist a  $p \times p$  orthogonal matrix  $L$  and a  $q \times q$  orthogonal matrix  $M$  such that  $(U_1, \dots, U_p)' = L\mathbf{Y}^{(1)}$  and  $(V_1, \dots, V_q)' = M\mathbf{Y}^{(2)}$  have correlation matrix of the form :

$$\begin{bmatrix} \mathbb{I}_p & P \\ P' & \mathbb{I}_q \end{bmatrix}$$

where  $\mathbb{I}_p$  is a  $p \times p$  identity matrix and  $P$  is a  $p \times q$  matrix of the form

$$[\text{diag}(\rho_1, \dots, \rho_m, 0, \dots, 0) | \mathbf{0}_{p \times (q-p)}].$$

Further, if  $L$  is chosen so that  $\rho_1^2 \geq \rho_2^2 \geq \dots \geq \rho_m^2$ , then  $(U_1, V_1)$  is the first pair of

canonical variable, and so on. Thus to find the first CC, one needs to maximize

$$\psi(\alpha, \beta) = \alpha' \Sigma_{12} \beta \text{ subject to } \alpha' \Sigma_{11} \alpha = 1 = \beta' \Sigma_{22} \beta, \text{ where } \alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^q. \quad (2.1)$$

This problem is usually solved by Lagrangian theory, which reduces to the equation :

$$\begin{bmatrix} -\lambda \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda \Sigma_{22} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = 0 \quad (2.2)$$

Then, the first CC,  $\rho_1$ , is the largest root of

$$\text{determinant} \begin{bmatrix} -\lambda \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda \Sigma_{22} \end{bmatrix} = 0 \quad (2.3)$$

Once  $\rho_1$  is found,  $\alpha_1$  and  $\beta_1$  are chosen to satisfy (2.2).

The first CC and coefficients can be found from (2.3) in several ways. For example, (2.2) is equivalent to  $(\Sigma \mathbf{B}^{-1} - \vartheta \mathbf{I}_n) \mathbf{B} \mathbf{x} = 0$  ; where  $\vartheta = \lambda + 1$  and  $\mathbf{B} = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$ . Hence, if  $\varsigma$  denotes the largest eigenvalue of  $\Sigma \mathbf{B}^{-1}$ , and  $\varphi$  denotes the corresponding eigenvector, then the first CC =  $\varsigma - 1$  and the canonical coefficients are given by  $\mathbf{B}^{-1} \varphi$ .

Alternatively,  $\rho_1$  is the square root of the largest eigenvalue of

$$\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \text{ or } \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$$

(the nonzero eigenvalues of these two matrices are the same.) The eigenvectors associated with this eigenvalue (there are two sets: one for  $\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$ , and one for  $\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$ ) then become the vector of the canonical coefficients. In fact, from (2.2) it follows that they can be obtained from one another by the formula

$$\alpha_1 = \frac{\Sigma_{11}^{-1} \Sigma_{12} \beta_1}{\rho_1}$$

and

$$\beta_1 = \frac{\Sigma_{22}^{-1} \Sigma_{21} \alpha_1}{\rho_1}$$

Consequently, there is no need to consider both the product matrices and one would naturally prefer the first product matrix when  $p < q$ .

To find  $\rho_r$  (for  $r \geq 2$ ) one places extra conditions on  $\alpha$  and  $\beta$ :

$$\alpha' \Sigma_{11} \alpha_i = 0 = \beta' \Sigma_{22} \beta_i \text{ for } i = 1, \dots, r - 1 \quad (2.4)$$

After some calculation it can be shown that, one ends up again with the equation (2.2). Now  $\rho_r$  is the largest root of (2.2) when there is a solution to (2.4).

## 2.3 Sample Canonical Correlation and its Sampling Distribution

In the previous section, it was observed that the population CC's are simply functions of the population covariance matrix. Hence the question of estimating CC arises naturally when  $\Sigma$  is unknown. Now, let  $\{Y_j : 1 \leq j \leq n\}$  be i.i.d.  $(p + q)$ -variate samples from the population; let  $S_n = \begin{bmatrix} S_{n:11} & S_{n:12} \\ S_{n:21} & S_{n:22} \end{bmatrix}$  denote the  $(p + q) \times (p + q)$ -dimensional sample covariance matrix based on the sample observations, i.e.

$$S_n = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y})(Y_j - \bar{Y})', \text{ where } \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$$

It is quite intuitive to use the sample covariance in defining sample CC's. In fact, under the assumption of normality of the population, maximum likelihood estimates (m.l.e.) of  $\rho$ 's and  $U, V$  can be obtained from (2.2) with the m.l.e.  $\tilde{\Sigma} = \frac{n-1}{n} S_n$  replacing  $\Sigma$ . Of course, when such an assumption is not valid, one can still use  $S_n$  as the method of moment estimator of  $\Sigma$ , and the sample CC obtained either way is the same.

It has been observed in the previous section that the squared population

CC's,  $\rho^2 (= \rho_1^2), \rho_2^2, \dots, \rho_p^2$ , are the eigenvalues of  $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1} = \Theta$  (say). Let

$$\mathbf{V}_n = \mathbf{S}_{n:12}\mathbf{S}_{n:22}^{-1}\mathbf{S}_{n:21}\mathbf{S}_{n:11}^{-1}$$

and  $r_n = r_{n:1}, r_{n:2}, \dots, r_{n:p}$  be the eigenvalues of  $\mathbf{V}_n$ , in decreasing order of magnitude. Then  $r_{n:i}$  represents the square of the  $i$ -th sample CC.

For finite sample size, the distribution of sample CC is extremely complicated. When the sample is drawn from a multivariate Normal distribution and the two groups are independent ( $\Leftrightarrow \Sigma_{12} = 0$ ), the sampling distribution is much more manageable. An elegant derivation of this can be found in Kshirsagar (1972). Of course, in such a case all population CC's are the 0's. Constantine (1963) obtained the density in terms of a hyper-geometric function of two matrix arguments, when the population CC's are different from the 0's.

Although, the sampling distribution of CC is usually studied under the assumption of multivariate normality of the observations, CCA is a meaningful study of dependence even when such assumption is not valid. It has to be kept in mind that in such a case, the canonical variables are only uncorrelated as opposed to being independent. The asymptotic distribution of the sample CC's, in a non-normal case, have been studied in Muirhead and Waternaux (1980). They proved the asymptotic joint normality of the sample CC's under the assumption that the population correlations are distinct and the fourth order moments of the population exist. The asymptotic variance of  $\sqrt{n}(r_{n:i} - \rho_i^2)$  is given by

$$4\rho_i^2(1 - \rho_i^2)^2 + \rho_i^4(\kappa_{i:4} + \kappa_{i+p:4}) + 2\rho_i^2(\rho_i^2 + 2)\kappa_{i,p+i:2,2} - 4\rho_i^3(\kappa_{i,p+i:3,1} + \kappa_{i,p+i:1,3}) \quad (2.5)$$

where  $\kappa$ 's are different fourth order cumulants of the population. The asymptotic covariances are given by even more cumbersome formulae. More importantly, since the asymptotic variance involves unknown cumulants of the population, a good estimate of the sample variance is a must for any statistical analysis (e.g. to find a confidence interval.) This calls for a resampling theory. This is reserved for chapter 5, where a detailed study of this is done along with the corresponding study for RCC.

## 2.4 Canonical Correlation as a “Smooth” Function and its Asymptotic Linearity

In this section, it is shown that CC is a “smooth” function of the covariance matrix. This in turn helps one to prove the asymptotic linearity. The purpose of this section is to establish the niceness of CC as a function, which in essence makes all relevant statistical methods “work.” The proof is by a method of existence rather than by explicit computation of the relevant functions; hence, although asymptotic normality follows as a consequence of this, one may not compute the asymptotic variance from such an approach. But this preliminary discussion is also useful for chapter 5, where the effectiveness of the jackknife and bootstrap methods for CC’s has been studied.

### 2.4.1 Canonical Correlation as a Function of Covariance matrix

First, note that, a  $(p + q) \times (p + q)$ -dimensional covariance matrix has only  $\frac{(p+q) \times (p+q+1)}{2}$  distinct elements. Hence if one considers a smooth function of this and makes a Taylor-series expansion (as in the following subsection), then it is proper and convenient to roll it out into a vector form consisting only of its distinct elements in a predetermined order and denote it by  $\hat{S}_n$  or  $\hat{\Sigma}$ . However, here,  $S_n$  and  $\hat{S}_n$  (and similarly other relevant matrices) are treated interchangeably, the particular form being obvious from the context.

Here, the focus is given on  $r_n$  or the squared first sample CC. Following exactly the same route, similar results for other eigenvalues and hence for CC’s of higher order can be obtained very easily.

Now let us look more closely at the function  $H$  which transforms the covari-

ance matrix  $\Sigma$  into the squared first CC,  $\rho^2$ . The goal is to establish that this is a "sufficiently smooth" function in the sense that it is twice differentiable and the second order partial derivatives are bounded in some neighborhood of  $\Sigma$ . To achieve this one writes  $\mathbf{H}$  as a composition of two functions ( $\mathbf{K}$  and  $\mathbf{G}$ ) which have similar properties. Let  $\mathbf{K}$  be the function which transforms  $\Sigma$  to  $\Theta$ . And let  $\mathbf{G}$  be the one which transforms  $\Theta$  to its largest eigenvalue. It is assumed here that  $\Theta$  has the unique largest eigenvalue (i.e the first population CC is unique). For a  $p \times p$  dimension positive definite (p.d.) matrix  $T$ , and scalar  $l$ , consider the function  $\mathbf{F}(T, l) = |T - l\mathbb{I}_p|$ , where  $|\cdot|$  stands for determinant and  $\mathbb{I}_p$  is the  $p$ -dimensional identity matrix. The following lemma is going to be helpful.

**Lemma 2.1** For a  $p \times p$  dimensional covariance matrix  $\mathbf{S}$ ,

$$\frac{\delta \mathbf{F}(\mathbf{S}, l)}{\delta l} = \sum_{i=1}^p |{}^i(\mathbf{S} - l\mathbb{I}_p)| \quad (2.6)$$

where  ${}^i(\mathbf{S} - l\mathbb{I}_p)$  is a  $p \times p$  matrix whose all but  $i$ -th columns are the same as that of  $(\mathbf{S} - l\mathbb{I}_p)$ , and the  $i$ -th column is  $(\underbrace{0, \dots, 0}_{i-1}, -1, 0, \dots, 0)'$ .

*Proof:* Note that the determinant of  $(\mathbf{S} - l\mathbb{I}_p)$  is the sum of  $2^p$  determinants where each column of these matrices are either from  $\mathbf{S}$  or from  $-l\mathbb{I}_p$ . The proof is by matching coefficients of  $l$ . For example, the coefficient of  $l^{p-1}$  is clearly  $p(-1)^p$  in both sides.

To match coefficients of general powers of  $l$ , there is a need to introduce some notations. For  $(i_1, \dots, i_r) \subseteq (1, \dots, p)$ , one can consider a submatrix which consists of  $(i_1, \dots, i_r)$ -th rows and columns of  $\mathbf{S}$ . Now, let all these submatrices be denoted by  $\mathbf{S}_{p-r, k}$ 's where the first index denotes the dimension of the submatrix, and the second stands for some pre-assigned ordering index. Then the coefficient of  $l^{r-1}$  in both sides is

$$r \times (-1)^r \sum_k |\mathbf{S}_{p-r, k}| \quad \text{for } r = 1, \dots, p-1. \quad \square$$

Without loss of generality, it may be assumed that  $\Theta = \text{diag}(\rho_1^2, \dots, \rho_p^2)$ .  
 Since  $\mathbf{V}_n \rightarrow \Theta$  a.s.,

$$\frac{\delta \mathbf{F}(\mathbf{V}_n, l)}{\delta l} = \sum_{i=1}^p |i(\mathbf{V}_n - \mathbb{I}_p)| \rightarrow (-1)^p \prod_{i=2}^p (\rho^2 - \rho_i^2),$$

and the last term is  $\neq 0$  (by assumption.) Hence by the implicit function theorem,  $\mathbf{G}$  is well defined in a (closed) neighborhood of  $\Theta$ .

Because  $\mathbf{F}(T, \mathbf{G}(T)) = 0$ , one gets

$$\delta \mathbf{G} / \delta t_{rs} = - \frac{\delta \mathbf{F} / \delta t_{rs}}{\delta \mathbf{F} / \delta \mathbf{G}},$$

where  $T = ((t_{rs}))$ . Since  $\mathbf{F}$  is a polynomial in both  $t_{rs}$  and  $l$ ,  $\delta \mathbf{G} / \delta t_{rs}$  is continuously differentiable (in a neighborhood where  $(\delta \mathbf{F} / \delta \mathbf{G}) \neq 0$ .) Also,

$$\begin{aligned} \delta^2 \mathbf{G} / \delta t_{rs}^2 &= - \frac{1}{(\delta \mathbf{F} / \delta \mathbf{G})^2} \{ (\delta \mathbf{F} / \delta \mathbf{G}) \times (\delta^2 \mathbf{F} / \delta^2 t_{rs}) - (\delta \mathbf{F} / \delta t_{rs}) \times (\delta^2 \mathbf{F} / \delta t_{rs} \delta \mathbf{G}) \} \\ &= \frac{(\delta \mathbf{F} / \delta t_{rs}) \times (\delta^2 \mathbf{F} / \delta t_{rs} \delta \mathbf{G})}{(\delta \mathbf{F} / \delta \mathbf{G})^2} \end{aligned}$$

Thus, by similar arguments, this is continuous (and hence in a closed neighborhood, also bounded.) Similar results hold true for  $(\delta^2 \mathbf{G} / \delta t_{rs} \delta t_{r's'})$ . This shows  $\mathbf{G}$  has bounded second derivatives in a neighborhood of  $\Theta$ . Since  $\mathbf{K}$  is a polynomial in its arguments, it is smooth up to any required degree, and hence  $\mathbf{H}$  has bounded second order partial derivatives in a neighborhood  $N_\epsilon$  of  $\Sigma$ .

Thus, the squared sample first CC  $r_n = \mathbf{G} \odot \mathbf{K}(\mathbf{S}_n) = \mathbf{H}(\mathbf{S}_n)$  is a "sufficiently" smooth function of  $\mathbf{S}_n$ , whose distinct coordinates are U-Statistics.

## 2.4.2 Asymptotic Linearity

*Definition* : A statistic  $T_n$  is *asymptotically linearizable* or has an asymptotic first order linear expansion if

$$T_n = \theta + \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{Y}_i, \theta) + o_p\left(\frac{1}{\sqrt{n}}\right) \quad (2.7)$$



where  $E_{\theta}\phi(\mathbf{Y}_i, \theta) = 0$ . A vector (or matrix) statistic is *asymptotically linearizable* if each component of it is so.

In the following, it will be shown that

$$r_n = \rho^2 + \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{Y}_i, \rho^2) + o_p\left(\frac{1}{\sqrt{n}}\right)$$

Clearly, then a similar expansion for  $\sqrt{r_n}$  holds as well, and consequently sample CC is asymptotically linear. This provides a nice representation for sample CC. In addition, one gets an alternative proof of asymptotic normality of  $r_n$ , and hence of sample CC. A higher order expansion could be considered to prove the validity of jackknifing in this context; however an alternative route has been taken in chapter 5 to do so.

First one should note that, sample covariance or variance is a U-Statistic and hence is asymptotically linearizable. The proof follows from Hoeffding decomposition. Thus,

$$S_{n:kl} = \Sigma_{kl} + \frac{1}{n} \sum_{i=1}^n \phi_{kl}(\mathbf{Y}_i, \Sigma_{kl}) + o_p\left(\frac{1}{\sqrt{n}}\right) \text{ for } k, l = 1, 2$$

Next, observe that  $S_{n:11}^{-1}$  and  $S_{n:22}^{-1}$  are also linearizable. This is because if  $S_n$  is linearizable, then so is  $S_n^{-1}$ . To see this, one notes that each element of  $S_n^{-1}$  is essentially a ratio of two determinants; and the elements of the matrices are asymptotically linearizable. The rest follows from fact that  $\frac{X_n}{Z_n}$  is asymptotically linearizable when so are  $X_n$  and  $Z_n$ , and  $Z_n \neq 0 a.s.$ . These imply (after simple multiplications) that  $V_n$  is linearizable ; i.e.

$$V_n = \Theta + \frac{1}{n} \sum_{i=1}^n \Phi(\mathbf{Y}_i, \Theta) + o_p\left(\frac{1}{\sqrt{n}}\right)$$

Hence  $r_n = \mathbf{G}(\mathbf{V}_n)$

$$\begin{aligned}
&= \mathbf{G}(\boldsymbol{\Theta}) + \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{\Phi}(\mathbf{Y}_i, \boldsymbol{\Theta}) + o_p\left(\frac{1}{\sqrt{n}}\right)\right)' \delta \mathbf{G}|_{\boldsymbol{\Theta}} \\
&\quad + \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{\Phi}(\mathbf{Y}_i, \boldsymbol{\Theta}) + o_p\left(\frac{1}{\sqrt{n}}\right)\right)' \nabla \mathbf{G}|_{\boldsymbol{\Theta}^*} \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{\Phi}(\mathbf{Y}_i, \boldsymbol{\Theta}) + o_p\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= \rho^2 + \frac{1}{n} \sum_{i=1}^n \xi(\mathbf{Y}_i, \rho^2) + o_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

where,  $\xi(\mathbf{Y}_i, \rho^2) = \boldsymbol{\Phi}(\mathbf{Y}_i, \boldsymbol{\Theta})' \delta \mathbf{G}|_{\boldsymbol{\Theta}}$ . This proves the asymptotic linearity of sample CC.

## 2.5 Interpretation, Usage, Disadvantages and Relation with other Multivariate Methods

### 2.5.1 Determining the Effective Number of Common Factors

Suppose the observed measurements (of  $Y^{(1)}$  and  $Y^{(2)}$ ) are linearly dependent on a number of common, uncorrelated unobservable factors. One can represent the structure by:

$$\begin{aligned}
Y^{(1)} &= \mathbf{A}_1 \mathbf{F} + \mathbf{G}_1 \\
Y^{(2)} &= \mathbf{A}_2 \mathbf{F} + \mathbf{G}_2
\end{aligned}$$

where  $\mathbf{F} = (F_1, \dots, F_m)'$  is the vector of common factors, and  $Cov(\mathbf{F}, \mathbf{G}_1) = \mathbf{0} = Cov(\mathbf{F}, \mathbf{G}_2) = Cov(\mathbf{G}_1, \mathbf{G}_2)$ . The goal is to determine the minimum  $m$ , for which such a representation is possible, and this  $m$  is known as the *effective number of common factors*.

It is known (for example, Rao 1974) that  $m$  is equal to the rank of the  $\Sigma_{12}$  matrix. But it is not possible to estimate  $m$  directly from  $S_{n:12}$ , since the rank of  $S_{n:12}$  is  $(p \wedge q)$  with probability 1 irrespective of the rank of  $\Sigma_{12}$ . One way of estimating  $m$  is by counting the number of sample CC's which are not too small. The intuition for this is that the number of positive (population) CC's is the same as the rank of  $\Sigma_{12}$  and one would expect the sample CC's, corresponding to zero population CC's, to be small. This problem is somewhat similar to factor analysis problems.

## 2.5.2 Prediction of Original Variables

More often CC is treated as a descriptive and exploratory tool. However, CCA is very closely related to the following prediction problem. In fact, they are so similar that often in the literature they are identified as the same problem.

Let  $Y^{(1)}$  and  $Y^{(2)}$  be as before. Consider the problem of determining the  $q$ -vector  $\nu^{(2)}$ , the  $s \times p$  matrix  $B$  and the  $q \times s$  matrix  $A$ , so that

$$Y^{(2)} - \nu^{(2)} - ABY^{(1)}$$

is small. Let us measure the size of this by

$$E\{[Y^{(2)} - \nu^{(2)} - ABY^{(1)}]'\Gamma^{-1}[Y^{(2)} - \nu^{(2)} - ABY^{(1)}]\} \quad (2.8)$$

where  $\Gamma$  is some positive definite matrix. Then the optimal choice of  $\nu^{(2)}$ ,  $B$ ,  $A$  is  $\nu_0^{(2)}$ ,  $B_0$ ,  $A_0$ , (Brillinger 1974), where

$$B_0 = [V_1 \dots V_s]'\Gamma^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}$$

$$A_0 = \Gamma^{1/2}[V_1 \dots V_s]$$

$$\nu_0^{(2)} = \mu^{(2)} - A_0B_0\mu^{(1)}$$

and the minimum error is

$$tr(\Sigma_{22}\Gamma^{-1}) - \sum_{j \leq s} \lambda_j,$$

where  $V_j$  is the  $j$ -th eigenvector of  $\Gamma^{-1/2} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Gamma^{-1/2}$ , and  $\lambda_j$  is the corresponding eigenvalue. Clearly, estimates of  $\mathbf{A}$  and  $\mathbf{B}$  can be used to predict  $Y^{(2)}$  from  $Y^{(1)}$ . One may note here, when  $\Gamma = \Sigma_{22}$ ,  $\lambda_j$  is the  $j$ -th CC. However, this problem as opposed to usual CCA, is not symmetric in the two groups.

### 2.5.3 Measure of Affine Dependence

Two random vectors  $\mathbf{U}$  and  $\tilde{\mathbf{U}}$  in the vector space  $(V, (\cdot, \cdot))$  are called affinely equivalent if  $\mathbf{U} = A\tilde{\mathbf{U}} + \mathbf{a}$  for some nonsingular linear transformation  $A$  and some vector  $\mathbf{a} \in V$ . Let us consider a measure of affine dependence between  $Y^{(1)}$  and  $Y^{(2)}$  which are functions of their mean  $\begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}$  and the covariance matrix  $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ . One expects this measure of affine dependence to be invariant of affine transformations on  $\mathbf{Y}$ . For example, if  $\tilde{Y}^{(1)}$  and  $\tilde{Y}^{(2)}$  are affinely equivalent to  $Y^{(1)}$  and  $Y^{(2)}$ , respectively, then the measures of affine dependence between  $Y^{(1)}$  and  $Y^{(2)}$  should be the same as those between  $\tilde{Y}^{(1)}$  and  $\tilde{Y}^{(2)}$ . Eaton (1983) showed that the eigenvalues of  $\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1}$  constitute the maximal invariant function under the action group generated by these affine linear transformations. Since CC's are simply the square roots (and hence 1-1 functions) of these eigenvalues, this simply means that any invariant measure of affine dependence has to be a function of the CC's.

The CC's have a natural interpretation as cosines of the angles between  $V_1 \oplus \{0\}$  and  $\{0\} \oplus V_2$  as subspaces of  $V_1 \oplus V_2$ . The geometrical interpretation for the sample CC's is discussed in Dempster (1969).

## 2.5.4 Canonical Loadings, Redundancy Coefficient

Canonical weights or the coefficients in the optimum CC's, express the importance of a variable from one set with regard to the other set in obtaining a maximum correlation between the two sets. Hence, these are comparable to multiple regression weights. However, one should not depend solely on these weights to interpret the results. A canonical loading is defined as the usual product-moment correlation between an original variable and its respective canonical variable. Hence it reflects the degree to which a variable is represented by its canonical variable. The computation of canonical loadings is fairly simple; one just needs to pre-multiply the vector of canonical coefficients by the proper intra-group correlation matrix. For example, canonical loadings for the  $j$ -th canonical variables are given by

$$\mathbf{R}_{11}\alpha_j \doteq r_{1*1}^j \text{ for the first group}$$

$$\mathbf{R}_{22}\beta_j \doteq r_{2*2}^j \text{ for the second group}$$

where  $\mathbf{R}_{ii}$ 's are the intra-group correlation matrices. These are very useful in interpretation of canonical analysis. A detailed description of canonical loadings and its usage can be found in Dillon and Goldstein (1984).

At this point, let us treat the first group of variables as the predictor set and the second group as the criterion set. The proportion of variance in the criterion set, explained by the  $j$ -th canonical variable of the second set, is given by

$$\frac{r_{2*2}^j r_{2*2}^j}{q} \doteq R_{(j)2}^2.$$

However, a more interesting problem is to find the proportion of variance in the criterion set which is explained by the predictor set. The squared CC's provide the variance shared by the canonical variates, but not the variance shared by the original predictor variables. This is where redundancy coefficients and redundancy analysis fit in. The redundancy coefficient, proposed by Stewart and Love (1968), represents the amount of variance in the criterion set that is *redundant* to the

amount of variance in the predictor set. It is computed from the information provided by the CC and  $R_{(j)2}^2$ . The computational formula for the redundancy coefficient associated with the  $j$ -th canonical variate is given by

$$R_{(j)2|1}^2 \doteq \rho_j^2 R_{(j)2}^2.$$

Stewart and Love (1968) also showed that the redundancy coefficient is equivalent to regressing each variable in the criterion set in turn on all the variables in the predictor set, and then averaging the  $q$  squared multiple correlation coefficients. Wollenberg (1977) considered redundancy analysis bypassing the CCA.

Canonical weights are used to assess the relationship between the original variables and the canonical variables. Canonical weights are analogous to beta weights in multiple regression analysis because they indicate the contribution of each variable to the respective within-set canonical variable. However, the weights do not necessarily reflect the relative importance of the different variables accurately. This may happen mainly due to the presence of multicollinearity. Thus, some variables may obtain a small, even negative, weight because of the presence of some other variables. This factor itself has received a lot of criticism from several experts. (This, at least the negative weightage problem, can be partially tackled in the RCC approach.) Also, because of multicollinearity, these coefficients become very unstable. Several authors have suggested the use of canonical loadings instead. But these also suffer from similar drawbacks. The main difficulties regarding CCA, have not been well-sorted, and one should exercise caution in interpreting results in any particular setting.

Another "drawback" of CC which is often cited in the literature, is the fact that the canonical variables, constructed by linearly combining the original variables, are artificial; they are not observable and hence very rarely one can attach physical meaning to these. Although the observation is certainly true, this criticism does not demand a lot of attention since the canonical variables are of interest in their own right. They help to deepen the understanding of the original variables and in some cases may even suggest new measures.

### **2.5.5 Canonical Correlation Analysis in Relation to other Multivariate Methods**

One of the reasons for the popularity of CC analysis is its ability to subsume so many other important multivariate methods. It brings together techniques like multiple correlation and regression analysis, canonical discriminant analysis, correspondence analysis, analysis of contingency tables and multivariate analysis of variance and covariance. Although there is some similarity between CCA and principal component analysis (PCA), the latter is an internal analysis as opposed to the former. However, some authors view CCA as a generalization of PCA and factor analysis as a more general form of CCA. But it seems that such a notion is more intuitive than formal. Multiple discriminant analysis can be shown to be a special case of canonical analysis (Takeuchi et al. 1981); in this case, canonical variables of one set turn out to be discriminators, while those of the other set provide the the optimal scores to be assigned to the different populations.

CCA has played a very important role in many problems from psychology, econometrics, anthropometry, botany and educational research. The works of Bartlett (1948), Barnett and Lewis (1963), Pearce and Holland (1960), Tintner (1952) and many others are relevant in this context.

As Kshirsagar (1972) remarked, if canonical analysis had no other use, it could at least be used as a descriptive and exploratory tool. It summarizes the complex relationships and provides a useful method of reduction in the dimensionality problem.

## **2.6 Generalized and Constrained Canonical Correlation**

### **2.6.1 Generalization to More Than Two Sets of Random Variables**

The idea of generalizing CC is due to Horst (1961). Kettenring (1971) unified several ways of extending the concept of CC for three or more sets of random variable. A more recent discussion is available in SenGupta (1983). The method calls for selection of canonical variables, one from each set, such that some function of their correlation matrix is maximized. The different methods considered are :

- 1) SUMCOR, the sum of correlation method, which maximizes the sum of the elements of the matrix.
- 2) SSQCOR, the sum of squared correlation method, which maximizes the sum of the squares of the elements of the correlation matrix.
- 3) MAXVAR, the maximum variance method, which maximizes the largest eigenvalue of the matrix.
- 4) MINVAR, the minimum variance method, which minimizes the smallest eigenvalue of the matrix.
- 5) GENVAR, the generalized variance method, which minimizes the generalized variance or determinant of the matrix.

Several computational algorithms are available for these methods. Although the theoretical studies to prove the convergence of these methods have not been done, they have been shown to work satisfactorily in practice.



## 2.6.2 When the Covariance Matrix is Singular

Previously, it has been assumed that all the relevant covariance matrices are non-singular. However, that is not necessary; the CC's obtained from any generalized inverse lead to the identical result. The work of Khatri (1976) and Rao (1981) are particularly relevant in this context. SenGupta (1991) unified various generalizations of CC's and considered the cases for appropriate singular covariance matrices. Jewell and Bloomfield (1983) showed the CC's obtained from a covariance are the same as those based on its inverse. This also has been generalized to cases when the covariance matrix is singular and different forms of generalized inverses are used. A good account of the problems related to this can be found in Baksalary et. al. (1992) and the references cited in this paper.

## 2.6.3 Constrained/restricted Canonical Correlations

During the course of this study, two papers have been found, which are somewhat similar in spirit to the present study. In both of these, the work is done only on ~~the~~ solving the optimization problems.

Desarbo et al. (1982) considered "constrained CC" for categorical data analysis. Their motivation of this study was mostly to simplify the interpretation of the canonical coefficients. They discussed response surface methodology to investigate the stability of the canonical coefficients. They found these to be remarkably unstable in the sense that there can be other pairs of linear combinations whose correlations are within "indifference or tolerance region" of the CC and yet the contributing coefficients are a lot different from the canonical coefficients. Hence they felt that the canonical coefficients should be modified/simplified for the ease of interpretation. This they achieved by allowing the coefficients to be -1, 0 or 1; and hence essentially the optimization is performed among a set of  $(3^p - 1)(3^q - 1)$  choices of the coefficients. They have also discussed computational aspect in this

case, comparing the complete enumeration with two algorithms which reduce the computation further. This type of discretization of coefficients is more applicable in categorical data. But, in general, this may have only some limited usage.

Yanai and Takane (1992) has discussed the algebra of restricted canonical correlations with additional linear constraints. To mimic their approach, slightly different notation and approach are being used in the following. Let  $\mathbf{z}$  be a  $n$ -dimensional random vector with covariance matrix proportional to the identity matrix; and let  $\mathbf{X}$  and  $\mathbf{Y}$  be centered  $n \times p$  and  $n \times q$  matrices respectively. Then to find the CC's between  $\mathbf{X}'\mathbf{z}$  and  $\mathbf{Y}'\mathbf{z}$ , one should solve any of the following equivalent equations:

$$\begin{aligned} P_{\mathbf{X}}\mathbf{Y}\beta &= \rho\mathbf{X}\alpha \text{ and } P_{\mathbf{Y}}\mathbf{X}\alpha = \rho\mathbf{Y}\beta \\ (P_{\mathbf{X}}P_{\mathbf{Y}})\mathbf{X}\alpha &= \rho^2\mathbf{X}\alpha \text{ and } P_{\mathbf{Y}}\mathbf{X}\alpha = \rho\mathbf{Y}\beta \\ (P_{\mathbf{Y}}P_{\mathbf{X}})\mathbf{Y}\beta &= \rho^2\mathbf{Y}\beta \text{ and } P_{\mathbf{X}}\mathbf{Y}\alpha = \rho\mathbf{X}\alpha. \end{aligned}$$

Where  $P_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ; with the inverse being replaced by any generalized inverse when the former does not exist. These equations show that the squared CC's are the set of nonzero (positive) eigenvalues of  $P_{\mathbf{X}}P_{\mathbf{Y}}$ .

Now consider the constrained CC problem when there are additional restrictions that:

$$\mathbf{A}'\alpha = \mathbf{0} \text{ and } \mathbf{C}'\beta = \mathbf{0}$$

where  $\mathbf{A}$  and  $\mathbf{C}$  are  $p \times r_1$  and  $q \times r_2$  dimensional matrices, respectively, with  $r_1 \leq p$  and  $r_2 \leq q$ . Then by theorem 3.1 of Yanai and Takane (1992), the solutions to the constrained problem are given by the following equations:

$$P_{\mathbf{X}\mathbf{B}}\mathbf{Y}\beta = \rho\mathbf{X}\alpha \text{ and } P_{\mathbf{Y}\mathbf{D}}\mathbf{X}\alpha = \rho\mathbf{Y}\beta \quad (2.9)$$

where  $\mathbf{B}$  and  $\mathbf{D}$  are  $p \times (p - r_1)$  and  $q \times (q - r_2)$  dimensional matrices, respectively, such that

$$\mathbf{A}'\mathbf{B} = \mathbf{0} \text{ and } \mathbf{C}'\mathbf{D} = \mathbf{0}$$

They also have other equivalent expressions for the equation (2.9).

## **Chapter 3**

# **Analysis of Restricted Canonical Correlation and Related Topics**

### **3.1 Introduction**

The calculation of restricted canonical correlations (RCC) is discussed in section 3.2. In connection with its analytic solution, a relation between RCC and CC's of different subvectors is obtained; this has played very important role in the statistical analysis of this work. Section 3.3 contains discussions about partially restricted canonical correlation, which is the maximum correlation between linear combinations of two sets of random variables when only some of the coefficients are restricted while the others are not. Naturally, it is meaningful to consider restricted versions of higher order CC's. This is attempted in section 3.4, taking two different approaches. Unfortunately, none of the methods lead to explicit solutions in general. CCA is not the only area in multivariate analysis where these kind of restrictions may be incorporated. The use of CC is particularly attractive because of its broad generality. However, restricted versions of some other multivariate methods may also be useful. A brief discussion of restricted

principal component is given in the last section of this chapter. Not surprisingly, the modifications (from traditional solutions) are very similar. Also considered are the restricted versions of part, partial and bipartial canonical correlation in section 3.5. Because of their similarity with RCC, the statistical properties of these restricted measures have not been pursued in the remaining chapters.

## 3.2 Evaluating Restricted Canonical Correlation

### 3.2.1 Calculation of Restricted Canonical Correlation using Kuhn-Tucker Theory

In this section, the goal is to discuss the algebraic solution for the RCC when the contributing coordinates are restricted to be nonnegative. Note that,

$$\text{RCC} = \sup\{\alpha' \Sigma_{12} \beta : \alpha' \Sigma_{11} \alpha = 1 = \beta' \Sigma_{22} \beta : \alpha \in \mathfrak{R}_p^+, \beta \in \mathfrak{R}_q^+\} \quad (3.1)$$

where  $\alpha = (\alpha_1, \dots, \alpha_p)'$ ,  $\beta = (\beta_1, \dots, \beta_q)'$ ; and

$\mathfrak{R}_t^+ = \{\mathbf{x} = ((x_i)) \in \mathfrak{R}_t : x_i \geq 0 \forall i = 1, \dots, t\}$ , and  $\mathfrak{R}_t$  denote the  $t$ -dimensional Euclidean space.

(3.1) can also be written as

$$\text{RCC} = \sup\{h(\mathbf{x}) = \mathbf{x}' \Sigma^* \mathbf{x} / \sqrt{\mathbf{x}' \Sigma_{11}^* \mathbf{x} \mathbf{x}' \Sigma_{22}^* \mathbf{x}} : x_i \geq 0, i = 1, \dots, l\} \quad (3.2)$$

where

$$\mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad l = p + q$$

$$\Sigma^* = \begin{bmatrix} 0 & \frac{1}{2} \Sigma_{12} \\ \frac{1}{2} \Sigma_{21} & 0 \end{bmatrix}; \quad \Sigma_{11}^* = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & 0 \end{bmatrix}; \quad \Sigma_{22}^* = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{22} \end{bmatrix}.$$

A related problem is :

$$\sup\{h(\mathbf{x}) : \mathbf{x} \in {}_0\mathfrak{R}_l^+\}, \quad (3.3)$$

where  ${}_0\mathfrak{R}_l^+ = \{\mathbf{x} \in \mathfrak{R}_l : x_i > 0\}$ .

**Lemma 3.1** *A maximum in problem (3.2) always exists.*

Proof: Let  $y^0 = \sup_{\mathbf{x} \in \mathfrak{R}_l^+} h(\mathbf{x})$ . Since  $h(\cdot)$  is scale-invariant, one also has  $y^0 = \sup_{\mathbf{x} \in \mathcal{U}_l^+} h(\mathbf{x})$ , where  $\mathcal{U}_l^+ = \{\mathbf{x} \in \mathfrak{R}_l : 0 \leq x_i \leq 1\}$ . So  $\exists \{\mathbf{x}_n\} \in \mathcal{U}_l^+$  such that  $h(\mathbf{x}_n) \rightarrow y^0$ . Since  $\mathcal{U}_l^+$  is bounded,  $\exists$  a subsequence  $\{n'\}$ , and  $\mathbf{x}^1 \in \mathcal{U}_l^+$  (since  $\mathcal{U}_l^+$  is closed) s.t.  $\mathbf{x}_{n'} \rightarrow \mathbf{x}^1$ . And, hence by continuity of  $h$ ,  $h(\mathbf{x}^1) = \lim_{n' \rightarrow \infty} h(\mathbf{x}_{n'}) = \lim_{n \rightarrow \infty} h(\mathbf{x}_n) = y^0$ .  $\square$

**Lemma 3.2** *A maximum in problem (3.3) may not exist; but if it does, then it is the same as the maximum in problem (3.1).*

Proof: The first part is true since  ${}_0\mathfrak{R}_l^+$  is an open set. The second conclusion follows because  ${}_0\mathfrak{R}_l^+$  is a dense subset of  $\mathfrak{R}_l^+$  and  $h$  is continuous.  $\square$

**Lemma 3.3** *The RCC or the maximum (supremum) in (3.1) must satisfy*

$$\left. \begin{aligned} \Sigma_{12}\beta - \rho \Sigma_{11}\alpha + \lambda^{(p)} &= 0 \\ \Sigma_{21}\alpha - \rho \Sigma_{22}\beta + \lambda^{(q)} &= 0 \end{aligned} \right\} \quad (3.4)$$

$$\alpha_i \geq 0; i = 1, \dots, p. \quad \beta_j \geq 0; j = 1, \dots, q \quad (3.5)$$

$$\lambda_i \geq 0; i = 1, \dots, (n+4) \quad (3.6)$$

$$\lambda_i \alpha_i = 0 = \lambda_{p+j} \beta_j; i = 1, \dots, p. j = 1, \dots, q \quad (3.7)$$

$$\alpha' \Sigma_{11} \alpha = 1 = \beta' \Sigma_{22} \beta \quad (3.8)$$

where  $\rho = RCC$ , and  $(\lambda^{(p)})' = (\lambda_1, \dots, \lambda_p)$ , and  $(\lambda^{(q)})' = (\lambda_{p+1}, \dots, \lambda_n)$  are the usual slack/surplus variables.

Proof: To solve (3.1), the Kuhn-Tucker Lagrangian theory is used. This states that, the optimal solution of the problem:

“Maximize  $f(\mathbf{x})$  subject to  $g_i(\mathbf{x}) \leq b_i, i = 1, \dots, m$  and  $\mathbf{x} \in \mathfrak{R}_n$ ”

must satisfy :

$$\nabla f(\mathbf{x}) - \sum_1^m \lambda_i \nabla g_i(\mathbf{x}) = 0 \quad (3.9)$$

$$g_i(\mathbf{x}) \leq b_i; i = 1, \dots, m \quad (3.10)$$

$$\lambda_i \geq 0; i = 1, \dots, m \quad (3.11)$$

$$\lambda_i [g_i(\mathbf{x}) - b_i] = 0; i = 1, \dots, m \quad (3.12)$$

This fits exactly into solving for (3.1) with

$$f(\mathbf{x}) = \mathbf{x}' \Sigma^* \mathbf{x}$$

$$n = l = p + q$$

$$m = l + 4$$

$$g_i(\mathbf{x}) = \mathbf{x}_i; i = 1, \dots, n.$$

$$g_{n+1}(\mathbf{x}) = \mathbf{x}' \Sigma_{11}^* \mathbf{x} = -g_{n+2}(\mathbf{x})$$

$$g_{n+3}(\mathbf{x}) = \mathbf{x}' \Sigma_{22}^* \mathbf{x} = -g_{n+4}(\mathbf{x})$$

$$b_i = 0; i = 1, \dots, n$$

$$b_{n+1} = b_{n+3} = 1 = -b_{n+2} = -b_{n+4}$$

Here,  $\nabla g_i(\mathbf{x}) = (\underbrace{0, \dots, 0}_{i-1}, -1, 0, \dots, 0), i = 1, \dots, n; \nabla f(\mathbf{x}) = \begin{pmatrix} \Sigma_{12} \beta \\ \Sigma_{21} \alpha \end{pmatrix};$

$\nabla g_{n+1}(\mathbf{x}) = \begin{pmatrix} 2\Sigma_{11} \alpha \\ 0 \end{pmatrix} = -\nabla g_{n+2}(\mathbf{x}); \nabla g_{n+3}(\mathbf{x}) = \begin{pmatrix} 0 \\ 2\Sigma_{22} \beta \end{pmatrix} = -\nabla g_{n+4}(\mathbf{x});$

so that the optimum solution to (3.1) must satisfy :

$$\Sigma_{12}\beta + 2(\lambda_{n+2} - \lambda_{n+1}) \Sigma_{11}\alpha + \lambda^{(p)} = 0, \quad (3.13)$$

$$\Sigma_{21}\alpha + 2(\lambda_{n+4} - \lambda_{n+3}) \Sigma_{22}\beta + \lambda^{(q)} = 0, \quad (3.14)$$

and the side restrictions (3.5) through (3.8).

Now (3.7) implies that  $\alpha'\lambda^{(p)} = 0 = \beta'\lambda^{(q)}$ . Using this and (3.8), one gets from premultiplying (3.13) and (3.14) by  $\alpha'$  and  $\beta'$  respectively,

$$2(\lambda_{n+2} - \lambda_{n+1}) = 2(\lambda_{n+4} - \lambda_{n+3}) = -\alpha'\Sigma_{12}\beta = -\rho$$

where  $\rho$  stands for the RCC. Substituting this in (3.13) and (3.14), it follows that the optimal solution to (3.1) has to satisfy (3.4), with the additional restrictions as given.  $\square$

**Lemma 3.4** *If the maximum for (3.3) exists, then it must satisfy*

$$\begin{pmatrix} -\rho\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\rho\Sigma_{22} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0 \quad (3.15)$$

**Proof:** Since the maximum in (3.3) exists, by lemma 3.2, it must satisfy (3.4). Further, since in this case  $\alpha_i > 0$ ,  $\beta_j > 0 \forall i, j$ , one obtains from (3.7) that  $\lambda_i = 0$ ,  $\forall i = 1, \dots, n$  i.e.  $\lambda^{(p)} = 0$  and  $\lambda^{(q)} = 0$ . Hence, the maximum RCC (if it exists), under the strictly positivity restriction on the coefficients, must satisfy

$$\Sigma_{12}\beta - \rho \Sigma_{11}\alpha = 0$$

$$\Sigma_{21}\alpha - \rho \Sigma_{22}\beta = 0$$

or, equivalently (3.15). It should be also noted that this is the same equation as in the case of usual CC.  $\square$

### 3.2.2 Notation for Subvector and “Proper” Submatrices

Let  $p$  and  $q$  be fixed integers as denoted previously. Also define for convenience  $N_k = \{1, 2, \dots, k\}$  for  $k$  any positive integer.

$W_k = \{a : \emptyset \neq a \subseteq N_k; \text{ with elements in } a \text{ written in natural order}\}$

$|a| = \text{cardinality of } a.$

For a  $p$ -component vector  $\mathbf{X}$ , and  $a \in W_p$ , let  ${}_a\mathbf{X}$  stand for the  $|a|$ -component vector consisting of those components of  $\mathbf{X}$  whose indices belong to  $a$ . Similarly, for a  $p \times q$  dimensional matrix  $\mathbf{S}$ ,  $a \in W_p$ ,  $b \in W_q$ , let  ${}_{a:b}\mathbf{S}$  represent the  $|a| \times |b|$ -dimensional submatrix of  $\mathbf{S}$ , consisting of those rows whose indices are in  $a$  and those columns whose indices are in  $b$ . So if

$$\mathbf{X} = (X_1, \dots, X_p)', \mathbf{S} = ((s_{ij})), \mathbf{a} = (i_1, \dots, i_l) \text{ and } \mathbf{b} = (j_1, \dots, j_k),$$

then

$${}_a\mathbf{X} = \begin{pmatrix} X_{i_1} \\ \vdots \\ X_{i_l} \end{pmatrix}; \quad {}_{a:b}\mathbf{S} = \begin{bmatrix} s_{i_1 j_1} & \dots & s_{i_1 j_k} \\ \vdots & \vdots & \vdots \\ s_{i_l j_1} & \dots & s_{i_l j_k} \end{bmatrix}$$

Throughout this work, these submatrices be referred to as “proper” submatrices, although this is not the traditional use of the terminology. Clearly, there are  $(2^p - 1) \times (2^q - 1) = \mathcal{P}\mathcal{Q}$  such “proper” submatrices of  $\mathbf{S}$ ; where  $\mathcal{P} = 2^p - 1$  and  $\mathcal{Q} = 2^q - 1$ .

### 3.2.3 Restricted Canonical Correlation is Canonical Correlation for some “Proper” Submatrix

First, note that, RCC

$$= \sup_{\substack{\alpha \in \mathcal{R}_p^+ \\ \beta \in \mathcal{R}_q^+}} \text{Correlation}(\alpha' \mathbf{Y}^{(1)}, \beta' \mathbf{Y}^{(2)}) \quad [\text{here sup can be replaced by max}]$$



$$= \max_{\substack{\mathbf{a} \in \mathcal{W}_p \\ \mathbf{b} \in \mathcal{W}_q}} \sup_{\substack{\alpha \in \mathbb{R}_+^+ \\ \beta \in \mathbb{R}_+^+}} \text{Correlation}(\alpha'_{\mathbf{a}} \mathbf{Y}^{(1)}, \beta'_{\mathbf{b}} \mathbf{Y}^{(2)}) \quad (3.16)$$

$$= \max_{\substack{\mathbf{a} \in \mathcal{W}_p \\ \mathbf{b} \in \mathcal{W}_q}} \max_{\substack{\alpha \in \mathbb{R}_+^+ \\ \beta \in \mathbb{R}_+^+}} \text{Correlation}(\alpha'_{\mathbf{a}} \mathbf{Y}^{(1)}, \beta'_{\mathbf{b}} \mathbf{Y}^{(2)}) \quad (3.17)$$

Now the supremum in (3.16) may not be attained for all  $\mathbf{a}$  and  $\mathbf{b}$  and hence cannot be replaced by maximum in general. However, when the supremum is not attained, then necessarily there exists subvectors  $\tilde{\mathbf{a}} \subseteq \mathbf{a}, \tilde{\mathbf{b}} \subseteq \mathbf{b}$  such that supremum is attained for  $\tilde{\mathbf{a}}$  and  $\tilde{\mathbf{b}}$ , and this supremum is the same as that for  $\mathbf{a}$  and  $\mathbf{b}$ . Thus it is alright to replace the supremum in (3.16) by the  $\max$  in (3.17), which stands for the maximum when it exists, while ignoring the subgroup when it does not. When the maximum exists, then by the lemma (3.4) the maximal correlation  $\rho$  has to satisfy (3.15) with the coefficients satisfying (3.5) and (3.8), where  $\Sigma_{11}, \Sigma_{12}$  and  $\Sigma_{22}$  are respectively replaced by  $\mathbf{a}:\mathbf{a}\Sigma_{11}, \mathbf{a}:\mathbf{b}\Sigma_{12}$  and  $\mathbf{b}:\mathbf{b}\Sigma_{22}$ . So for all "proper" submatrices of  $\Sigma$ , one needs to check whether there exists coefficients (corresponding to maximal correlation  $\rho$ , obtained from (3.15) ) which satisfies the side restrictions (3.5) and (3.8). It may be noted here that usually (3.5) is the more difficult condition to satisfy. However if  $\rho$  is the single root of the equation (3.15), i.e. it is an eigenvalue of single multiplicity (which is often the case with the sample covariance matrices) then this is easy to check. By (3.16), the RCC is the maximal such  $\rho$  (which has corresponding coefficients satisfying (3.5) and (3.8) ).

Since a solution of (3.15) has to be one (first or higher) of the canonical correlations, the following lemma evolves from the above discussion.

**Lemma 3.5** *If set of squared CC's corresponding to*

$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$  *is represented by*  $\text{setcc}^2(\Sigma_{11}, \Sigma_{12}, \Sigma_{22})$ , *then*

$$\text{RCC}^2(\Sigma_{11}, \Sigma_{12}, \Sigma_{22}) \in \bigcup_{\mathbf{a} \in \mathcal{W}_p, \mathbf{b} \in \mathcal{W}_q} \{\text{setcc}^2(\mathbf{a}:\mathbf{a}\Sigma_{11}, \mathbf{a}:\mathbf{b}\Sigma_{12}, \mathbf{b}:\mathbf{b}\Sigma_{22})\}.$$

Or equivalently, the squared RCC between  $Y^{(1)}$  and  $Y^{(2)}$  is equal to one of the squared CC's between  ${}_a Y^{(1)}$  and  ${}_b Y^{(2)}$  for some  $a \in W_p$ ,  $b \in W_q$ .

The proof follows from the above discussion. It is necessary to consider the squares of both RCC and CC's, because from (3.15), one can only solve for  $\rho^2$ , not  $\rho$ . This does not cause any concern for CC (since it is always nonnegative); but in the case of RCC, one needs to look at the coefficients to decide whether it should be the positive or the negative root. It is simple to characterize the situations when this added precaution is not necessary. Clearly, this is not needed when RCC is nonnegative or equivalently, at least one element of  $\Sigma_{12}$  is nonnegative.

An interesting question remains open: that is whether the squared RCC is the squared first CC of some "proper" submatrix. The answer seems to be in the affirmative based on some examples. Although, clearly the answer to this has important relevance from computational point of view, this does not affect the theoretical study for the rest of the work. In this study, this has not been assumed to be true, and the computer programs do check for all CC's for all subgroups. Proving this result seems to be too difficult. A natural first attempt is to prove that if the canonical coefficients corresponding to the first CC are not all nonnegative, then the same would be true for higher order CC's. This however is not even true.

### 3.3 Partially Restricted Canonical Correlation

In the beginning of this chapter, the calculation of RCC was done when *all* the canonical coefficients are restricted to be nonnegative. In this section, the attention is devoted to the analysis, when only *some* of the coefficients have nonnegativity constraints. This is useful in the case of some other inequality restrictions as well. One example of this would be a problem with the restriction  $\alpha_1 \leq \dots \leq \alpha_p$ ;  $\beta_1 \leq \dots \leq \beta_q$ . As opposed to the monotone restriction in section

1.3, it is not required that  $\beta_1 \geq 0$  or  $\alpha_1 \geq 0$  here. In this case, the following would come in handy because  $\tilde{\alpha}_1$  and  $\tilde{\beta}_1$  (of section 1.4) will be unrestricted in that case.

After appropriate renumbering, any partially restricted problem (with non-negativity restriction) can be transformed to the situation where the first  $p_1 (\geq 0)$  coefficients of the first set and first  $q_1 (\geq 0)$  coefficients of the second set are restricted to be nonnegative, while the remaining coefficients are unrestricted. Let  $p = p_1 + p_2$ , and  $q = q_1 + q_2$ . It can be shown following exactly the same type of calculations as in (3.4), that the optimal PRCC and the coefficients, under the nonnegativity constraints, must satisfy

$$\left. \begin{aligned} \Sigma_{12}\beta - \rho \Sigma_{11}\alpha + \begin{pmatrix} \lambda^{(p_1)} \\ 0 \end{pmatrix} &= 0 \\ \Sigma_{21}\alpha - \rho \Sigma_{22}\beta + \begin{pmatrix} \lambda^{(q_1)} \\ 0 \end{pmatrix} &= 0 \end{aligned} \right\} \quad (3.18)$$

The calculation and argument proceed as in section 3.2; but instead of taking maximum over all "proper" subsets, now the maximum is over only those subsets which contain at least the last  $p_2$  indices for the first set and the last  $q_2$  indices of the second set. More formally,

$$\text{PRCC} = \max_{\substack{\mathbf{a} \in \mathcal{W}_p^{p_1} \\ \mathbf{b} \in \mathcal{W}_q^{q_1}}} \max_{\alpha\beta}^* \text{Correlation}(\alpha'_{\mathbf{a}} Y, \beta'_{\mathbf{b}} Z) \quad (3.19)$$

where,  $\mathcal{W}_k^l = \{\mathbf{a} : N_k \setminus N_l \subseteq \mathbf{a} \subseteq N_k\}$ , and  $\max^*$  denotes the maximum (when exists) over  $\alpha, \beta$  satisfying

$$\alpha_i > 0 \forall i \in (\mathbf{a} \cap N_{p_1}) \text{ and } \beta_j > 0 \forall j \in (\mathbf{b} \cap N_{q_1}).$$

From this, PRCC can be calculated using the same formulae as before.

From (3.19) it is clear that if  $p_1 = 0 = q_1$ , then PRCC reduces to the usual CC; while if  $p_1 = p$  and  $q_1 = q \Leftrightarrow (p_2 = 0 = q_2)$ , then one gets RCC. Thus, as one would expect, PRCC is a more general form of RCC and CC.

### 3.4 Higher Order Restricted Canonical Correlation

In traditional canonical correlation analysis, the higher order CC's play a very important role, especially when some of them are not negligible compared to the first CC. The higher order CC's have been discussed in section 2.2. Naturally it is of interest to explore the extent by which the first RCC can be supplemented by its higher order analogues.

Presently, two possible approaches for such an extension will be considered. But irrespective of their differences, in general one can note some particular features about higher order (in particular second) RCC through the following examples.

**Example 3.1** This example shows that a higher order RCC may not even exist. Consider

$$\Sigma_{11} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix} = \Sigma_{22}; \quad \text{and} \quad \Sigma_{21} = \begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix}$$

For this problem, the first RCC=.5, with  $\alpha^{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \beta^{(1)}$ . Then  $\langle \alpha_1 Y_1 + \alpha_2 Y_2, \alpha^{(1)'} Y^{(1)} \rangle = (\alpha_1 \ \alpha_2) \Sigma_{11} \alpha^{(1)} = \alpha_1 + \frac{1}{2} \alpha_2$  which is 0 (under the nonnegativity restriction on  $\alpha_i$ 's) only when  $\alpha_1 = 0 = \alpha_2$ . Thus there does not exist any representative, which is orthogonal to the first canonical variable. Hence the second RCC does not exist.

Note: Of course,  $\min(p, q)$  CC's (some of them may be 0) always exist. In this case, the first CC = 1, whereas the second CC = 1/3.

Since the nonnegativity condition puts one more constraint on the coefficients, it may be tempting to think that the number of RCC's will always be one less than the number of CC's. That however, is also not the case as displayed by this following simple example.

**Example 3.2** Consider

$$\Sigma_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \Sigma_{22}; \text{ and } \Sigma_{21} = \begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix}$$

For this problem, the first RCC=.5= second RCC, with  $\alpha^{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \beta^{(1)}$ ; and  $\alpha^{(2)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \beta^{(2)}$ .

While defining higher order RCC, one has to face a dilemma, which is presented by the following interesting property of usual CC's. If a linear combination of random variables in one set is orthogonal with a canonical vector, then it is also orthogonal to its counterpart in the other group. This is because if  $\alpha^{(i)}, \beta^{(i)}$  are canonical variables (of any order) corresponding to a CC  $\rho^{(i)}$  and  $\alpha'Y^{(1)} \perp \alpha^{(i)'}Y^{(1)}$ , then

$$0 = \langle \alpha'Y^{(1)}, \alpha^{(i)'}Y^{(1)} \rangle = \alpha' \Sigma_{11} \alpha^{(i)} \quad (3.20)$$

$$- \rho^{(i)} \Sigma_{11} \alpha^{(i)} + \Sigma_{12} \beta^{(i)} = 0 \quad (3.21)$$

Hence

$$\begin{aligned} \langle \alpha'Y^{(1)}, \beta^{(i)'}Y^{(2)} \rangle &= \alpha' \Sigma_{12} \beta^{(i)} \\ &= \rho^{(i)} \alpha' \Sigma_{11} \alpha^{(i)} \quad \text{by (3.21)} \\ &= 0 \quad \text{by (3.20)} \end{aligned}$$

which proves the orthogonality between  $\alpha'Y^{(1)}$  and  $\beta^{(i)'}Y^{(2)}$ .

Evidently, one may or may not like to enforce the restriction that a canonical variable of a group of variables be orthogonal to the previous (lower order) canonical variables of the other group. It does not matter in the usual canonical analysis, because of what was just observed. However, there is a significant difference between the two approaches in the restricted case, as will be seen presently. Unfortunately, it was not possible to calculate the explicit solution under general circumstances in either approach.

Let  $\rho_{(1)}$  denote the first RCC with corresponding canonical coefficients  $\alpha_{(1)}$  and  $\beta_{(1)}$ . Recall that,  $\exists \lambda_{(1)}^{(p)}, \lambda_{(1)}^{(q)}$  which satisfy

$$\Sigma_{12}\beta_{(1)} - \rho_{(1)} \Sigma_{11}\alpha_{(1)} + \lambda_{(1)}^{(p)} = 0 \quad (3.22)$$

$$\Sigma_{21}\alpha_{(1)} - \rho_{(1)} \Sigma_{22}\beta_{(1)} + \lambda_{(1)}^{(q)} = 0 \quad (3.23)$$

with additional side restrictions (3.5) through (3.8).

### 3.4.1 Approach 1

In this approach, a higher order restricted canonical variable is orthogonal to the lower order restricted canonical variables of the same group, but not necessarily to those of the other group. Here, attempt is made to do the calculation for the second RCC (RCC2) only.

Note that the problem is the same as in (3.2), except for the added restriction that  $\alpha'Y^{(1)} \perp \alpha'_{(1)}Y^{(1)}$ , and  $\beta'Y^{(2)} \perp \beta'_{(1)}Y^{(2)}$  i.e.

$$\begin{aligned} \text{RCC2} = \sup\{ & \mathbf{x}'\Sigma^*\mathbf{x} : \mathbf{x}'\Sigma_{11}^*\mathbf{x} = 1 = \mathbf{x}'\Sigma_{22}^*\mathbf{x}; \mathbf{x}'\Sigma_{11}^*\mathbf{x}_{(1)} = 0 = \mathbf{x}'\Sigma_{22}^*\mathbf{x}_{(1)}; \\ & x_i \geq 0, i = 1, \dots, l; \} \text{ where } \mathbf{x}_{(1)} = (\alpha_{(1)}' \beta_{(1)}')'. \end{aligned}$$

(3.24)

Using Kuhn-Tucker theory as before, with  $m = p + q + 8$  instead, and

$$\begin{aligned}
g_{n+5}(\mathbf{x}) &= \mathbf{x}' \boldsymbol{\Sigma}_{11}^* \mathbf{x}_{(1)} = -g_{n+6}(\mathbf{x}) \\
g_{n+7}(\mathbf{x}) &= \mathbf{x}' \boldsymbol{\Sigma}_{22}^* \mathbf{x}_{(1)} = -g_{n+8}(\mathbf{x}) \\
b_{n+5} = b_{n+6} &= 0 = b_{n+7} = b_{n+8} \\
\nabla g_{n+5}(\mathbf{x}) &= \boldsymbol{\Sigma}_{11}^* \mathbf{x}_{(1)} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}_{(1)} \\ \mathbf{0} \end{pmatrix} = -\nabla g_{n+6}(\mathbf{x}); \\
\nabla g_{n+7}(\mathbf{x}) &= \boldsymbol{\Sigma}_{22}^* \mathbf{x}_{(1)} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\Sigma}_{22} \boldsymbol{\beta}_{(1)} \end{pmatrix} = -\nabla g_{n+8}(\mathbf{x}).
\end{aligned}$$

Thus the optimal restricted RCC2 and the corresponding restricted canonical coefficients must satisfy (in  $\rho$ ,  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ ):

$$\boldsymbol{\Sigma}_{12} \boldsymbol{\beta} + 2(\lambda_{n+2} - \lambda_{n+1}) \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} + (\lambda_{n+6} - \lambda_{n+5}) \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}_{(1)} + \boldsymbol{\lambda}^{(p)} = \mathbf{0} \quad (3.25)$$

$$\boldsymbol{\Sigma}_{21} \boldsymbol{\alpha} + 2(\lambda_{n+4} - \lambda_{n+3}) \boldsymbol{\Sigma}_{22} \boldsymbol{\beta} + (\lambda_{n+8} - \lambda_{n+7}) \boldsymbol{\Sigma}_{22} \boldsymbol{\beta}_{(1)} + \boldsymbol{\lambda}^{(q)} = \mathbf{0} \quad (3.26)$$

and the additional restrictions as before.

Multiplying (3.25) by  $\boldsymbol{\alpha}'$  and (3.26) by  $\boldsymbol{\beta}'$ , and using the additional restrictions, one obtains

$$\boldsymbol{\alpha}' \boldsymbol{\Sigma}_{12} \boldsymbol{\beta} + 2(\lambda_{n+2} - \lambda_{n+1}) = 0$$

$$\boldsymbol{\beta}' \boldsymbol{\Sigma}_{21} \boldsymbol{\alpha} + 2(\lambda_{n+4} - \lambda_{n+3}) = 0$$

So, if the second RCC is denoted by  $\rho_{(2)}$ , then  $2(\lambda_{n+2} - \lambda_{n+1}) = -\rho_{(2)} = 2(\lambda_{n+4} - \lambda_{n+3})$ . Similarly multiplying (3.25) by  $\boldsymbol{\alpha}'_{(1)}$  and (3.26) by  $\boldsymbol{\beta}'_{(1)}$ , one gets

$$\boldsymbol{\alpha}'_{(1)} \boldsymbol{\Sigma}_{12} \boldsymbol{\beta} + (\lambda_{n+6} - \lambda_{n+5}) + \boldsymbol{\alpha}'_{(1)} \boldsymbol{\lambda}^{(p)} = 0 \quad (3.27)$$

$$\boldsymbol{\beta}'_{(1)} \boldsymbol{\Sigma}_{21} \boldsymbol{\alpha} + (\lambda_{n+8} - \lambda_{n+7}) + \boldsymbol{\beta}'_{(1)} \boldsymbol{\lambda}^{(q)} = 0 \quad (3.28)$$

Now by (3.22),  $\boldsymbol{\beta}'_{(1)} \boldsymbol{\Sigma}_{21} \boldsymbol{\alpha} = \rho_{(1)} \boldsymbol{\beta}'_{(1)} \boldsymbol{\Sigma}_{22} \boldsymbol{\beta} - \boldsymbol{\beta}'_{(1)} \boldsymbol{\lambda}_{(1)}^{(q)} = -\boldsymbol{\beta}'_{(1)} \boldsymbol{\lambda}_{(1)}^{(q)}$ , and similarly  $\boldsymbol{\beta}'_{(1)} \boldsymbol{\Sigma}_{21} \boldsymbol{\alpha} = \boldsymbol{\alpha}'_{(1)} \boldsymbol{\lambda}_{(1)}^{(p)}$ . But none of the terms  $\boldsymbol{\alpha}_{(1)}' \boldsymbol{\lambda}_{(1)}^{(p)}$ ,  $\boldsymbol{\beta}'_{(1)} \boldsymbol{\lambda}_{(1)}^{(q)}$ ,  $\boldsymbol{\alpha}' \boldsymbol{\lambda}_{(1)}^{(p)}$  and  $\boldsymbol{\beta}' \boldsymbol{\lambda}_{(1)}^{(q)}$  is zero in general. However, if

$$\{i : \boldsymbol{\alpha}_{(1),i} = 0\} = \{i : \boldsymbol{\alpha}_i = 0\} \text{ and } \{j : \boldsymbol{\beta}_{(1),j} = 0\} = \{j : \boldsymbol{\beta}_j = 0\}, \quad (3.29)$$

then these terms do vanish because of (3.7). In such a case,

$$(\lambda_{n+6} - \lambda_{n+5}) = 0 = (\lambda_{n+8} - \lambda_{n+7}), \text{ because of (3.27) and (3.28),}$$

and hence by (3.25) and (3.26), the second RCC and the coefficients satisfy the same basic equations as the first RCC (with some additional constraints.) Having found the first RCC already, finding the second RCC is going to be almost trivial because one needs to look only at that “proper” subvector which supplies the first RCC. But (3.29) seems to be more of an exception than a rule, based on a few examples where the solutions can be found explicitly. Hence this is not very useful from practical point of view. However, it may also be noted that, there is little hope of finding explicit solution from (3.25, 3.26) otherwise. So the second approach is considered in the following.

### 3.4.2 Approach 2

In this approach, the canonical variables of one set is required to be orthogonal to those of the same set as well as to the lower order ones of the other set. So, for example, to find the second RCC, the maximization problem is the same as the one in (3.24), with the added restrictions

$$\alpha' \Sigma_{12} \beta_{(1)} = 0 = \alpha_{(1)}' \Sigma_{12} \beta \text{ or } \mathbf{x}' \Sigma_{12}^* \mathbf{x}_{(1)} = 0 = \mathbf{x}' \Sigma_{21}^* \mathbf{x}_{(1)},$$

where

$$\Sigma_{12}^* = \begin{bmatrix} \mathbf{0} & \Sigma_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}; \quad \Sigma_{21}^* = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \Sigma_{21} & \mathbf{0} \end{bmatrix}.$$

Using Kuhn-Tucker theory as before, with  $m = p + q + 12$  instead, and along with the previous notations, consider

$$\begin{aligned} g_{n+9}(\mathbf{x}) &= \mathbf{x}' \Sigma_{12}^* \mathbf{x}_{(1)} = -g_{n+10}(\mathbf{x}) \\ g_{n+11}(\mathbf{x}) &= \mathbf{x}' \Sigma_{21}^* \mathbf{x}_{(1)} = -g_{n+12}(\mathbf{x}) \\ b_{n+9} = b_{n+10} &= 0 = b_{n+7} = b_{n+8} \end{aligned}$$



$$\nabla g_{n+9}(\mathbf{x}) = \Sigma_{12}^* \mathbf{x}_{(1)} = \begin{pmatrix} \Sigma_{12} \boldsymbol{\beta}_{(1)} \\ 0 \end{pmatrix} = -\nabla g_{n+10}(\mathbf{x});$$

$$\nabla g_{n+11}(\mathbf{x}) = \Sigma_{21}^* \mathbf{x}_{(1)} = \begin{pmatrix} 0 \\ \Sigma_{21} \boldsymbol{\alpha}_{(1)} \end{pmatrix} = -\nabla g_{n+12}(\mathbf{x});$$

Then the optimal restricted RCC2 and the corresponding restricted canonical coefficients must satisfy (in  $\rho$ ,  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ ):

$$\left. \begin{aligned} \Sigma_{12} \boldsymbol{\beta} + 2(\lambda_{n+2} - \lambda_{n+1}) \Sigma_{11} \boldsymbol{\alpha} + (\lambda_{n+6} - \lambda_{n+5}) \Sigma_{11} \boldsymbol{\alpha}_{(1)} \\ + (\lambda_{n+10} - \lambda_{n+9}) \Sigma_{12} \boldsymbol{\beta}_{(1)} + \lambda^{(p)} = 0 \end{aligned} \right\} \quad (3.30)$$

$$\left. \begin{aligned} \Sigma_{21} \boldsymbol{\alpha} + 2(\lambda_{n+4} - \lambda_{n+3}) \Sigma_{22} \boldsymbol{\beta} + (\lambda_{n+8} - \lambda_{n+7}) \Sigma_{22} \boldsymbol{\beta}_{(1)} \\ + (\lambda_{n+12} - \lambda_{n+11}) \Sigma_{21} \boldsymbol{\alpha}_{(1)} + \lambda^{(q)} = 0, \end{aligned} \right\} \quad (3.31)$$

and the additional restrictions as before.

Multiplying (3.30) by  $\boldsymbol{\alpha}'$ , and using the additional restrictions, one gets

$$\boldsymbol{\alpha}' \Sigma_{12} \boldsymbol{\beta} + 2(\lambda_{n+2} - \lambda_{n+1}) = 0$$

Similarly, from (3.31), it follows

$$\boldsymbol{\beta}' \Sigma_{21} \boldsymbol{\alpha} + 2(\lambda_{n+4} - \lambda_{n+3}) = 0$$

Hence, as before,  $2(\lambda_{n+2} - \lambda_{n+1}) = -\rho_{(2)} = 2(\lambda_{n+4} - \lambda_{n+3})$ . But, Multiplying (3.30) by  $\boldsymbol{\alpha}'_{(1)}$ , only gives:

$$(\lambda_{n+6} - \lambda_{n+5}) + (\lambda_{n+10} - \lambda_{n+9}) \rho_{(1)} + \boldsymbol{\alpha}'_{(1)} \lambda^{(p)} = 0$$

and there does not seem to be a way out to eliminate the surplus variables. Hence, one cannot find explicit solution even when (3.29) is true and one would naturally prefer the first approach.

Unfortunately, it was not possible to find the explicit solutions under general circumstances in either approach. Of course, as noted through example 3.1, the

second RCC (unlike the first) does not exist always, and hence it is important to remember that (3.25, 3.26) and (3.30, 3.31) in simplified forms provide only necessary conditions . Because of these difficulties no attempts have been made to proceed with any further higher order RCC.

## 3.5 Other Restricted Measures

### 3.5.1 Restricted Principal Component

As opposed to canonical correlation, principal component analysis is an internal analysis. For example, to find the first principal component of a  $q$ -component random vector  $X$ , we seek the normalized linear combination of components of  $X$ , which has highest variance. That is, if  $\Sigma$  represents the covariance matrix of  $X$ , then the problem is to

$$\text{maximize } \beta' \Sigma \beta \text{ subject to } \sum_{j=1}^q \beta_j^2 = 1 \quad (3.32)$$

It can be shown that the solution is the largest eigenvalue of  $\Sigma$ . The higher order principal components are defined as those with maximal variance subject to being orthogonal to the earlier components. It turns out that these maximal variances are also eigenvalues of  $\Sigma$ .

Similar to RCC one can consider the restricted version of the problem (3.32). Next, the solution of the restricted principal component problem is discussed, when the  $\beta_j$ 's are restricted to be nonnegative. The solution is very similar to that for the corresponding RCC case. This is probably not very surprising considering the fact that many people view canonical correlation theory as a generalization of principal component theory.

According to the Kuhn-Tucker Lagrangian theory, the solution( $\rho$ ) to the

restricted problem must satisfy:

$$\Sigma\beta - \rho\beta + \lambda = 0. \quad (3.33)$$

$$\lambda_j\beta_j = 0 \quad j = 1, \dots, q. \quad (3.34)$$

$$\sum_{j=1}^q \beta_j^2 = 1, \quad \beta_j \geq 0, \quad j = 1, \dots, q. \quad (3.35)$$

Thus, following the same sequence of arguments as in RCC, the maximal variance of the principal component is the largest eigenvalue of any “proper” submatrix of  $\Sigma$  which has a corresponding eigenvector satisfying (3.35). One can study the corresponding sample version and its sampling distributions; but since this is very similar to restricted canonical analysis, it is omitted here.

### 3.5.2 Restricted and Unrestricted Part, Partial and Bipartial Canonical Correlations

Rao(1969) generalized the concept of the partial correlation coefficient to the partial canonical correlation between two sets of variables  $Y^{(1)}$  and  $Y^{(2)}$  with respect to a third set of variables  $Y^{(3)}$ . This partial canonical correlation is defined as canonical correlation between  $Y^{(1)}$  and  $Y^{(2)}$ , after the effect of  $Y^{(3)}$  is removed from both of them. Essentially, this amounts to the change that in CCA,

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \text{ is replaced by } \begin{pmatrix} \Sigma_{11.3} & \Sigma_{12.3} \\ \Sigma_{21.3} & \Sigma_{22.3} \end{pmatrix} \quad (3.36)$$

where

$$\Sigma_{ij.3} = \Sigma_{ij} - \Sigma_{i3}\Sigma_{33}^{-1}\Sigma_{3j}$$

Timm and Carlson (1976) extended this to part and bipartial canonical correlations. Both of these are natural extensions of (univariate) part and bipartial correlations and deal with conditional distributions. Unlike partial correlation,

these are not symmetric measures between the two main variables. In the case of part correlation, the effect of the third variable is removed from one of the principal variables, but not from the other. And in the case of bipartial correlation, the effect of the third variable is removed from one principal variable, while the effect of a fourth is removed from the other variable.

The extensions to CC versions of these are entirely natural and predictable. In the case of part canonical correlation, there are three groups of random variables  $Y^{(1)}$ ,  $Y^{(2)}$  and  $Y^{(3)}$ , and one is interested in finding CC between  $Y^{(1)}$  and  $Y^{(2)}$  after the effect of  $Y^{(3)}$  is removed from  $Y^{(1)}$ . In bipartial canonical correlation, there is an additional group of variables  $Y^{(4)}$ , whose effect is removed from  $Y^{(2)}$  before finding the canonical correlations. The discussion about finding these are given in details in Tim and Carlson (1976); but it again amounts to replacing  $\Sigma$  appropriately. For part canonical correlation, the appropriate replacement is

$$\begin{pmatrix} \Sigma_{11.3} & \Sigma_{12.3} \\ \Sigma_{21.3} & \Sigma_{22} \end{pmatrix} \text{ in place of } \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad (3.37)$$

And in the case of bipartial canonical correlation, one should use

$$\begin{pmatrix} \Sigma_{11.3} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \Sigma_{22.4} \end{pmatrix} \text{ in place of } \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad (3.38)$$

where  $\hat{\Sigma}_{12} = \Sigma_{12} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{32} - \Sigma_{14}\Sigma_{44}^{-1}\Sigma_{42} + \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{34}\Sigma_{44}^{-1}\Sigma_{42}$ , which is the covariance matrix between  $Y^{(1)} - \hat{Y}^{(1)}$  and  $Y^{(2)} - \hat{Y}^{(2)}$ , where  $\hat{Y}^{(1)}$  is obtained by regressing  $Y^{(1)}$  on  $Y^{(3)}$  and  $\hat{Y}^{(2)}$  is obtained by regressing  $Y^{(2)}$  on  $Y^{(4)}$ .

Naturally, one can consider the restricted versions of all these type of CC. This means, as before, that there may be some additional restrictions on the canonical coefficients of  $Y^{(1)}$  and  $Y^{(2)}$ . Since it has already been noted that the only difference (from usual CC) in the methodology to find these are in the adjustments of the covariance matrix, it is clear that to find the restricted versions (with nonnegativity constraints) of partial or part or bipartial CC's, one only needs to make appropriate adjustments as in (3.36), (3.37) and (3.38), and then follow the usual steps as in usual RCC.

# Chapter 4

## Properties and Sampling Distribution

### 4.1 Introduction

In section 4.2, some of the properties of RCC are explored. The upper and lower bounds of RCC are found; some examples are cited when these bounds are attained.

Sample restricted canonical correlation is obtained when in the calculation of section 3.2, population covariance matrix  $\Sigma$  is replaced by the sample covariance matrix  $S_n$ . The intuitive justification is the same as in the case of traditional canonical correlation; i.e. it provides the maximum likelihood estimate of population RCC when the sample is drawn from a multivariate normal population (because it does not change the calculation, if we use  $\frac{n-1}{n}S_n$  instead of  $S_n$ .) Otherwise, at least a method of moment estimator of  $\Sigma$  is being used. Of course, it would have been nice to have an alternative candidate which is an unbiased estimator of population RCC. But, as in the case of usual CC, this is not available.

Section 4.3 studies the asymptotic distribution of sample RCC. This uses the representation of RCC in terms of CC's which is also described in this section.

## 4.2 Properties of Restricted Canonical Correlation

When  $p = q = 1$ , RCC clearly reduces to the usual product-moment correlation. It may be noted here that this is not quite the case with traditional CC since, if  $Correlation(X, Y) = \rho$  then  $CC(X, Y) = |\rho|$ . Thus, in a way, RCC is a more natural generalization of correlation than CC.

### 4.2.1 Upper and Lower Bounds

In this section, for simplicity,  $\Sigma$  is taken as the correlation matrix. Also, denote the largest element of  $\Sigma_{12}$  by  $\sigma_{i_0 j_0}$ .

**Lemma 4.1**

$$\sigma_{i_0 j_0} \leq RCC \leq CC$$

**Proof:** The proof is trivial because maximum over a larger set is larger. □

**Note:** Of course, it is clear that the upper bound is valid for any form of restrictions. So also is the lower bound as long as  $(0, \dots, 0, 1, \dots, 0)$  type of coefficients are allowed in the restrictions.

Next, a brief study is done to explore when these bounds in Lemma 4.1 are attained. As usual, only the case with nonnegativity restriction is considered here. A matrix is written as  $\geq (\leq) \mathbf{0}$  if all elements of the matrix are nonnegative (nonpositive.)

**Lemma 4.2** *If either*

(a)  $\Sigma_{12} \leq \mathbf{0}$  or

(b)  $\Sigma_{11} \geq \mathbf{0}, \Sigma_{22} \geq \mathbf{0}$  and  $\sigma_{i_0 j_0}$  is the only positive element of  $\Sigma_{12}$ ,  
then RCC attains the lower bound  $\sigma_{i_0 j_0}$ .

Proof:

$$\text{correlation}(\alpha'Y, \beta'Z) = \frac{\alpha' \Sigma_{12} \beta}{\sqrt{(\alpha' \Sigma_{11} \alpha)(\beta' \Sigma_{22} \beta)}} = (\star)$$

and the numerator  $\sum_{i=1}^p \sum_{j=1}^q \alpha_i \beta_j \sigma_{ij}^{12}$ . So under (a),

$$\begin{aligned} (\star) &\leq \left( \frac{\sum \alpha_i}{\alpha' \Sigma_{11} \alpha} \right) \times \left( \frac{\sum \beta_j}{\beta' \Sigma_{22} \beta} \right) \times \sigma_{i_0 j_0}; \text{ since } \sigma_{ij}^{12} \leq \sigma_{i_0 j_0} \forall i, j \\ &\leq \sigma_{i_0 j_0}; \text{ since } (\sum \alpha_i)^2 = \sum \sum \alpha_i \alpha_j \geq \sum \sum \alpha_i \alpha_j \sigma_{ij}^{11} = \alpha' \Sigma_{11} \alpha; \\ &\quad (\sum \beta_j)^2 \geq \beta' \Sigma_{22} \beta \text{ and } \sigma_{i_0 j_0} \leq 0 \end{aligned}$$

To prove the result under (b), it has been assumed without loss of generality, that  $\alpha_{i_0} = 1 = \beta_{j_0}$ . Then

$$\begin{aligned} (\star) &\leq \frac{\sigma_{i_0 j_0}}{\sqrt{(\alpha' \Sigma_{11} \alpha)(\beta' \Sigma_{22} \beta)}}; \text{ since } \sigma_{ij}^{12} \leq 0 \text{ for } (i, j) \neq (i_0, j_0) \\ &\leq \sigma_{i_0 j_0}; \text{ since } \sigma_{i_0 j_0} \geq 0 \text{ and } \alpha' \Sigma_{11} \alpha \geq \sum \alpha_i^2 \geq \alpha_{i_0}^2 = 1 \end{aligned}$$

□

Note: It seems that if one element ( $\sigma_{i_0 j_0}$ ) of  $\Sigma_{12}$  is “significantly” larger than the others, (but  $\Sigma$  is still nonnegative definite) then  $\text{RCC} = \sigma_{i_0 j_0}$

**Lemma 4.3** Suppose all inter-block correlations are the same, say  $\rho (> 0)$ , and all intra-block correlations are also the same i.e.

$$\Sigma_{12} = \rho \mathbb{J}_{p \times q}, \quad \Sigma_{11} = \rho_1 \mathbb{J}_{p \times p} (1 - \rho_1) I_p, \quad \Sigma_{22} = \rho_2 \mathbb{J}_{q \times q} + (1 - \rho_2) I_q.$$

Then the  $RCC = \rho pqab = CC$ , where

$$a = \frac{1}{\sqrt{\rho_1 p^2 + (1 - \rho_1)p}}, \quad b = \frac{1}{\sqrt{\rho_2 q^2 + (1 - \rho_2)q}}.$$

Proof: Simple calculation shows that the CC is  $\rho pqab$  and the canonical coefficients consist of equal weight  $a$  from  $Y_i^{(1)}$ 's and  $b$  from  $Y_j^{(2)}$ 's. The rest follows by noting that both  $a$  and  $b$  are positive.  $\square$

Note 1: In this case  $minimum RCC = \rho$ . Also similar results hold when  $\rho < 0$ .

Note 2: This simple proof illustrates an important (albeit trivial) point regarding finding RCC under any type of restrictions. That is, since calculation of RCC is in general much more difficult than that of CC, it may be sensible to calculate CC, and then check if the contributing coefficients satisfy the relevant restrictions.

Note 3: Neither lemma 4.2, nor lemma 4.3 provide necessary conditions for bounds in lemma 4.1 to attain.

## 4.2.2 Invariance under Nonsingular Transformation

CC remains unchanged under any nonsingular linear transformation; but RCC does not. This can be illustrated by the following example:

**Example 4.1** Let  $p = 2 = q$ ;  $Y^{(1)}$  and  $Y^{(2)}$  have the following covariance structure:



$$\Sigma_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \Sigma_{22}; \text{ and } \Sigma_{12} = \begin{bmatrix} .4 & 0 \\ 0 & -.1 \end{bmatrix}$$

It is easy to verify that  $CC(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) = .4 = RCC(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ . Now consider the effect of multiplying  $\mathbf{Y}^{(1)}$  by  $P$  and  $\mathbf{Y}^{(2)}$  by  $Q$ , where

$$P = \mathbb{I}_2 = -Q.$$

Since both  $P$  and  $Q$  are nonsingular, the CC between  $P\mathbf{Y}^{(1)}$  and  $Q\mathbf{Y}^{(2)}$  remains the same as before. But the covariance matrix between them is  $-\Sigma_{12}$  and the RCC is only 0.1.

Next, a closer look is taken at the characteristic of the matrices  $P$  and  $Q$ , which keep RCC invariant in the sense discussed above. Mathematically speaking, the goal is to find characteristics of  $P$  and  $Q$  such that  $RCC(Y^{(1)}, Y^{(2)}) = RCC(P \cdot Y^{(1)}, Q \cdot Y^{(2)})$ . The only obvious way to ensure this is to have a  $P$  such that

$$\{\alpha \geq 0\} \iff \{P'\alpha \geq 0\} \tag{4.1}$$

(and a similar  $Q$ .) Clearly (4.1) holds if elements of  $P$  are all nonnegative. Thus, any nonsingular matrix with all nonnegative entries keeps RCC invariant.

It may be noted that RCC remains unchanged under the product of a permutation matrix. This is, of course, obvious even without the observation in the previous paragraph. Because the effect of multiplying the random variables by a permutation matrix is the same as that of relabeling, it does not change the problem or the solution. Also, since an orthonormal matrix with all nonnegative elements has to be a permutation matrix, it is the only type of orthonormal matrix, which *always* keeps RCC invariant.

## 4.3 Asymptotic Distribution of Sample Restricted Canonical Correlation

### 4.3.1 Notations and Assumptions

It follows from lemma 3.5, that the squared RCC equals the square of the first or higher CC of some proper submatrix. A  $(p' + q') \times (p' + q')$ -dimensional “proper” submatrix of  $\Sigma$  has  $\min(p', q')$  CC’s. Hence, there are

$$\mathcal{K} = \sum_{i=1}^p \sum_{j=1}^q \binom{p}{i} \binom{q}{j} (i \wedge j)$$

squared CC’s which are possible candidates of the squared RCC.

Now index the different “proper” submatrices (as defined in chapter 3) in any fixed order. As noted before, there are  $\mathcal{P}\mathcal{Q}$  such submatrices (recall,  $\mathcal{P} = 2^{p-1}$ .)

Let  $\rho_{r,s}^2$  denote the squared  $r$ -th canonical correlation of the  $s$ -th “proper” submatrix of  $\Sigma$ ; where  $s = 1, 2, \dots, \mathcal{P}\mathcal{Q}$ , and  $r = 1, \dots, \min(p', q')$ , if the  $s$ -th “proper” submatrix is of dimension  $(p' + q') \times (p' + q')$ . Similarly, let  $X_{(r,s),n}$  denote its sample version, i.e. the squared  $r$ -th canonical correlation of the  $s$ -th “proper” submatrix of  $\mathbf{S}_n$ . Although such double-indexing describes the situation more clearly, this notation is somewhat clumsy. Hence, for the sake of notational convenience,  $\rho_{r,s}^2$  and  $X_{(r,s),n}$  will be denoted by  $\rho_k^2$  and  $X_{k,n}$  from now on; and the index  $k$  runs from 1 to  $\mathcal{K}$ . It may be noted here that if RCC is the first CC of some submatrix, then this slight notational inconvenience does not arise and  $\mathcal{K} = \mathcal{P}\mathcal{Q}$ .

Accordingly the class of  $(p+q) \times (p+q)$ -dimensional covariance matrices can be partitioned into  $\mathcal{K}$  classes, and one may say that a matrix  $\mathbf{S}$  belong to the  $k$ -th class if the squared RCC of  $\mathbf{S}$  is the  $k$ -th (in the sense of ordering as described above) squared CC of  $\mathbf{S}$ .

Also let  $T_n$  and  $\theta$  denote the squared sample and population RCC respectively.

It has been shown that  $\theta = \rho_{\check{k}}^2$  for some  $\check{k}$ . The following assumption is crucial for most of the statistical properties discussed in the remaining part of this work.

**Assumption 1**

$$\check{k} \text{ is unique i.e. } \rho_k^2 \neq \theta \text{ for } k \neq \check{k} \quad (4.2)$$

Also, for the rest of the work, it will be assumed without loss of generality, that  $\check{k} = 1$ . This is only for the sake of convenience in notation.

### 4.3.2 Representing Sample Restricted Canonical Correlation in terms of Sample Canonical Correlations

Recall from Chapter 3, (replacing  $\Sigma$  by  $S_n$  in lemma 3.5) that sample squared RCC is always equal to sample squared CC for some “subvector”, i.e.

$$\forall n \text{ and } \forall \omega \in \Omega, T_n(\omega) = X_{k,n}(\omega) \text{ for some } k = k(\omega, n) \in \{1, 2, \dots, \mathcal{K}\}.$$

So if

$$A_{k,n}^0 = \{\omega : T_n(\omega) = X_{k,n}(\omega)\}, \quad k = 1, 2, \dots, \mathcal{K}; \quad (4.3)$$

then  $\bigcup_{k=1}^{\mathcal{K}} A_{k,n}^0 = \Omega$ . Further, for each  $n$  one can define a partition  $\{A_{k,n}, k = 1, 2, \dots, \mathcal{K}\}$  of  $\Omega$  by

$$A_{1,n} = A_{1,n}^0 = \{\omega : T_n(\omega) = X_{1,n}(\omega)\} \quad (4.4)$$

$$A_{k,n} = A_{k,n}^0 \setminus \bigcup_{i=1}^{k-1} A_{i,n} \quad (4.5)$$

Thus,

$$T_n = \sum_{k=1}^{\kappa} \mathbb{I}_{A_{k,n}} \times X_{k,n} \quad (4.6)$$

### 4.3.3 Preliminary Results

From Muirhead and Waternaux (1980) it follows that, if  $\rho_k$  is a distinct population CC, then

$$\sqrt{n}(X_{k,n} - \rho_k^2) \xrightarrow{\mathcal{L}} N(0, \eta_k^2) \quad (4.7)$$

where  $\eta_k^2$  is given by the appropriate asymptotic variance formula, which is of the form (2.5). Also, since  $X_{k,n}$  is a continuous function of the sample covariance matrix, for any  $k$ ,

$$X_{k,n} \xrightarrow{a.s.} \rho_k^2 \quad (4.8)$$

Also let,

$$\gamma^2 = \eta_k^2$$

The following lemma will be useful in proving the main theorem of this chapter. This will be required in chapter 5 as well.

**Lemma 4.4** *Given a sequence of events  $A_n$ , the following are equivalent*

- (i)  $\mathbb{I}_{A_n} \xrightarrow{\mathcal{L}} 0$ .
  - (ii)  $\mathbb{I}_{A_n} \xrightarrow{P} 0$ .
  - (iii)  $\mathbb{I}_{A_n} = o_p(n^\alpha)$  for some  $\alpha$ .
  - (iv)  $\mathbb{I}_{A_n} \xrightarrow{L^1} 0$  i.e  $P(A_n) \rightarrow 0$ .
- and are implied by*
- (v)  $\mathbb{I}_{A_n} \xrightarrow{a.s.} 0$ .

**Proof:** In general, (iv) and (v) each implies (ii). Also (i) and (ii) are equivalent since the limit is constant. To see that for a sequence of indicator functions

functions (ii), (iii) and (iv) are equivalent, one simply observes that for  $0 < \epsilon < 1$  and any  $\alpha$ ,

$$\{\omega : n^{-\alpha} \mathbb{I}_{A_n}(\omega) > \epsilon\} = \{\omega : \mathbb{I}_{A_n}(\omega) > \epsilon\} = \{\omega : \mathbb{I}_{A_n}(\omega) = 1\}$$

so that,  $P(n^{-\alpha} \mathbb{I}_{A_n} > \epsilon) = P(\mathbb{I}_{A_n} > \epsilon) = P(\mathbb{I}_{A_n} = 1) = P(A_n)$ .  $\square$

**Corollary 4.4.1** *This lemma remains true if the limiting constant is 1 instead of 0 ( add +1 to the R.H.S. of (iii) ).*

#### 4.3.4 Main Results

**Lemma 4.5**  $T_n \xrightarrow{a.s.} \theta$

**Proof:** Define the following function on the class of  $(p+q) \times (p+q)$ -dimensional positive definite matrices :

$$f(A) = f(A_{11}, A_{12}, A_{22}) = \sup_{x_i \geq 0, \forall i} \frac{x' A^* x}{\sqrt{x' A_{11}^* x \ x' A_{22}^* x}}$$

where

$$A^* = \begin{bmatrix} 0 & \frac{1}{2} A_{12} \\ \frac{1}{2} A_{21} & 0 \end{bmatrix}; \quad A_{11}^* = \begin{bmatrix} A_{11} & 0 \\ 0 & 0 \end{bmatrix}; \quad A_{22}^* = \begin{bmatrix} 0 & 0 \\ 0 & A_{22} \end{bmatrix}.$$

Since  $f$  is continuous and  $S_n \xrightarrow{a.s.} \Sigma$ , one gets  $f(S_n) \xrightarrow{a.s.} f(\Sigma)$ . Now to complete the proof, just note that  $\{f(S_n)\}^2 = T_n$  and  $\{f(\Sigma)\}^2 = \theta$ .  $\square$

**Lemma 4.6** *Under assumption (1),  $\mathbb{I}_{A_{1,n}} \xrightarrow{a.s.} 1$ , where  $A_{1,n}$  is defined as in (4.4).*

**Proof:** The objective is to show that  $\exists \Omega_0 \subseteq \Omega$ , such that  $\forall \omega \in \Omega_0, \mathbb{I}_{A_{1,n}}(\omega) \rightarrow 1$   
i.e.  $\forall \omega \in \Omega_0, \exists N(\omega)$  such that  $n \geq N(\omega) \Rightarrow \mathbb{I}_{A_{1,n}}(\omega) = 1$   
i.e. to show that  $\forall \omega \in \Omega_0$ ,

$$\exists N(\omega) \text{ such that } n \geq N(\omega) \Rightarrow T_n(\omega) = X_{1,n}(\omega) \quad (4.9)$$

**Claim :**  $\Omega_0 = \{\omega : T_n(\omega) \rightarrow \theta; X_{k,n} \rightarrow \rho_k^2 \forall k = 1, 2, \dots, \mathcal{K}\}$  does the job. To see this, first note that (4.8) and lemma 4.5 imply that  $P(\Omega_0) = 1$ . Fix any  $\omega \in \Omega_0$ . Suppose, if possible, (4.9) does not hold true. Since  $\forall n$  &  $\forall \omega \in \Omega_0 (\subseteq \Omega)$ ,  $T_n(\omega) = X_{k,n}(\omega)$  for some  $k = k(n, \omega) \in \{1, 2, \dots, \mathcal{K}\}$  and  $\mathcal{K}$  is finite; this would imply that  $\exists k_1 \geq 2$  and a subsequence  $\{n_l\}$  such that

$$T_{n_l}(\omega) = X_{k_1, n_l}(\omega) \quad \forall l \geq 1$$

This implies

$$\lim_{l \rightarrow \infty} T_{n_l}(\omega) = \lim_{l \rightarrow \infty} X_{k_1, n_l}(\omega) = \rho_{k_1}^2 \neq \theta$$

because of the assumption 1. But this implies that  $\lim_{n \rightarrow \infty} T_n(\omega) \neq \theta$  which is a contradiction to  $\omega \in \Omega_0$ . This completes the proof.  $\square$

**Corollary 4.6.2** Under the same assumption (1),  $\mathbb{I}_{A_{k,n}} \xrightarrow{a.s.} 0$  for  $k \geq 2$

**Theorem 4.7** Under assumption (1),  $\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} N(0, \gamma^2)$

**Proof:** From representation (4.6) it follows :

$$\sqrt{n}(T_n - \theta) = \mathbb{I}_{A_{1,n}} \times \sqrt{n}(X_{1,n} - \theta) + \sum_{k=2}^{\mathcal{K}} \mathbb{I}_{A_{k,n}} \times \sqrt{n}(X_{k,n} - \theta)$$

Since by assumption 1,  $\rho_k^2 \neq \theta$  for  $k > 1$ ,  $X_{1,n}$  corresponds to a distinct population CC, and hence by (4.7),

$$\sqrt{n}(X_{1,n} - \theta) \xrightarrow{\mathcal{L}} N(0, \eta_1^2)$$

Further, since all CC's are bounded by 1, one obtains

$$\sqrt{n}(X_{k,n} - \theta) = O_p(\sqrt{n}).$$

Thus, the result will follow via Slutsky's theorem, provided it is shown:

(a)  $\mathbb{I}_{A_{1,n}} \xrightarrow{P} 1$ .

(b)  $\mathbb{I}_{A_{k,n}} = o_p(\frac{1}{\sqrt{n}})$ .

Both (a) and (b) follows from lemma 4.6, corollary 4.6.2, lemma 4.4, and corollary 4.4.1. □

# Chapter 5

## Resampling methods in the context of Restricted and usual Canonical Correlations

### 5.1 Introduction

In chapter 2 and 4, it was observed that the asymptotic variances of sample CC and sample RCC involve unknown fourth order moments of the population. Hence, for any statistical purposes, (viz. to find a confidence interval) a good estimate of the sample variance is required. It is natural to think of the resampling methods in such a situation, which is the focus of this chapter. The relevant jackknifing method is discussed in more details here, because not only it does a good job of variance estimation, but it also provides a good estimate of the bias. In general bootstrap involves slightly more computation, although its performance is better for estimating the variance of sample CC and RCC.

Section 5.2 deals with verification of jackknifing for the usual CC. Next



in section 5.3, the same is studied when the statistic involved is the RCC. The effectiveness of the usual bootstrap for both CC and RCC is briefly discussed in section 5.4.

In this chapter, the resampling methods are investigated from the perspective of asymptotic theory. To supplement this, a detail study of their performance in finite sample size case is done in chapter 6 through simulation.

## 5.2 Jackknifing Canonical Correlation

The idea behind jackknife is as follows: given a set of  $n$  i.i.d. replicates and a statistic  $T_n$ , computing the statistic from a selection of  $(n - 1)$  observations out of possible  $n$  (i.e. deleting one observation each time) enables one to get  $n$  replicates (albeit, highly dependent) of the statistic (more strictly of  $T_{n-1}$ .) Although, Quenouille (1949) introduced jackknifing primarily for the purpose of bias reduction, now it has become more popular for estimating the standard error of a statistic. Miller (1974) provides a good review paper on jackknifing. The work of Arvesen (1969) and Sen (1977) is relevant in the context of present study.

If  $T_{n,-i}$  denotes the jackknife replicate of  $T_n$  deleting the  $i$ -th observation, then the jackknife estimate of the bias and the variance of  $T_n$  are respectively given by

$$(n - 1)\{\bar{T}_n^J - T_n\} = \hat{B}^J(T_n), \text{ and}$$

$$\frac{n - 1}{n} \sum_{i=1}^n \{T_{n,-i} - \bar{T}_n^J\}^2 = \hat{V}_n^J(T_n),$$

where  $\bar{T}_n^J = \frac{1}{n} \sum_{i=1}^n T_{n,-i}$ .  $\hat{V}_n^J(T_n)$  will simply be written as  $\hat{V}_n^J$  when the statistic is understood. Matching the notations in chapter 2, some further notations are introduced in the following.

$S_{n,-i}$ : sample covariance matrix based on  $\{Y_j : 1 \leq j \leq n, j \neq i\}$ .

$U_n = (U_n^1, \dots, U_n^t)' = (\hat{S}_n)$ .

$$\begin{aligned} \mathbf{U}_n &= (U_n^1, \dots, U_n^t)' = (\dot{\mathbf{S}}_n). \\ \mathbf{U}_{n,-i} &= (U_{n,-i}^1, \dots, U_{n,-i}^t)' = (\dot{\mathbf{S}}_{n,-i}) \text{ for } i = 1, \dots, n. \\ r_n &= \mathbf{H}(\mathbf{S}_n); r_{n,-i} = \mathbf{H}(\mathbf{S}_{n,-i}) \end{aligned}$$

Note that, each distinct element of  $\mathbf{S}_n$  is a U-Statistics with a kernel of degree two. For example, if the  $m$ -th coordinate of  $\mathbf{Y}_i$  is denoted by  $Y_{i:m}$ , for  $i = 1, \dots, n$ ;  $m = 1, \dots, l$ , then the  $(r, s)$ -th element of  $\mathbf{S}_n$  can be represented as:

$$s_{n:rs} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} f((Y_{i_1:r}, Y_{i_1:s}), (Y_{i_2:r}, Y_{i_2:s}))$$

where

$$f((x_1, y_1), (x_2, y_2)) = \frac{1}{2}(x_1 - x_2)(y_1 - y_2).$$

Hoeffding (1948) has shown that, if  $f^g$  denotes the kernel of  $U_n^g$ , then the asymptotic covariance between  $U_n^h$  and  $U_n^g$  is  $\xi_1^{hg} = 4Cov(f_1^h(Y), f_1^g(Y))$ , where  $f_1^g(Y) = \mathbf{E}f^g(Y, Z|Y)$ . Further, let  $\xi = ((\xi_1^{kl}))$ . Thus the asymptotic variance of  $\sqrt{n}r_n$  is given by  $\sigma^2 = (\delta\mathbf{H}|_{\Theta})'\xi(\delta\mathbf{H}|_{\Theta})$ . This should match with the asymptotic variance formula (2.5).

Let  $r_n^i = n \times r_n - (n-1) \times r_{n,-i}$  represent the pseudo-values (of the squared CC) for  $i = 1, \dots, n$ .

The bias-corrected jackknifed estimator of squared CC is

$$r_n^J = \frac{1}{n} \sum_{i=1}^n r_n^i = n \times r_n - \frac{n-1}{n} \sum_{i=1}^n r_{n,-i}$$

and

$$V_n^J = \frac{1}{n(n-1)} \sum_{i=1}^n (r_n^i - r_n^J)^2 = \frac{n-1}{n} \sum_{i=1}^n \left\{ r_{n,-i} - \frac{1}{n} \sum_{j=1}^n r_{n,-j} \right\}^2$$

is the jackknife estimator of the variance of the squared sample CC.

There are two main results in this section. The first one shows that the bias-corrected jackknifed estimate of the squared CC is asymptotically normal. The second one proves the strong consistency of the jackknife estimate of variance

**Lemma 5.1**

$$(n-1) \sum_{i=1}^n (U_{n,-i}^k - U_n^k)(U_{n,-i}^l - U_n^l) \xrightarrow{a.s.} \xi_1^{kl} \quad (5.1)$$

**Theorem 5.2** *The bias-corrected jackknife estimate of the squared CC is asymptotically normal with the same mean and asymptotic variance as the sample squared CC, i.e.*

$$\sqrt{n}(r_n^J - \rho^2) \xrightarrow{L} N(0, \sigma^2) \quad (5.2)$$

**Proof:**

$$\begin{aligned} r_n^i &= n\mathbf{H}(\mathbf{U}_n) - (n-1)\mathbf{H}(\mathbf{U}_{n,-i}) = \mathbf{H}(\mathbf{U}_n) \\ &\quad - (n-1)\left\{(\mathbf{U}_{n,-i} - \mathbf{U}_n)' \delta\mathbf{H}|_{\mathbf{U}_n} + \frac{1}{2}(\mathbf{U}_{n,-i} - \mathbf{U}_n)' \nabla\mathbf{H}|_{\zeta_{i,n}}(\mathbf{U}_{n,-i} - \mathbf{U}_n)\right\} \end{aligned}$$

where  $\zeta_{i,n}$  lies between (coordinatewise)  $\mathbf{U}_{n,-i}$  and  $\mathbf{U}_n$ . Since  $\frac{1}{n} \sum_{i=1}^n \mathbf{U}_{n,-i} = \mathbf{U}_n$ , we get

$$r_n^J = \mathbf{H}(\mathbf{U}_n) - \frac{n-1}{2n} \sum_{i=1}^n (\mathbf{U}_{n,-i} - \mathbf{U}_n)' \nabla\mathbf{H}|_{\zeta_{i,n}}(\mathbf{U}_{n,-i} - \mathbf{U}_n), \text{ and hence}$$

$$\sqrt{n}(r_n^J - \rho^2) = \sqrt{n}\{\mathbf{H}(\mathbf{U}_n) - \mathbf{H}(\hat{\theta})\} - \frac{n-1}{2\sqrt{n}} \sum_{i=1}^n (\mathbf{U}_{n,-i} - \mathbf{U}_n)' \nabla\mathbf{H}|_{\zeta_{i,n}}(\mathbf{U}_{n,-i} - \mathbf{U}_n)$$

Now expanding  $\mathbf{H}(\mathbf{U}_n)$  around  $\rho^2 = \mathbf{H}(\hat{\theta})$ , one gets the first term on the R.H.S

$$= \sqrt{n}(\mathbf{U}_n - \hat{\theta})' \delta\mathbf{H}|_{\zeta_n};$$

where  $\zeta_n$  lies between  $\mathbf{U}_n$  and  $\hat{\theta}$ , and hence it converges to  $\hat{\theta}$  a.s. Thus by continuity of  $\delta\mathbf{H}$ , Slutsky's theorem and asymptotic normality of multivariate U-statistic (Hoeffding 1948), it follows that the first term is asymptotically normal with mean 0 and variance  $(\delta\mathbf{H}|_{\hat{\theta}})' \xi(\delta\mathbf{H}|_{\hat{\theta}}) = \sigma^2$ .  $\nabla\mathbf{H}|_{\zeta_{i,n}}$  is bounded, and hence the second term converges to 0 by Lemma 5.1.  $\square$

**Theorem 5.3** *Jackknife estimate of variance for the squared CC is strongly consistent, i.e.*

$$n \times V_n^J \xrightarrow{a.s.} \sigma^2.$$

**Proof:** Now, a first order Taylor series expansion of  $\mathbf{H}(U_{n,-i})$  around  $U_n$  will be sufficient.

$$\begin{aligned} nV_n^J &= (n-1) \sum_{i=1}^n [\mathbf{H}(U_{n,-i}) - \frac{1}{n} \sum_{k=1}^n \mathbf{H}(U_{n,-k})]^2 \\ &= (n-1) \sum_{i=1}^n [(\mathbf{U}_{n,-i} - \mathbf{U}_n)' \delta \mathbf{H}|_{\hat{\theta}} + (\mathbf{U}_{n,-i} - \mathbf{U}_n)' \{\delta \mathbf{H}|_{\zeta_{i,n}} - \delta \mathbf{H}|_{\hat{\theta}}\}]^2 \\ &\quad - \frac{1}{n} \sum_{k=1}^n (\mathbf{U}_{n,-k} - \mathbf{U}_n)' \{\delta \mathbf{H}|_{\zeta_{k,n}} - \delta \mathbf{H}|_{\hat{\theta}}\}]^2 \\ &\quad \text{where } \zeta_{k,n} \text{ lies between } U_n \text{ and } U_{n,-k} \\ &= (n-1) \sum_{i=1}^n \underbrace{\{(\mathbf{U}_{n,-i} - \mathbf{U}_n)' \delta \mathbf{H}|_{\hat{\theta}}\}^2}_{(*)} \\ &\quad + (n-1) \underbrace{\sum_{i=1}^n [(\mathbf{U}_{n,-i} - \mathbf{U}_n)' \{\delta \mathbf{H}|_{\zeta_{i,n}} - \delta \mathbf{H}|_{\hat{\theta}}\}]^2 - \frac{1}{n} \sum_{k=1}^n (\star_k)^2}_{(**)} \\ &\quad + 2(n-1) \sum_{i=1}^n (*) \times (**). \end{aligned}$$

By Cauchy-Schwarz inequality, it is enough to show that the first term converges to  $\sigma^2$ , and the second term converges to 0. But the second term is  $\leq (n-1) \sum_{i=1}^n (\star)_i^2$  and for any given  $\epsilon > 0$ , this is

$$\leq \epsilon^2 (n-1) \sum_{i=1}^n \sum_{j=1}^t (\mathbf{U}_{n,-i}^j - \mathbf{U}_n^j)^2 \leq 2\epsilon^2 \sum_{j=1}^t \xi_1^{jj}$$

for large enough  $n$  almost surely. The first inequality holds because of continuity of  $\mathbf{H}$  and the corollary B.2.3 of appendix, as  $\zeta_{k,n}$  lies between  $U_n$  and  $U_{n,-k}$ ; and the second inequality follows from (5.1), and hence the second term converges to

to 0 a.s. The first term

$$= (\delta \mathbf{H} |_{\hat{\theta}})' \left\{ (n-1) \sum_{i=1}^n (\mathbf{U}_{n,-i} - \mathbf{U}_n)(\mathbf{U}_{n,-i} - \mathbf{U}_n)' \right\} (\delta \mathbf{H}(\hat{\theta}))$$

which converges to  $\sigma^2$ , by lemma 5.1 and Slutsky's theorem.  $\square$

### 5.3 Jackknifing Restricted Canonical Correlation

Essentially the required results follow from the corresponding results for usual canonical correlation and the representation of RCC in terms of CC's of "proper" submatrices as described in (4.6). Of course, some details need to be checked which is done in this section. Notations are as introduced in earlier chapters. Also, as before, the validation of assumption 1 is required i.e. it is assumed that the squared RCC equals the squared CC for *unique* "proper" submatrix (and without loss of generality, corresponding to the first index.) In accordance with previous section, some further notations are introduced:

$X_{k,n,-j}$  is the  $k$ -th (in the sense described in section 4.3.1) squared CC based on  $\{\mathbf{Y}_t : 1 \leq t \leq n, t \neq j\}$ ,  $k = 1, 2, \dots, \mathcal{K}$ ,  $j = 1, \dots, n$ .

$T_{n,-j}$  is the corresponding sample squared RCC.

Further, let

$$A_{k,n,-j} = \{\omega : T_{n,-j}(\omega) = X_{k,n,-j}(\omega)\}$$

except for appropriate disjointifications, exactly as in (4.3) and (4.4). Also define:

$$B_n = A_{1,n} \bigcap_{i=1}^n A_{1,n,-i}$$

The following lemma will be extremely useful in proving the theorems of this section. The proof of this follows as a corollary to the lemma 5.6; but an independent proof is also provided.

**Lemma 5.4** *Under the assumption (1),  $P(B_n) \rightarrow 1$ .*

**Proof:** It has been noted before that both RCC and CC are continuous functions of the covariance matrix. These are the relevant eigenvalues; the canonical coefficients are smooth functions of the eigenvectors, and hence these are continuous as well. This implies,

$\exists$  open set  $N(\Sigma)$  containing  $\Sigma$ , such that, if  $S \in N(\Sigma)$ ,

then squared RCC ( $S$ ) = first (in the sense of section 4.3.1) squared  $CC(S)$  (5.3)

Since, each coordinate of  $S_n$  is a U-Statistic with kernel of degree 2, and  $N(\Sigma)$  must contain an open rectangle (or product of open intervals) containing  $\Sigma$ , by lemma B.1 of appendix, it follows that

$$P(S_n \in N(\Sigma), S_{n,-i} \in N(\Sigma), \forall i = 1, \dots, n) \rightarrow 1 \quad (5.4)$$

Now, (5.3) and (5.4) clearly implies lemma 5.4. □

Just like the previous section, the goal here is two-fold. First it is shown that the bias-corrected jackknife estimate of RCC is asymptotically normal; then the strong consistency of the jackknife estimate of variance of RCC is proved.

**Theorem 5.5** *The bias-corrected jackknife estimate of the squared RCC is asymptotically normal with the same mean and asymptotic variance as usual sample squared RCC; i.e.*

$$\sqrt{n}(T_n^J - \theta) \xrightarrow{L} N(0, \gamma^2) \quad (5.5)$$

**Proof:** Recall the representation (4.6) in chapter 4

$$T_n = \sum_{k=1}^{\kappa} \mathbf{I}_{A_{k,n}} \times X_{k,n}$$

Similarly,

$$T_{n,-j} = \sum_{k=1}^{\kappa} \mathbb{I}_{A_{k,n,-j}} \times X_{k,n,-j}$$

On  $(A_{k,n} \cap_{i=1}^n A_{k_i,n,-i})$ ,  $T_n = X_{k,n}$ , and  $T_{n,-i} = X_{k_i,n,-i}$ . So

$$\begin{aligned} \sqrt{n}T_n^J &= n^{-1/2} \sum_{i=1}^n \{nT_n - (n-1)T_{n,-i}\} \\ &= \mathbb{I}_{B_n} \times n^{-1/2} \sum_{i=1}^n \{nX_{1,n} - (n-1)X_{1,n,-i}\} \\ &\quad + \mathbb{I}_{B_n^c} \times R_n \end{aligned}$$

where

$$R_n = \sum_{\substack{k, k_1, \dots, k_n \\ \text{not all 1's}}}^{\kappa} \mathbb{I}_{[A_{k,n} \cap_{i=1}^n A_{k_i,n,-i}]} \times n^{-1/2} \sum_{i=1}^n \{nX_{k,n} - (n-1)X_{k_i,n,-i}\}$$

Since CC's are bounded by 1,  $R_n = O_p(n^{5/2})$ . By lemma 5.4,  $\mathbb{I}_{B_n} = 1 - o_p(1)$ . This also implies,  $\mathbb{I}_{B_n^c} = o_p(n^{-5/2})$  by lemma 4.4. So by Slutsky's theorem, the result follows from theorem 5.2, the corresponding result for CC.  $\square$

The weak consistency of jackknife estimate of variance for RCC, follows from lemma 5.4 and theorem 5.3, as will be obvious from the proof of the next theorem. However, to prove strong consistency, one requires

**Lemma 5.6**  $\mathbb{I}_{B_n} \xrightarrow{a.s.} 1$

Or, equivalently

$$P[\liminf(B_n)] = 1.$$

**Proof:** First, observe that, by corollary B.2.3 of appendix,

$$\sup_{0 \leq j \leq n} |T_{n,-j} - \theta| \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty. \quad (5.6)$$

and for any fixed  $i = 1, \dots, \mathcal{K}$

$$\sup_{0 \leq j \leq n} |X_{i,n,-j} - \rho_i^2| \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty. \quad (5.7)$$

Now let

$$\Omega_1 = \{\omega : \sup_{0 \leq j \leq n} |T_{n,-j}(\omega) - \theta| \rightarrow 0, \sup_{0 \leq j \leq n} |X_{i,n,-j}(\omega) - \rho_i^2| \rightarrow 0, \forall i\}$$

By (5.6) and (5.7),  $P(\Omega) = 1$ . Suppose, if possible, there exists  $\omega_1 \in \Omega_1$ , such that

$$\mathbb{I}_{B_n}(\omega_1) \not\rightarrow 1 \text{ as } n \rightarrow \infty.$$

Then for each integer  $n$ ,  $\exists$  integer  $i_n : 0 \leq i_n \leq n$  and

$$T_{n,-i_n}(\omega_1) \neq X_{1,n,-i_n}(\omega_1).$$

Since  $\mathcal{K}$  is finite, this implies  $\exists$  subsequence  $\{n'\}$  and index  $k_1 > 1$ , such that

$$T_{n',-i_{n'}}(\omega_1) = X_{k_1,n',-i_{n'}}(\omega_1).$$

But since  $\omega_1 \in \Omega_1$ , the L.H.S. converges to  $\theta$  as  $n \rightarrow \infty$ , and the R.H.S. converges to  $\rho_{k_1}^2$ , which is not equal to  $\theta$ . This leads to a contradiction.  $\square$

Note: It can be verified that  $\liminf(B_n)$  is a symmetric set, and hence by Hewitt-Savage zero-one law, must have probability either 0 or 1. Unfortunately, this along with lemma 5.4, does not prove lemma 5.6, because the possibility that  $P[\mathbb{I}_{B_n} \text{ does not converge}] = 1$  cannot be ruled out. Hence, a different approach was taken in the above proof.

**Theorem 5.7** *The jackknife estimate of variance of the squared RCC is strongly consistent; i.e.*

$$n \times V_n^J \xrightarrow{a.s.} \gamma^2$$



Proof:

$$\begin{aligned} n \times V_n^J &= (n-1) \sum_{i=1}^n \left\{ T_{n,-i} - \frac{1}{n} \sum_{j=1}^n T_{n,-j} \right\}^2 \\ &= \mathbb{I}_{B_n} \times nV_n^J(X_{1,n}) + \mathbb{I}_{B_n^c} \times R_n^* \end{aligned}$$

where  $nV_n^J(X_{1,n}) = (n-1) \sum_{i=1}^n \left\{ X_{1,n,-i} - \frac{1}{n} \sum_{j=1}^n X_{1,n,-j} \right\}^2$ . And by the theorem 5.3,

$$nV_n^J(X_{1,n}) \xrightarrow{a.s.} \gamma^2 \quad (5.8)$$

Lemma 5.6 and (5.8) complete the proof.  $\square$

## 5.4 Bootstrap for Restricted and usual Canonical Correlations

Efron (1979) introduced another resampling method known as bootstrapping. Although several variations of Efron's method are now available; it is observed that the original bootstrap is sufficient for the present purpose. Given a set of sample of size  $n$ , this method computes the values of the statistic from the bootstrap resamples, which are drawn i.i.d. with replacement from the set of original sample, i.e. from the empirical distribution function  $F_n$ . This resampling can be done extensively to provide bootstrap replicates of a statistic  $T_n$ , from which one obtains the bootstrap estimates of the entire distribution of  $T_n$ . Several authors have proved the effectiveness of this algorithm under different conditions; the most relevant one for this study is that of Beran and Srivastava (1985). They showed that the bootstrapped sampling distribution of a smooth function (with continuous first derivative) of sample covariance matrix, converges to the right population distribution, i.e. if

$$J_n(F) = \mathbf{L}[\sqrt{n}\{g(\mathbf{S}_n) - g(\boldsymbol{\Sigma})\} | F]$$

and  $g$  is continuously differentiable, then

$$J_n(F_n) \rightarrow N(0, \Omega_F)$$

which is the limit of  $J_n(F)$  as  $n \rightarrow \infty$ . Hence it is reasonable to estimate the sampling distribution of  $g(S_n)$  by the appropriate bootstrapped distribution.

The required smoothness of CC as a function of sample covariance has already been established in chapter 2. Hence the effectiveness of bootstrap follows. The same should hold for RCC because of (4.6) and lemma 4.6. The advantage of bootstrap over jackknifing is that, entire bootstrapped distribution converges to the target distribution. But for bootstrap to work, there may be a heavy price to pay in terms of computation.

# Chapter 6

## Simulation Study

### 6.1 Introduction

In this chapter, the objective is to supplement the asymptotic theory discussed in chapter 4 and 5. Simulating from different MN and MP populations, attempts have been made to investigate sampling distribution CC and RCC for various sample sizes. Also, a closer look is taken at the performances of the jackknife and the bootstrap regarding estimating variance and reduction in bias.

In section 6.2, it is discussed how data has been simulated from MN and MP populations. General parameters for this simulation study are also stated here. Section 6.3 mainly consists of ten sets of figures of sampling distributions of RCC and CC for different cases. The jackknife and the bootstrap method have been compared in section 6.4 and 6.5. In section 6.4, the focus is on variance estimation of sample RCC and CC. In section 6.5, it is investigated whether the “bias-corrected” jackknife estimate has, in fact, any significantly less bias. This section is not driven by any theoretical result, since no computation of the bias of RCC or CC is available.

Computations were done on the CONVEX C240 series super computer, "Gibbs" and on the SUN workstation in the departmental computing laboratory. SAS IML was used for programming; the program codes are provided in the Appendix.

## 6.2 Generating Samples from MP and MN Distribution, and choice of simulation parameters

The definition and the characteristic of the MP distribution is discussed in Appendix A. A  $(p+q)$ -variate MP distribution is characterized by  $2^{p+q}-1$  parameters, and to retain generality, the data from such a distribution should be generated from the same number of independent univariate Poisson random variables. However, a simpler model is considered here for generating the MP random vectors by using only  $(p+q)(p+q-1)/2$  univariate Poisson variables which account for the second moments of the distribution.

For example, with  $p=2$ , and  $q=3$ , 10 independent Poisson variables  $U_{ij}; 1 \leq i < j \leq 5$  are generated. Suppose, these have means  $\sigma'_{ij}$ s. Next constructed are

$$Y_i = \sum_{j=1}^{i-1} U_{ji} + \sum_{j=i+1}^5 U_{ij}, \quad i = 1, \dots, 5$$

Then, clearly,  $Y_i$  has a (univariate) Poisson distribution with mean

$$\mu_i = \sum_{\substack{j=1 \\ j \neq i}}^5 \sigma_{ij}$$

and

$$\text{Covariance}(Y_i, Y_j) = \sigma_{ij}.$$

Jointly,  $\mathbf{Y}$  has multivariate Poisson distribution. For the first and the second covariance matrices,  $\mu_i = 1 \forall i$  and hence  $\Sigma$  represents the correlation matrix as well.

To generate samples from a MN distribution with covariance matrix  $\Sigma$ , first i.i.d. normal variates

$$\tilde{\mathbf{Y}} = (\tilde{Y}_1 : \dots : \tilde{Y}_l)'$$

are generated. Then  $\mathbf{Y} = (Y_1 : \dots : Y_l)'$ , constructed from

$$\mathbf{Y} = \Sigma^{1/2} \tilde{\mathbf{Y}}$$

provide a random sample from  $N_l(\mathbf{0}, \Sigma)$ .

Throughout the simulation work, three different population covariance matrices were dealt with. The first of them is

$$\Sigma_1 = \left[ \begin{array}{cc|ccc} 1 & 1/2 & 1/6 & 1/6 & 1/6 \\ 1/2 & 1 & 1/6 & 1/6 & 1/6 \\ \hline 1/6 & 1/6 & 1 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/3 & 1 & 1/3 \\ 1/6 & 1/6 & 1/3 & 1/3 & 1 \end{array} \right].$$

This covariance matrix provides a simple case where the population RCC and population CC are the same (here, equal to  $\frac{1}{\sqrt{15}} \approx 0.2582$ , by lemma 4.3.) Although the sample RCC is not *always* the same as the sample CC, from the asymptotic distribution theory of Chapter 4, it follows that both are asymptotically normal with the same asymptotic mean and variance. Hence, it is interesting to compare their behavior for finite sample size. The choice of  $p = 2$  and  $q = 3$  here, and in the following example, is to have a fairly simple (to speed up the simulation) yet nontrivial case where  $p \neq q$ .

The second covariance matrix is given by:

$$\Sigma_2 = \left[ \begin{array}{cc|ccc} 1 & .4 & .3 & .15 & .15 \\ .4 & 1 & 0 & .35 & .25 \\ \hline .3 & 0 & 1 & .3 & .4 \\ .15 & .35 & .3 & .1 & .2 \\ .15 & .25 & .4 & .2 & 1 \end{array} \right] .$$

This covariance matrix is such that RCC and the CC are not the same. Computing RCC and CC for this matrix gives:

$$CC \approx 0.4685 \text{ and } RCC \approx 0.3968$$

The third covariance matrix was chosen to study the distribution of RCC when assumption 1 does not hold.

$$\Sigma_3 = \left[ \begin{array}{cc|cc} 1 & 0 & .4 & 0 \\ 0 & 1 & 0 & .4 \\ \hline .4 & 0 & 1 & 0 \\ 0 & .4 & 0 & 1 \end{array} \right] .$$

Clearly for this case, both RCC and CC are equal to 0.4.

Choice of the parameters: The observations were drawn from either a multivariate Poisson distribution or a multivariate Normal distribution. The sample size ( $n_s$ ) was either 50, 100, 200, 500, 750, or 1000. For each sample of size  $n_s$ , jackknife replicates are obtained by deleting one observation each time; these replicates provide estimate of bias and variance by the formulae given in chapter 5. Also, for each simulation, i.i.d. bootstrap replicates of size ( $n_B$ ) 1000 were obtained from the original sample. From these replicates, the estimates are computed. For each simulation, the bootstrap estimate of variance is the sample variance of the bootstrap replicates. And the bootstrap estimate of bias is the difference between the average of bootstrap replicates and the original estimate. The

final answers (given in tables) are the averages over different simulations. These simulations were performed under each of the 3 covariance matrices introduced above. For convenience, these are taken to be correlation matrices.

### 6.3 Finite Sample Distribution

The goal here is to have data-driven validation of theorem 4.7. The important issue is when the “asymptotics kick in”. The corresponding distributions for CC are included as well. Figures in the next ten pages show the simulated sampling distributions of RCC and CC for different sample sizes when the samples are obtained from either MP or MN population. The label  $n_s$  and  $n_{sim}$  in the figures represent the sample size and the simulation size respectively.

The pictures with simulation size 1000 (as opposed to 5000) are understandably more bumpy. Also, since the graphs are drawn after binning the data (bin width=0.01), the “bin effect” is visible in some cases. But generally, the data supports the asymptotic result that was obtained in Chapter 4. The degree of non-normality is substantial for some small sample sizes, but the asymptotic normality effect takes over for sample sizes 500 (conservatively) and higher. For the third covariance matrix, figures are provided only for sample sizes 50, 100, 200, and 400 and MN distribution. From these, it appears that the above observations remain true for the third covariance matrix. This indicates that possibly the assumption 1 is not necessary for the asymptotic normality of RCC. This is not very surprising, because for this covariance, population RCC and CC are the same.

figure 6.1  
 Sampling distribution  
 of RCC  
 MP population  
 first covariance matrix

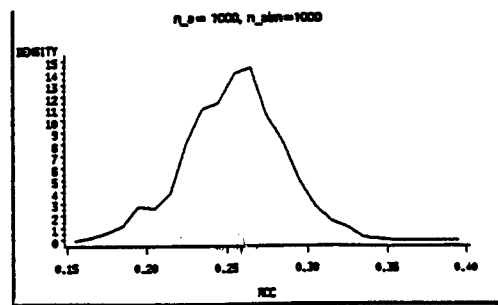
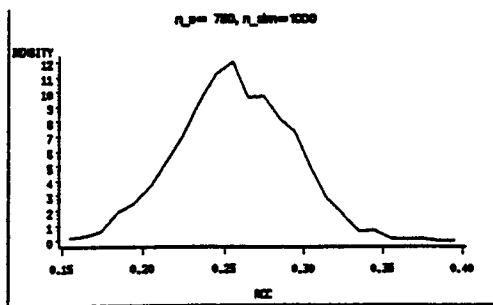
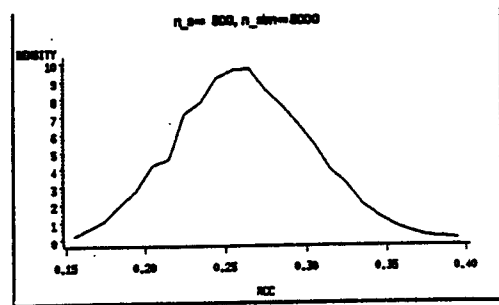
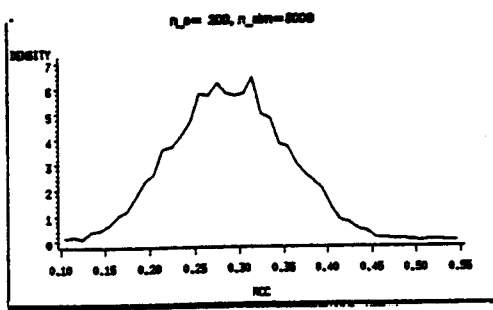
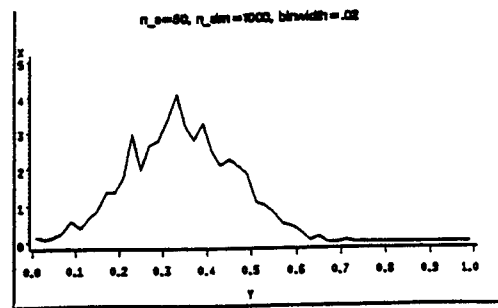




figure 6.2  
 Sampling distribution  
 of CC  
 MP population  
 first covariance matrix

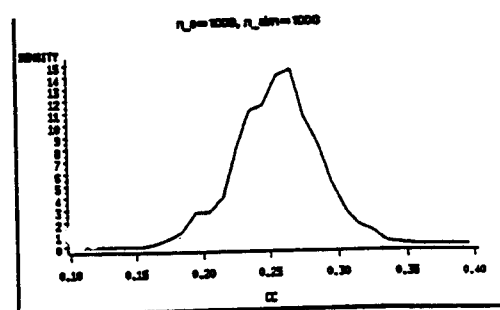
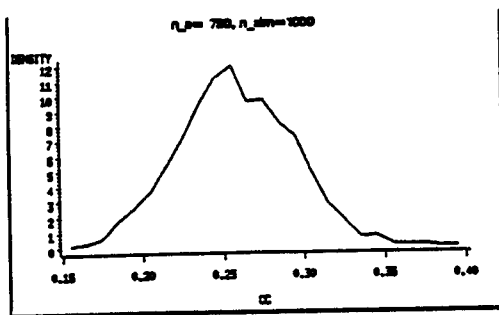
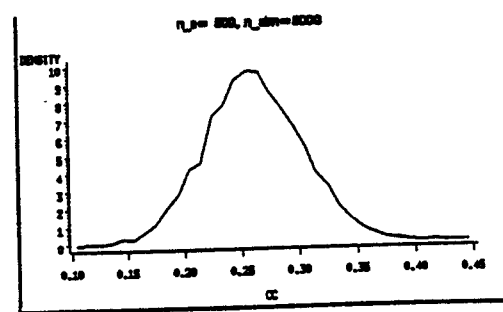
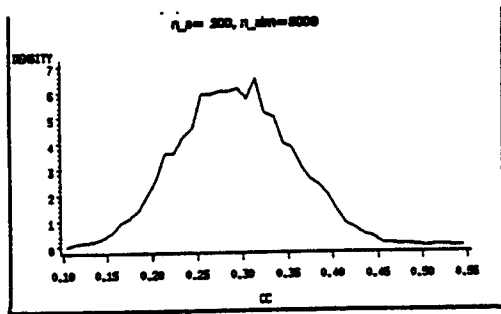
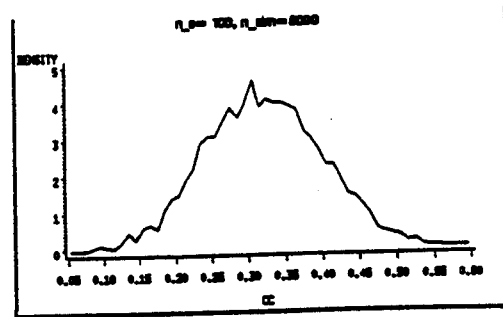


figure 6.3  
 Sampling distribution  
 of RCC  
 MN population  
 first covariance matrix

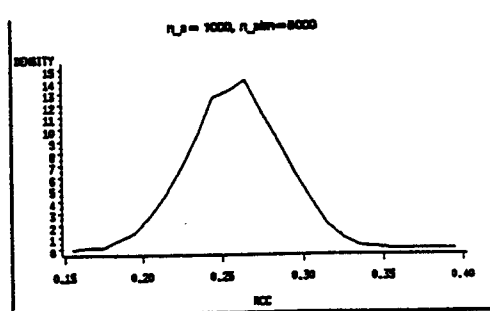
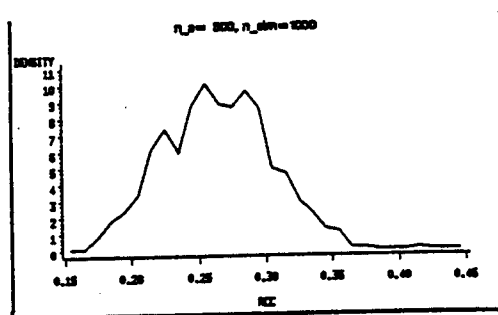
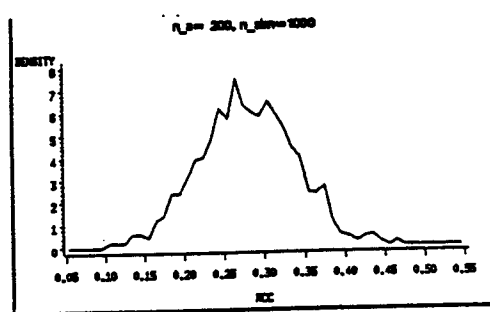
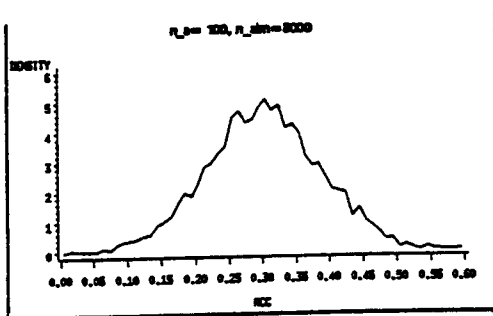
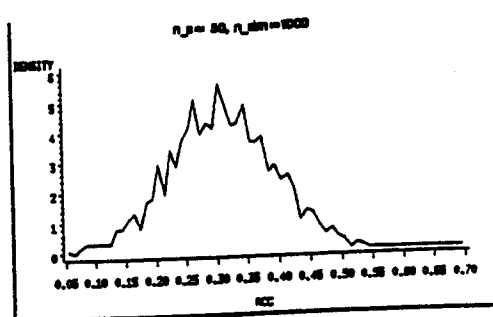


figure 6.4  
 Sampling distribution  
 of CC  
 MN population  
 first covariance matrix

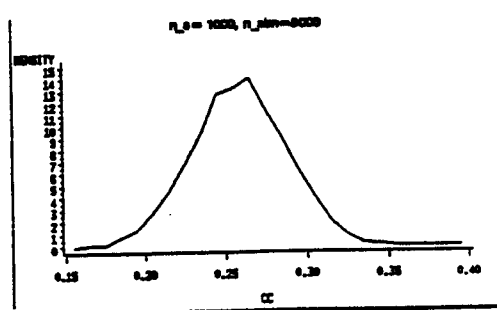
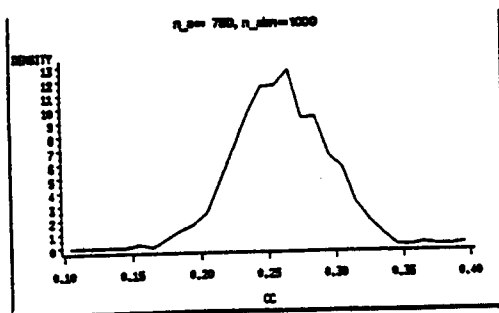
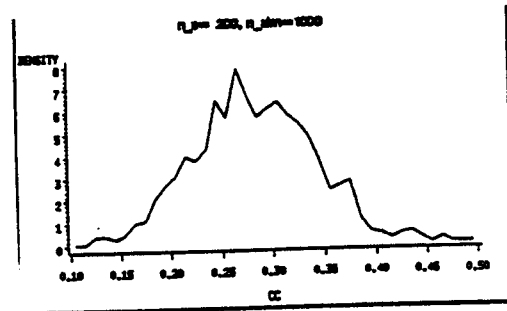
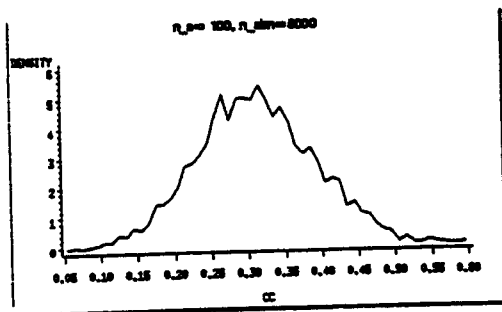
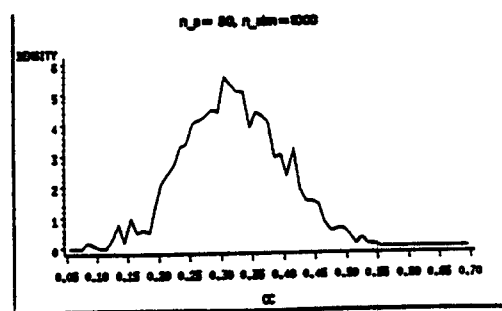
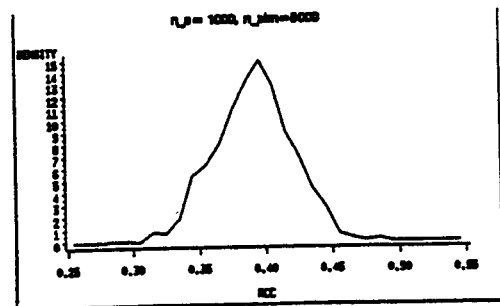
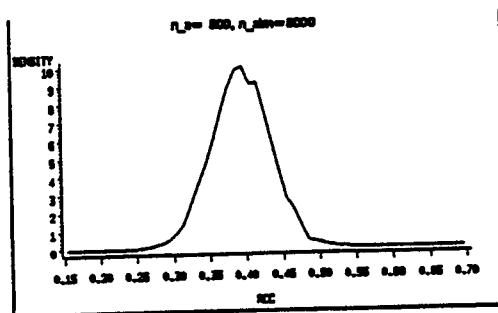
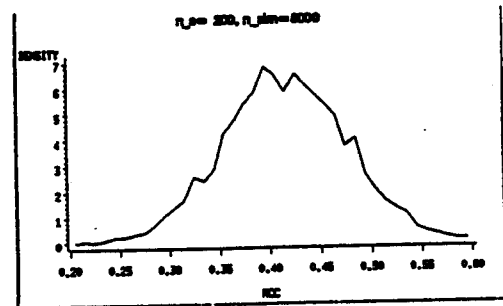
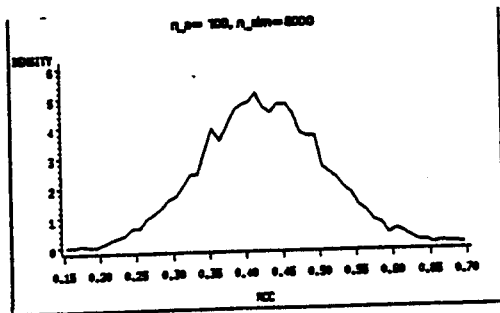
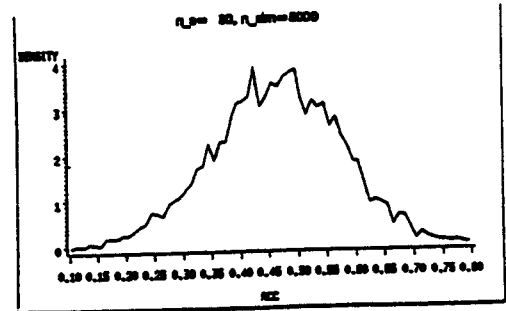


figure 6.5  
 Sampling distribution  
 of RCC  
 MP population  
 second covariance matrix



**figure 6.6**  
**Sampling distribution**  
**of CC**  
**MP population**  
**second covariance matrix**

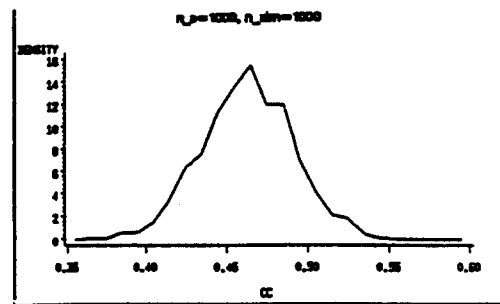
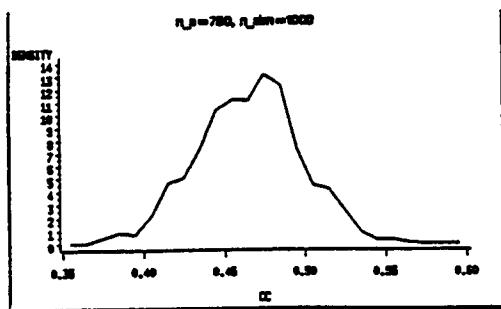
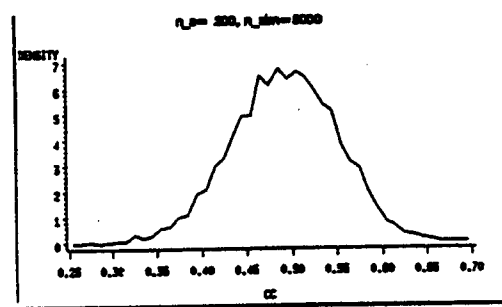
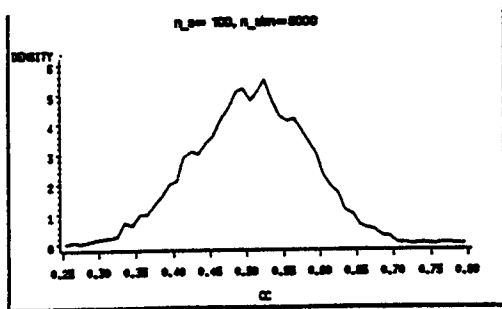
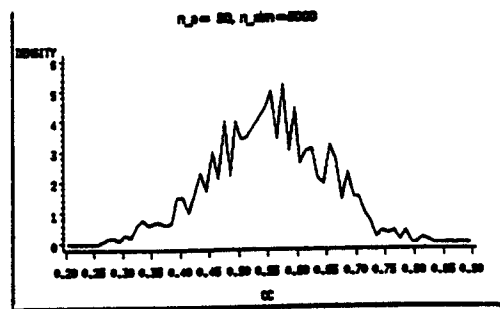


figure 6.7  
 Sampling distribution  
 of RCC  
 MN population  
 second covariance matrix

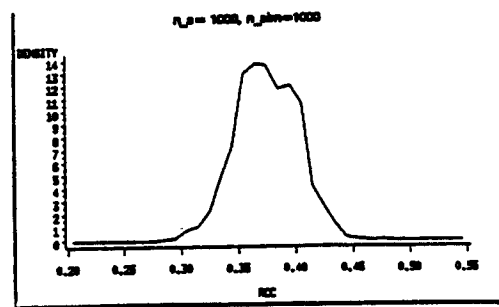
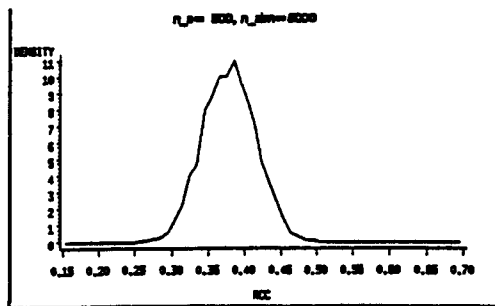
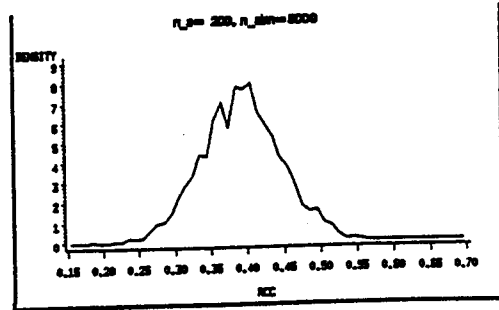
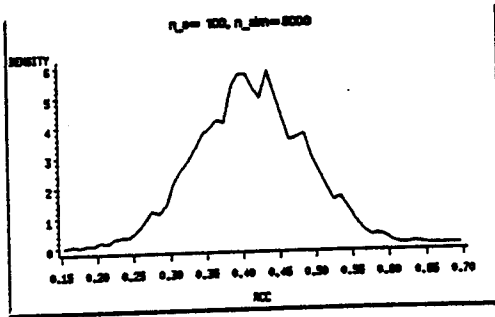
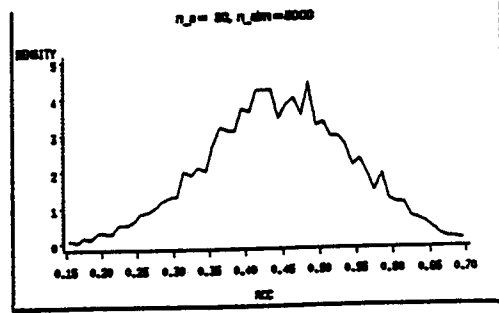


figure 6.8  
Sampling distribution  
of CC  
MN population  
second covariance matrix

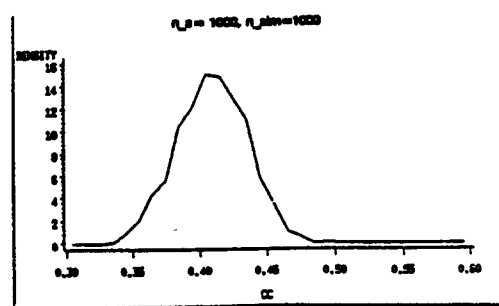
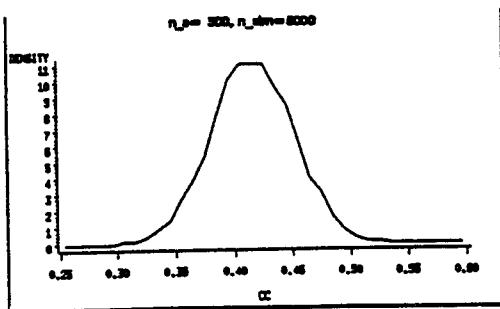
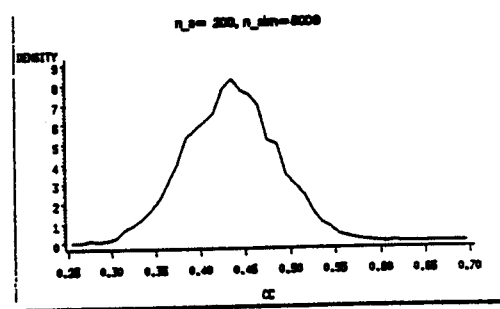
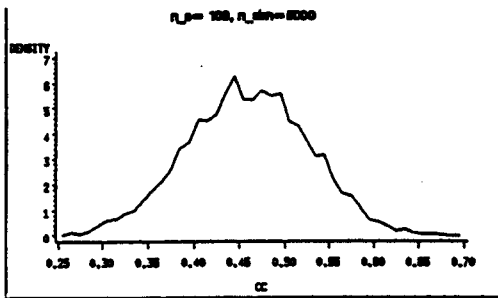
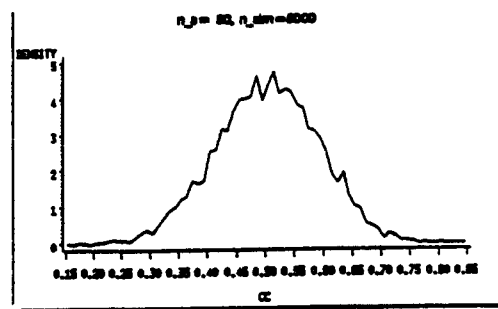


figure 6.9  
Sampling distribution  
of RCC  
MN population  
third covariance matrix

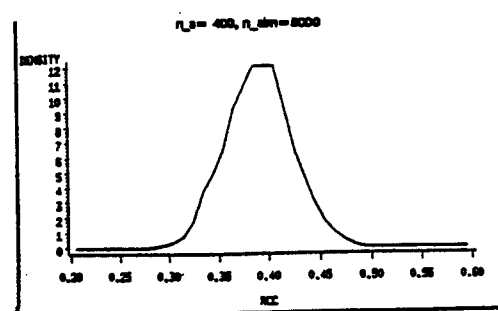
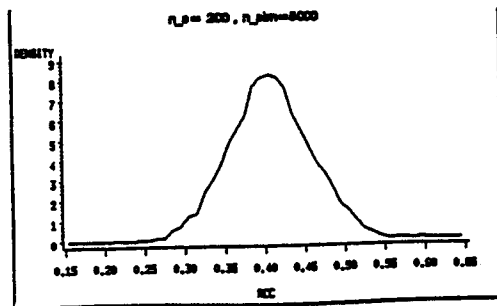
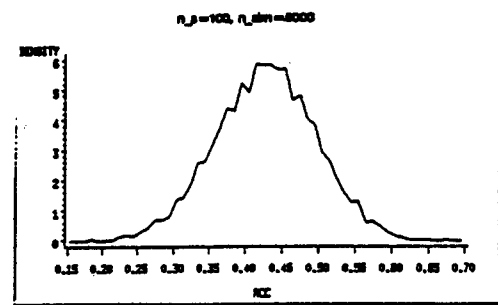
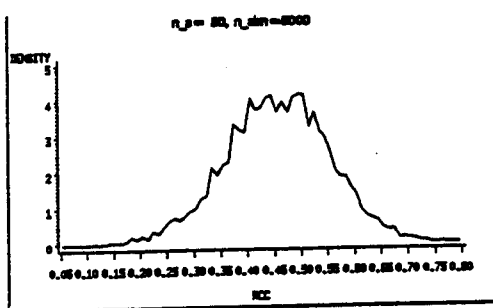
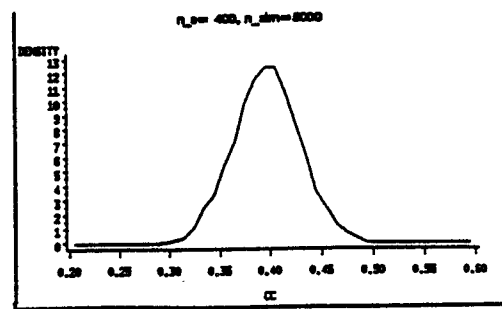
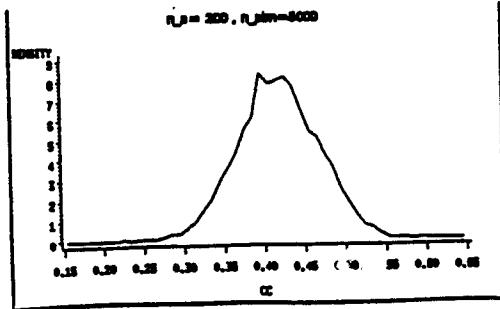
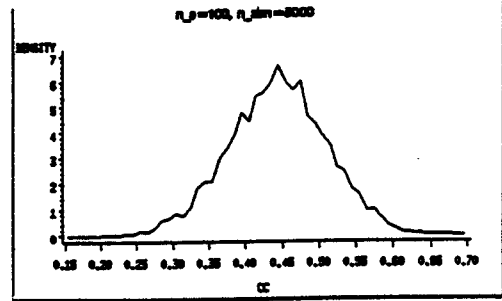
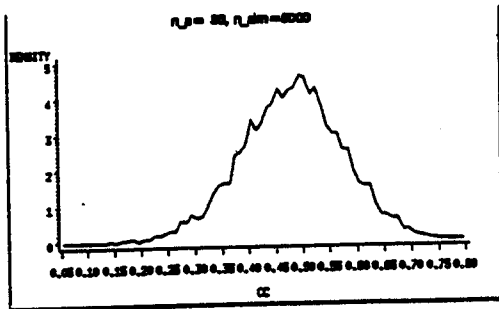




figure 6.10  
Sampling distribution  
of CC  
MN population  
third covariance matrix



## 6.4 Comparing Performances of the Resampling methods in Estimating Sample Variance

In this section, the goal is to study the performance of the jackknife and the bootstrap estimate of variance for sample RCC and sample CC for finite sample size. The formula for the jackknife estimate of variance is given in chapter 5. As for bootstrap, if the bootstrap replicates of  $T_n$  are denoted by  $T_{n,1}^*, \dots, T_{n,B}^*$ , then the bootstrap estimate of variance is given by:

$$\frac{1}{B-1} \sum_{i=1}^B (T_{n,i}^* - \bar{T}_n^*)^2$$

where  $\bar{T}_n^* = \frac{1}{B} \sum_{i=1}^B T_{n,i}^*$ .

In spite of the theoretical justification provided in the earlier section, there is a need to explore the situation when the sample size is not “so large”.

Table 6.1 and 6.2 provide the variance estimators of RCC and CC, respectively, when the sample is drawn from a MP population with the first covariance matrix  $\Sigma_1$ . The numbers in the parenthesis (in the charts) indicate the number of simulations (n\_sim) on which it was based on. For the simulated variance the tabulated figure is simply the variance of n\_sim simulated values of the statistic. For the jackknife and the bootstrap estimate, the tabulated figure represents the simple average of n\_sim respective estimates of variance. Bootstrap resample size is taken to be 1000.

The performances of both jackknife and bootstrap estimators are very good, especially for sample size 500 and higher. The variations for these sample sizes are due to sampling fluctuations. However, for smaller sample sizes, bootstrap does a significantly better job, although possibly at the cost of somewhat higher computation. Recall that for  $\Sigma_1$ , population CC and RCC are the same. This is reflected in the tables for sample sizes 750 and 1000. But there is some difference in the picture for smaller sample sizes like 50 or 100, which is expected.

As one moves over to tables 6.3 and 6.4, which are for MN population, the similar features of jackknife and bootstrap estimators can be observed. The noticeable difference from the earlier tables is that these two tables are very similar even for smaller sample sizes. The only exception to this are the bootstrap estimates for sample size 100, where the estimate in the case of CC is not nearly as good as that for RCC.

Table 6.5 through 6.8 represent the same as the four earlier tables; except now the population covariance matrix is  $\Sigma_2$ . For this population covariance matrix, RCC is not equal to CC, although the assumption 1 is still valid. These tables make it more apparent that the jackknife estimate of variance overshoots the target for smaller sample sizes. This matches with the theoretical result of Efron and Stein (1981), which shows that in general, the jackknife estimate of variance overestimates the target sample variance. The most interesting feature in table 6.7 and 6.8 (MN population) is that the bootstrap estimates are excellent even for sample size 50 and 100. The rest of the picture is very similar to that of the previous ones.

Table 6.9 and 6.10 is for the case when the samples are drawn from MN population with the third covariance matrix. The estimates are tabulated for sample sizes 50, 100, 200 and 500. From these, it does not seem that the performance of the resampling methods is any worse than the earlier situations. This is in parity with what was observed for sampling distributions in the previous section.

Table 6.1: Estimate of Var(RCC)  
 Multivariate Poisson population  
 The first Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.0140845 (1000)	0.0205719 (50)	0.0133118 (50)
100	0.007365 (5000)	0.0106648 (50)	0.0071625 (50)
200	0.004183 (5000)	0.004847 (1000)	0.0042208 (100)
500	0.00176 (5000)	0.0019593 (50)	0.0017812 (100)
750	0.001212 (1000)	0.0012969 (50)	0.0012061 (50)
1000	0.000877 (1000)	0.0009554 (50)	0.0009301 (50)

Table 6.2: Estimate of Var(CC)

Multivariate Poisson population  
 The first Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.0107541 (1000)	0.0220913 (50)	0.0106293 (50)
100	0.006346 (5000)	0.0113518 (50)	0.0065513 (50)
200	0.00396 (5000)	0.0049158 (1000)	0.0039332 (100)
500	0.001755 (5000)	0.0019632 (50)	0.0017561 (100)
750	0.001211 (1000)	0.0012971 (50)	0.0012017 (50)
1000	0.000876 (1000)	0.0009553 (50)	0.0009295 (50)

Table 6.3: Estimate of Var(RCC)

Multivariate Normal population  
The first Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.0127449 (5000)	0.0174378 (50)	0.0117443 (50)
100	0.007052 (5000)	0.0086276 (50)	0.0066271 (50)
200	0.003538 (1000)	0.0044168 (50)	0.0037989 (50)
500	0.001597 (1000)	0.0017409 (50)	0.0016539 (50)
750	0.001062 (1000)	0.0011541 (50)	0.0011024 (50)
1000	0.000838 (5000)	0.0008636 (50)	0.0008396 (50)

Table 6.4: Estimate of Var(CC)

Multivariate Normal population  
The first Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.0097001 (5000)	0.0174245 (50)	0.0093012 (50)
100	0.006138 (5000)	0.0088072 (50)	0.0057676 (50)
200	0.003413 (1000)	0.0044216 (50)	0.0034839 (50)
500	0.001593 (5000)	0.0017426 (50)	0.0016329 (50)
750	0.001061 (1000)	0.001154 (50)	0.0010978 (50)
1000	0.000838 (5000)	0.0008634 (50)	0.0008374 (50)

Table 6.5: Estimate of Var(RCC)

Multivariate Poisson population  
The second Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.012062 (5000)	0.015672 (50)	0.0111898 (50)
100	0.0067031 (5000)	0.0078288 (50)	0.0064401 (50)
200	0.003652 (5000)	0.0040313 (50)	0.0035001 (50)
500	0.001589 (5000)	0.0016525 (50)	0.0015341 (50)
750	0.000995 (1000)	0.0011161 (50)	0.0010451 (50)
1000	0.000823 (1000)	0.0008546 (50)	0.0007916 (50)

Table 6.6: Estimate of Var(CC)

Multivariate Poisson population  
The second Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.009711 (5000)	0.013579 (50)	0.0081902 (50)
100	0.0058868 (5000)	0.007431 (50)	0.0050273 (50)
200	0.003419 (5000)	0.003756 (50)	0.0031611 (50)
500	0.001544 (5000)	0.0016074 (50)	0.0014998 (50)
750	0.001047 (1000)	0.0010868 (50)	0.0010206 (50)
1000	0.000785 (1000)	0.0008272 (50)	0.0007943 (50)

Table 6.7: Estimate of Var(RCC)

Multivariate Normal population  
The second Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.010003 (5000)	0.0146193 (50)	0.0098318 (50)
100	0.005507 (5000)	0.0072462 (50)	0.0056143 (50)
200	0.002944 (5000)	0.0036626 (50)	0.002986 (50)
500	0.0013131 (5000)	0.0014705 (50)	0.0012675 (50)
750	0.000918 (1000)	0.0009257 (50)	0.00087 (50)
1000	0.000716 (1000)	0.0007259 (50)	0.0006626 (50)

Table 6.8: Estimate of Var(CC)

Multivariate Normal population  
The second Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.0079041 (5000)	0.0134894 (50)	0.0077126 (50)
100	0.004565 (5000)	0.0065777 (50)	0.0045617 (50)
200	0.002613 (5000)	0.0034733 (50)	0.0025779 (50)
500	0.0011975 (5000)	0.001563 (50)	0.0011592 (50)
750	0.000822 (1000)	0.0008466 (50)	0.0008102 (50)
1000	0.000648 (1000)	0.0006778 (50)	0.0006128 (50)

Table 6.9: Estimate of Var(RCC)

Multivariate Normal population  
The third Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.0092764 (5000)	0.013008 (50)	0.0086782 (50)
100	0.00477 (5000)	0.0065201 (50)	0.0048928 (50)
200	0.0025235 (5000)	0.0035114 (50)	0.0026218 (50)
500	0.0009228 (50)	0.0014159 (50)	0.0011461 (50)

Table 6.10: Estimate of Var(CC)

Multivariate Normal population  
The third Covariance matrix

<i>sample size</i>	<i>simulated variance</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.0085378 (5000)	0.0125524 (50)	0.0078283 (50)
100	0.004437 (5000)	0.0066084 (50)	0.0047414 (50)
200	0.0024047 (5000)	0.003513 (50)	0.0026631 (50)
500	0.0011691 (50)	0.0014061 (50)	0.0011286 (50)



## 6.5 Bias Reduction for Canonical Correlation and Restricted Canonical Correlation

In this section, the interest is to study the performance of the jackknife and bootstrap estimates of bias of the sample CC and RCC. In fact, the bias corrected jackknife and bootstrap estimates are considered here, and they are compared with the original estimates. The formula of jackknife estimate of bias is as described in chapter 5. The bootstrap estimate of bias is given by  $\hat{E}_{F_n}(T_n^*) - \theta(F_n)$ . Hence the bias corrected bootstrap estimate of RCC is  $2T_n - \hat{E}_{F_n}(T_n^*)$ , and the last term is approximated by  $\bar{T}_n^* = \frac{1}{B} \sum_{i=1}^B T_{n,i}^*$ .

The choice of the parameters is as described before. The tables provide the “estimates” of RCC or CC, rather than the bias. By “estimate”, one means that corresponding to the column of simulated average, one has the simple average of `n_sim` (1000 or 5000) original estimates (sample RCC and CC.) For the jackknife column, this refers to the average over `n_sim` (here, this is usually 50) simulated replicates of bias-corrected jackknife estimate of RCC or CC. Similarly, for the bootstrap column, it refers to the average over `n_sim` (usually 50) “bootstrap estimates”, where for each simulation, the “bootstrap estimate” simply refers to the average of `n_B` bootstrap replicates of sample RCC or CC.

The general features of the data provided in the following tables show that both the sample CC and sample RCC have significant positive bias for sample sizes less than 200. Of course, the bias goes down as the sample size increases and it is barely noticeable for sample size 750 and 1000. Jackknifing does an excellent job of reducing this bias as the bias-corrected jackknife estimates are remarkably closer to the target for small sample sizes. The performance of the bootstrap is quite comparable; although for sample sizes 50 and 100, jackknife seems to work slightly better.

Table 6.11: Estimate of RCC  
Multivariate Poisson population  
The first Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.3595421 (1000)	0.2677187 (50)	0.2859487 (50)
100	0.3297957 (1000)	0.2861678 (50)	0.3166759 (50)
200	0.288904 (5000)	0.2600668 (1000)	0.2694279 (100)
500	.270241 (5000)	0.2659456 (50)	0.2759455 (100)
750	0.266949 (1000)	0.2563418 (50)	0.2606443 (50)
1000	0.264571 (1000)	0.2596893 (50)	0.2586034 (50)

Table 6.12: Estimate of CC

Multivariate Poisson population  
The first Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.3974844 (1000)	0.2457028 (50)	0.2710151 (50)
100	0.340491 (1000)	0.27724 (50)	0.2983459 (50)
200	0.2914531 (5000)	0.2555381 (1000)	0.2597665 (100)
500	0.270308 (5000)	.2652422 (50)	.2752074 (100)
750	0.266959 (1000)	0.2562171 (50)	0.2603602 (50)
1000	0.264577 (1000)	0.2596765 (50)	0.2586512 (50)

Table 6.13: Estimate of RCC

Multivariate Normal population  
The first Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.3494836 (5000)	0.2667595 (50)	0.2923938 (50)
100	0.311106 (5000)	0.25963 (50)	0.2646331 (50)
200	0.288701 (1000)	0.2504751 (50)	0.2535769 (50)
500	0.273393 (1000)	0.2631003 (50)	0.2635439 (50)
750	0.269396 (1000)	0.2633955 (50)	0.2636561 (50)
1000	0.268179 (1000)	0.2651156 (50)	0.2650349 (50)

Table 6.14: Estimate of CC

Multivariate Normal population  
The first Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.3788287 (5000)	0.2376932 (50)	0.2789343 (50)
100	0.320236 (5000)	0.2447502 (50)	0.2531109 (50)
200	0.290368 (1000)	0.2474803 (50)	0.2470323 (50)
500	0.273503 (1000)	0.2626595 (50)	0.2626318 (50)
750	0.269414 (1000)	0.2632444 (50)	0.2633254 (50)
1000	0.268184 (5000)	0.2650072 (50)	0.2648893 (50)

Table 6.15: Estimate of RCC

Multivariate Poisson population  
The second Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.460588 (5000)	0.3999115 (50)	0.4031173 (50)
100	0.430442 (5000)	0.3934334 (50)	0.4004533 (50)
200	0.414041 (5000)	0.3967461 (50)	0.3941454 (50)
500	0.402746 (5000)	0.4060295 (50)	0.4127208 (50)
750	0.402019 (1000)	0.4089961 (50)	0.4204441 (50)
1000	0.400537 (1000)	0.4003661 (50)	0.4049725 (50)

Table 6.16: Estimate of CC

Multivariate Poisson population  
The second Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.55772 (5000)	0.4848791 (50)	0.5077199 (50)
100	0.5145316 (5000)	0.4925893 (50)	0.5092311 (50)
200	0.49157 (5000)	0.4908573 (50)	0.5185382 (50)
500	0.476784 (5000)	0.4783313 (50)	0.4884502 (50)
750	0.474621 (1000)	0.4779874 (50)	0.4840491 (50)
1000	0.472773 (1000)	0.4715559 (50)	0.4730235 (50)

Table 6.17: Estimate of RCC

Multivariate Normal population  
The second Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.449994 (5000)	0.3860778 (50)	0.3910302 (50)
100	0.41943 (5000)	0.37686 (50)	0.3753686 (50)
200	0.401147 (5000)	0.3726186 (50)	0.3678591 (50)
500	0.3889738 (5000)	0.3725544 (50)	0.3747715 (50)
750	0.38522 (1000)	0.379539 (50)	0.3781769 (50)
1000	0.384987 (1000)	0.3831143 (50)	0.3819569 (50)

Table 6.18: Estimate of CC

Multivariate Normal population  
The second Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.511775 (5000)	0.4076308 (50)	0.4238424 (50)
100	0.4445549 (5000)	0.3943105 (50)	0.4008965 (50)
200	0.4251509 (5000)	0.3866343 (50)	0.3968099 (50)
500	0.4255187 (5000)	0.4088523 (50)	0.4102269 (50)
750	0.420936 (1000)	0.4098006 (50)	0.4131896 (50)
1000	0.4217597 (1000)	0.416223 (50)	0.4159928 (50)

Table 6.19: Estimate of RCC

Multivariate Normal population  
The third Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.4927176 0.46545(5000)	0.4633863 (50)	0.459714 (50)
100	0.43679 (5000)	0.3839884 (50)	0.4041279 (50)
200	0.414646 (5000)	0.3847468 (50)	0.389045 (50)
500	0.4016535	0.3906379 (50)	0.3903485 (50)

Table 6.20: Estimate of CC

Multivariate Normal population  
The third Covariance matrix

<i>sample size</i>	<i>simulated average</i>	<i>jackknife estimate</i>	<i>bootstrap estimate</i>
50	0.48873 (5000)	0.4634067 (50)	0.4677923 (50)
100	0.43679 (5000)	0.4093445 (50)	0.4170501 (50)
200	0.42705 (5000)	0.3896175 (50)	0.3957414 (50)
500	0.408184	0.392592 (50)	0.3943236 (50)

# Appendix A

## Multivariate Poisson Distribution

In general, if the marginals are univariate Poisson variables, then one can refer to the joint distribution as multivariate Poisson (MP). However, there have been several attempts in the literature in making this class more structured (and hence narrower). A good account of this is provided in Johnson and Kotz (1969). The most standard approach is that of Teicher (1954), which is discussed briefly in the following.

Let  $U_1, \dots, U_k$  be independent Poisson random variables with mean  $\lambda_1, \dots, \lambda_k$  respectively; where  $k = 2^p - 1$  for some positive integer  $p$ . Let  $B$  be a  $p \times k$  dimensional binary matrix where columns of  $B$  consist of all  $p$ -dimensional binary column vectors except for the null one (i.e. which consist of all 0's). Let  $\mathbf{U} = (U_1 \dots U_k)'$ . Then  $\mathbf{X} = B \times \mathbf{U}$  is said to have a  $p$ -dimensional MP distribution. Let  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)' = E(\mathbf{X})$ . Clearly, the distribution is characterized by  $2^p - 1$  parameters  $\lambda_1, \dots, \lambda_k$  which typically represent  $E\{\prod_1^p (X_i - \mu_i)^{\gamma_i}\}$ , where  $\gamma_i$ 's are binary, but not all 0's. This is very unlike the multivariate Normal distribution which is characterized by only  $\frac{p(p+1)}{2}$  parameters, namely the first and second

order moments. The characteristic function of a MP distribution is of the form:

$$\phi(t_1, \dots, t_p) = \exp\{-A_p + \sum_i^p a_i z_i + \sum_{i < j} a_{ij} z_i z_j + \dots + a_{12\dots p} z_1 z_2 z_p\};$$

where  $z_j = \exp\{it_j\}$  and  $A_p$  is a constant such that  $\phi(0, \dots, 0) = 1$ . For more details see Teicher(1954). If  $\mathbf{X}$  has a multivariate Poisson distribution with mean vector  $(\mu_1, \dots, \mu_p)'$ , then

$$0 \leq \text{Correlation}(X_i, X_j) \leq \min\left(\sqrt{\frac{\mu_i}{\mu_j}}, \sqrt{\frac{\mu_j}{\mu_i}}\right).$$

Thus correlation, between component variables which jointly have MP distribution, is not only nonnegative but is also bounded above (nontrivially). This restricts the scope of applicability of the MP distribution as a model. Obviously, the problem of upper bound is especially severe if the marginal intensities  $\mu_i$  and  $\mu_j$  are not of similar magnitude, since  $\min\left(\sqrt{\frac{\mu_i}{\mu_j}}, \sqrt{\frac{\mu_j}{\mu_i}}\right) \ll 1$  in such a case.



# Appendix B

## Some Results on Jackknifing U-Statistics with a kernel of degree 2

The results discussed here are true for general kernels with little modifications. However only this special case has been required in this present work; also it is easy to follow the proofs for this simple case.

Let  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$  be i.i.d (multivariate) observations from an unknown distribution. Let  $U_n$  be a U-Statistics with a kernel of degree 2 based on the  $\mathbf{Y}_i$ 's; i.e.

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \leq \alpha_1 < \alpha_2}^n f(\mathbf{Y}_{\alpha_1}, \mathbf{Y}_{\alpha_2}) \quad (\text{B.1})$$

So the jackknife replicates are given by,

$$U_{n,-i} = \frac{1}{\binom{n-1}{2}} \sum_{\substack{\alpha_1 < \alpha_2 \\ \neq i}}^n f(\mathbf{Y}_{\alpha_1}, \mathbf{Y}_{\alpha_2}); \quad i = 1, \dots, n. \quad (\text{B.2})$$

**Lemma B.1** Let  $E(U_n) = \xi$ . Then for any  $\epsilon > 0$ ,

$$P[U_n, U_{n-1}, \dots, U_{n-n} \in (\xi - \epsilon, \xi + \epsilon)] \rightarrow 1. \quad (\text{B.3})$$

Proof: By Hoeffding (1948), it follows that  $U_n$  is weakly (in fact strongly) consistent for  $\xi$ . i.e.

$$P[U_n \in (\xi - \epsilon/2, \xi + \epsilon/2)] \rightarrow 1. \quad (\text{B.4})$$

So, it is enough to show that,

$$P[\max_{1 \leq i \leq n} (U_{n,-i} - U_n) > \epsilon/2] \rightarrow 0. \quad (\text{B.5})$$

Now, from (B.1) and (B.2), we get

$$\begin{aligned} \binom{n}{2} U_n &= \binom{n-1}{2} U_{n,-i} + \sum_{\substack{\alpha=1 \\ \neq i}}^n f(\mathbf{Y}_i, \mathbf{Y}_\alpha) \\ \text{i.e. } U_{n,-i} &= \binom{n}{2} \binom{n-1}{2}^{-1} U_n - \binom{n-1}{2}^{-1} \sum_{\substack{\alpha=1 \\ \neq i}}^n f(\mathbf{Y}_i, \mathbf{Y}_\alpha) \\ &= \frac{n}{n-2} U_n - \frac{2}{n-2} Z_i \end{aligned} \quad (\text{B.6})$$

where  $Z_i = \frac{1}{n-1} \sum_{\substack{\alpha=1 \\ \neq i}}^n f(\mathbf{Y}_i, \mathbf{Y}_\alpha)$ . Now  $Z_i$ 's are exchangeable random variables with bounded variance. So

$$\begin{aligned} P\left\{ \left( \max_{1 \leq i \leq n} \frac{2Z_i}{n-2} \right) > \epsilon \right\} &\leq n P\left\{ \frac{2Z_1}{n-2} > \epsilon \right\} \\ &\leq \frac{4n}{\epsilon^2 (n-2)^2} \times \text{Var}(Z_1) \\ &\rightarrow 0. \end{aligned} \quad (\text{B.7})$$

Clearly, (B.5) follows from (B.6) and (B.7).  $\square$

**Lemma B.2**

$$\sup_{0 \leq j \leq n} |U_{n,-j} - \xi| \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty. \quad (\text{B.8})$$

Proof: Let  $\mathcal{C}_n$  denote the  $\sigma$ -field generated by  $\{Y_{n:1}, Y_{n:2}, \dots, Y_{n:n}, Y_{n+k}, k \geq 1\}$ , where  $Y_{n:i}$  denote the  $i$ -th order statistics of  $Y_1, \dots, Y_n$ . Bhattacharya and Sen (1977) have shown that:

$$n(n-1)\mathbf{E}[(U_{n-1} - U_n)^2 | \mathcal{C}_n] \rightarrow 4\zeta \text{ as } n \rightarrow \infty \quad (\text{B.9})$$

where  $\zeta = \text{var}\{\mathbf{E}\{f(Y_i, Y_\alpha) | Y_i\}\}$ . But, L.H.S. of (B.9) is

$$(n-1) \sum_{i=1}^n (U_{n,-i} - U_n)^2$$

and hence,

$$\max_{1 \leq i \leq n} (U_{n,-i} - U_n)^2 \leq \sum_{i=1}^n (U_{n,-i} - U_n)^2 \xrightarrow{a.s.} 0 \quad (\text{B.10})$$

But, also  $U_n \xrightarrow{a.s.} \xi$ . This, along with (B.10) proves (B.8).  $\square$

**Corollary B.2.3** *If  $U_n$  is a vector-valued ( $k$ -variate)  $U$ -statistic and  $f$  is a real-valued function on  $R_k$ , then*

$$\sup_{0 \leq j \leq n} |f(U_{n,-j}) - f(\xi)| \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty. \quad (\text{B.11})$$

*when  $f$  is differentiable and the partial derivatives of  $f$  are bounded.*

Proof: By the lemma, it trivially follows that:

$$\sup_{0 \leq j \leq n} \|U_{n,-j} - \xi\|_{L^p} \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty. \quad (\text{B.12})$$

Now,

$$f(U_{n,-j}) - f(\xi) = (U_{n,-j} - \xi)'(\delta f / \delta \xi)|_{\xi_0}$$

and the result follows from (B.12) and the assumption.  $\square$

# Appendix C

## Computer Programs

### C.1 Program to Compute Restricted Canonical Correlation

```
proc iml;
start rcc(cancorr,rc.up,cf_a,cf_b,s11,s22,s12);
p=nrow( s11);
q=nrow( s22);
s =( s11|| s12)//( s12' || s22);
b_inv =(inv( s11)||J( p, q,0))//(J( q, p,0)||inv( s22) );
F=root(b_inv);
M=F * s * F';
call eigen(var,ver,M);
miu=var-1;
cancorr=max(miu);
rc.up=-1;

    do ip=1 to (2 ** p)-1;
```

```

cc=1;
re_p=ip;
do j=p to 1 by -1;
If (re_p >= 2 ** (j-1)) & cc=1 then do;
sp=j;
cc=cc+1;
re_p=re_p - 2 ** (j-1);
end;
else if (re_p >= 2 ** (j-1)) & cc>1 then do;
sp=sp||j;
re_p=re_p - 2 ** (j-1);
end;
end;

      do iq=1 to (2 ** q)-1;
cc=1;
re_q=iq;
do j=q to 1 by -1;
If (re_q >= 2 ** (j-1)) & cc=1 then do;
sq=j;
cc=cc+1;
re_q=re_q - 2 ** (j-1);
end;
else if (re_q >= 2 ** (j-1)) & cc>1 then do;
sq=sq||j;
re_q=re_q - 2 ** (j-1);
end;
end;
ss11=s11[sp,sp];
ss22=s22[sq,sq];
ss12=s12[sp,sq];

```

```

ss=(ss11||ss12)/(ss12' ||ss22);
pp=nrow(ss11);
qq=nrow(ss22);
b_inv =(inv(ss11)||J(pp,qq,0))/(J(qq,pp,0)|(inv(ss22)) );
F=root(b_inv);
M=F * ss * F';
call eigen(var,ver,M);
miu=var-1;
ver=ss * F' * ver;
l=b_inv * ver;
nn=pp+qq;
do mm=1 to nn ;
x=l[,mm];
alpha=x[1:pp,];
beta=x[(pp+1):nn,];
a=alpha' * ss11 * alpha;
b=beta' * ss22 * beta;
if(all(x>0) | all(x<0)) &( abs(a-b)<.00001)& (miu[mm,]>rc_up) then do;
x=abs(x)/(b ** 0.5);
cf_a =J(p,1,0);
cf_b=J(q,1,0);
do i=1 to pp;
cf_a[sp[,i,]]=x[i,];
end;
do i=1 to qq;
cf_b[sq[,i,]]=x[(pp+i),];
end;

rc_up=miu[mm,];
end; /* corresponding to if..then do */
end; /* corresponding to mm loop */

```

```
end; /* corresponding to jp loop */
end; /* corresponding to ip loop */
finish rcc; quit;
```

## C.2 Program to Generate sample Restricted Canonical Correlation Replicates to Compute Simulated Bias and Variance

```
/* define the correlation matrix ( upper part only) */
p=2 ;
q=3 ;
npq=p+q ;
r=J(npq,npq,0);
do i=1 to (p-1);
do j=(i+1) to p;
r[i,j]=1/p ;
end;
end;
do i=1 to (q-1);
do j=(i+1) to q;
r[p+i,p+j]=1/q;
end;
end;
do i=1 to p;
do j=1 to q;
r[i,p+j]=1/(p * q);
end;
end;
```

```

/* end define R. Now simulating observations */
/* n_s=sample size
n_sim=simulation size
i_s=sample index
i_sim=simulation index */

n_s=500 ;
n_sim=1000;
bin=.01;
n_class=1/bin ;

seed=1243;
freqcc=J(n_class,1,0);
freqrcc=J(n_class,1,0);

do i_sim=1 to n_sim; /* start loop1 */
do i_s=1 to n_s; /* start loop2 */

u=J(npq,npq,0);
do i=1 to (npq-1); /* start loop3 */
do j=(i+1) to npq; /* start loop4 */
If (r[i,j] > 0) then do; /* start loop5 */
T=1;
do while (T > 0) ; /* start loop6 */
ex= -log(1-uniform(seed))/r[i,j];
seed= seed + 7 ;
If ex<T then u[i,j]=u[i,j]+1;
T=T-ex ;
end ; /* end loop6 */
end ; /* end loop5 */

```



```

end ; /* end loop4 */
end ; /* end loop3 */
x= (u+u') * J(npq,1);
if i_s=1 then y=x;
else y=y||x;
end ; /* end loop2 */
Xbar=Y * J(n_s,n_s,1)/n_s; /* Xbar is npq x n_s dimen */
S=(Y-Xbar) * (Y-Xbar)'/(n_s-1);
s11=S[1:p,1:p];
s22=s[(p+1):(p+q),(p+1):(p+q)];
s12=s[1:p,(p+1):(p+q)];
Call RCC(cancorr,rescorr, cfa,cfb,S11,s22,s12);
If i_sim=1 then do;
sm_st_cc=cancorr;
sm_st_rc=rescorr;
end;
else do;
sm_st_cc=sm_st_cc//cancorr;
sm_st_rc=sm_st_rc//rescorr;
end;
ind=floor(n_class * cancorr) ;
freqcc[ind,1]=freqcc[ind,1]+1;
ind=floor(n_class * rescorr) ;
freqrcc[ind,1]=freqrcc[ind,1]+1;

end; /* end loop 1 */

freqcc=freqcc/n_sim/bin;
freqrcc=freqrcc/n_sim/bin;

```

```

    /* print statements */
print "This study was done with sample size= " n_s;
print "simulation size= " n_sim;

```

```

    filename out1 'freqden3.rcc';
file out1;
do i=1 to n_class;
put (freqrcc[i,1]) 3.2 +2 @;
put /;
end;
closefile out1;

```

```

    filename out2 'freqden3.cc';
file out2;
do i=1 to n_class;
put (freqcc[i,1]) 3.2 +2 @;
put /;
end;
closefile out2;

```

### C.3 Program for Jackknifing Restricted Canonical Correlation

```

/* n_s=sample size
n_sim=simulation size
i_s=sample index
i_sim=simulation index */

```

```

n_s=750 ;
n_sim=50 ;

seed=1243;
cc_jk=J(n_sim,1);
vr_cc_jk=J(n_sim,1);
rc_jk=J(n_sim,1);
vr_rc_jk=J(n_sim,1);

      do i_sim=1 to n_sim; /* start loop1 */
do i_s=1 to n_s; /* start loop2 */

u=J(npq,npq,0);
do i=1 to (npq-1); /* start loop3 */
do j=(i+1) to npq; /* start loop4 */
If (r[i,j] > 0)
then do; /* start loop5 */
T=1;
do while (T > 0) ; /* start loop6 */
ex= -log(1-uniform(seed))/r[i,j];
seed= seed + 7 ;
If ex<T then u[i,j]=u[i,j]+1;
T=T-ex ;
end ; /* end loop6 */
end ; /* end loop5 */
end ; /* end loop4 */
end ; /* end loop3 */
x= (u+seed, * J(npq,1);
if i_s=1 then y=x;
else y=y||x;

```

```

end ; /* end loop2 */
Xbar=Y*J(n_s,n_s,1)/n_s; /* Xbar is npq x n_s dimen */
S=(Y-Xbar)*(Y-Xbar)'/(n_s-1);
s11=S[1:p,1:p];
s22=s[(p+1):(p+q),(p+1):(p+q)];
s12=s[1:p,(p+1):(p+q)];
Call RCC(cancorr,rescorr, cfa,cfb,S11,s22,s12);
If i_sim=1 then do;
sm_st_cc=cancorr;
sm_st_rc=rescorr;
end;
else do;
sm_st_cc=sm_st_cc//cancorr;
sm_st_rc=sm_st_rc//rescorr;
end;
ns_1=1;
do k=2 to n_s;
ns_1=ns_1||k;
end;
do j_ind=1 to n_s; /* start loop7 */
col_ind=remove(ns_1,j_ind);
Jdata=Y[,col_ind]; /* jdata has npq rows (n_s-1) col */
ns1=n_s-1;

      XbarJ=jdata*J(ns1,ns1,1)/ns1; /* Xbarj is npq x ns1 dimen */
SJ=(Jdata-Xbarj)*(Jdata-Xbarj)'/(ns1-1);
sj11=SJ[1:p,1:p];
sj22= SJ[(p+1):(p+q),(p+1):(p+q)];
SJ12= SJ[1:p,(p+1):(p+q)];
/* jkrp_cc = set of cc, from jackknife resamples. similarly jkrp_rc */

```

```

Call RCC(cancorr, rescorr, cfa, cfb, sJ11, sJ22, sJ12);
If J_ind =1 then do;
jkrp_cc=cancorr;
jkrp_rc=rescorr;
end;
else do;
jkrp_cc= jkrp_cc//cancorr;
jkrp_rc= jkrp_rc//rescorr;
end;
end; /* end loop7 */
jkrp_avc =jkrp_cc* J(n_s,1)/n_s;
jkrp_avr =jkrp_rc* J(n_s,1)/n_s;
cc_jk[i_sim]=n_s*sm_st_cc[i_sim] - ns1*jkrp_avc;
rc_jk[i_sim]=n_s*sm_st_rc[i_sim] - ns1*jkrp_avr;
vr_cc_jk[i_sim]=(jkrp_cc-jkrp_avc)'*(jkrp_cc-jkrp_avc) * ns1 /n_s;
vr_rc_jk[i_sim]=(jkrp_rc-jkrp_avr)'*(jkrp_rc-jkrp_avr) * ns1 /n_s;

end; /* end loop1 */

sm_av_cc=sm_st_cc*J(n_sim,1)/n_sim;
sm_av_rc=sm_st_rc*J(n_sim,1)/n_sim;
sm_vr_cc=(sm_st_cc-sm_av_cc)' * (sm_st_cc- sm_av_cc)/n_sim;
sm_vr_rc=(sm_st_rc-sm_av_rc)' * (sm_st_rc- sm_av_rc)/n_sim;
av_cc_jk= cc_jk*J(n_sim,1)/n_sim;
av_rc_jk= rc_jk*J(n_sim,1)/n_sim;
av_vr_cj=vr_cc_jk*J(n_sim,1)/n_sim;
av_vr_rj=vr_rc_jk*J(n_sim,1)/n_sim;
/* print statements */
print "This study was done with sample size= " n_s;
print "simulation size= " n_sim;

```

```

print "simulated average of CC" sm_av_cc;
print "simulated average of RCC" sm_av_rc;
print "simulated variance of CC " sm_vr_cc;
print "simulated variance of RCC" sm_vr_rc;
print "bias corrected jackknifed est(avg) of CC " av_cc_jk;
print "bias corrected jackknifed est(avg) of RCC" av_rc_jk;
print "Jack-knife (avg) estimate of var(cc)" av_vr_cj;
print "Jack-knife (avg) estimate of var(rcc)" av_vr_rj;
n_avvrcj=n_s*av_vr_cj;
print "Jack-knife (avg) est. of pop. var(cc)" n_avvrcj;
n_avvrrj=n_s*av_vr_rj;
print "Jack-knife (avg) est. of pop. var(rcc)" n_avvrrj;
quit;

```

## C.4 Program for Bootstrapping Restricted Canonical Correlation

```

proc iml;
/* n_s=sample size
n_sim=simulation size
n_b=bootstrap size
i_b=bootstrap index
i_s=sample index
i_sim=simulation index */
n_s=750 ;
n_sim=50 ;
n_b=1000;
seed=1243; av_cc_bt=J(n_sim,1);

```

```

vr_cc.bt=J(n_sim,1);
av_rc.bt=J(n_sim,1);
vr_rc.bt=J(n_sim,1);

do i_sim=1 to n_sim; /* start loop1 */
do i_s=1 to n_s; /* start loop2 */

u=J(npq,npq,0);
do i=1 to (npq-1); /* start loop3 */
do j=(i+1) to npq; /* start loop4 */
If (r[i,j] > 0)
then do; /* start loop5 */
T=1;
do while (T > 0) ; /* start loop6 */
ex= -log(1-uniform(seed))/r[i,j];
seed= seed + 1 ;
If ex<T then u[i,j]=u[i,j]+1;
T=T-ex ;
end ; /* end loop6 */
end ; /* end loop5 */
end ; /* end loop4 */
end ; /* end loop3 */
x= (u+u') * J(npq,1);
if i_s=1 then y=x;
else y=y||x;
end ; /* end loop2 */
/* generate the bootstrap resamples Xboot(i,j) here */
do i_b=1 to n_b; /* start loop7 */
seed= seed +1 ;
rand=int(n_s*uniform(J(1,n_s,seed)) +1 );
bt_data=Y[,rand];

```

```

Xbar=bt_data*J(n_s,n_s,1)/n_s; /* Xbar is npq x n_s dimen */
Sboot= (bt_data-Xbar)*(bt_data-Xbar)'/(n_s-1);
sboot11=sboot[1:p,1:p];
sboot22= sboot[(p+1):(p+q),(p+1):(p+q)];
sboot12=sboot[1:p,(p+1):(p+q)];
/* st_cc[.] = set of cc, from bootstrap resamples;similarly st_rc */
Call RCC(cancorr,rescorr,cfa,cfb,Sboot11,sboot22,sboot12);
If i_b =1 then do;
st_cc=cancorr;
st_rc=rescorr;
end;
else do;
st_cc=st_cc//cancorr;
st_rc=st_rc//rescorr;
end;
end; /* end loop7 */
/* print st_cc; */
av_cc_bt[i_sim]= st_cc'* J(n_b,1)/n_b;
vr_cc_bt[i_sim]= (st_cc-av_cc_bt[i_sim])'*(st_cc-av_cc_bt[i_sim])/n_b;
av_rc_bt[i_sim]= st_rc'* J(n_b,1)/n_b;
vr_rc_bt[i_sim]= (st_rc-av_rc_bt[i_sim])'*(st_rc-av_rc_bt[i_sim])/n_b;

end; /* end loop1 */

av_av_cb=av_cc_bt'*J(n_sim,1)/n_sim;
av_av_rb=av_rc_bt'*J(n_sim,1)/n_sim;
av_vr_cb=vr_cc_bt'*J(n_sim,1)/n_sim;
av_vr_rb=vr_rc_bt'*J(n_sim,1)/n_sim;
/* print statements */
print "This study was done with sample size= " n_s;

```



```

print "simulation size= " n_sim;
print "bootstrap size" n_b;
print "Bootstrap(average) Estimate of canonical correlation" av_av_cb;
print "Bootstrap(av) estimate of restricted CC" av_av_rb ;
print "Bootstrap estimate of var(rcc)" av_vr_rb;
print "Bootstrap estimate of var(cc)" av_vr_cb;
run;
quit;

```

## C.5 Simulating from Multivariate Normal Distribution

```

/* define the correlation matrix ( upper part only) */
r = 1 || .4 || .3 || .15 || .15;
r=r/(.4 || 1 || 0 || .35 || .25);
r=r/(.3 || 0 || 1 || .3 || .4);
r=r/(.15 || .35 || .3 || 1 || .2);
r=r/(.15 || .25 || .4 || .2 || 1);
G=root(r);

seed=1243;
freqcc=J(n_class,1,0);
freqrcc=J(n_class,1,0);

do i_sim=1 to n_sim; /* start loop1 */
do i_s=1 to n_s; /* start loop2 */

x=J(npq,1,0);
do i=1 to npq; /* start loop3 */

```

```
x[i]=normal(seed);
end; /* end loop3 */
x=G*x;
seed= seed + 7 ;
if i_s=1 then y=x;
else y=y||x;
end ; /* end loop2 */
Xbar=Y*J(n_s,n_s,1)/n_s; /* Xbar is npq × n_s dimen */
/* and so on ... */

end; /* end loop 1 */
```

# Bibliography

- [1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*. John-Wiley.
- [2] Arvesen, J.N.(1969). Jackknifing U-statistics. *Ann. Math. Statist* 40: pp 2076-2100.
- [3] Baksalary, J.K., Puntanen, S. and Yanai, H. (1992) Canonical correlations associated with symmetric reflexive generalized inverses of the dispersion matrix. *Linear Algebra and its applications* 176: pp 61-74.
- [4] Barnett, V.D. and Lewis, T. (1963) A study of relation between G.C.E. and degree results. *J. Roy. Stat. Soc. A* 126: pp 187-226.
- [5] Bartlett, M.S. (1948) A note on the statistical estimation of demand and supply relations from time series. *Econometrica* 16: pp 323-330.
- [6] Beran, R. and Srivastava, M.S. (1985). Bootstrap tests and covariance regions for functions of a covariance matrix. *Ann. Statist.* 13: pp 95-115.
- [7] Bhattacharya, B.B. and Sen, P.K. (1977) Weak convergence of the Rao-Blackwell estimator of a distribution function. *Ann. Prob.* 5: pp 500-510.
- [8] Brillinger, D.R. (1974). *Time Series: data analysis and theory*. Holt, Rinehart and Winston, Inc.

- [9] Brillinger, D.R. (1988). Some statistical methods for random process data from seismology and neurophysiology. *Ann. Statist.* 16: pp 1-54.
- [10] Constantine, A.G. (1963). Some non-central distribution problems in multivariate analysis. *Ann. Math. Statist.* 34: pp 1270-1285.
- [11] Dempster, A.P. (1969) *Continuous Multivariate Analysis*. Addison-Wesley.
- [12] DeSarbo, W. S., Hausman, R.E., Lin, S. and Thompson, W. (1982) Constrained canonical correlation. *Psychometrika* 47: pp 489-516.
- [13] Dillon, W.R. and Goldstein, M.(1984) *Multivariate Analysis: methods and applications*. John-Wiley.
- [14] Eaton, M.L.(1983) *Multivariate Statistics: a vector space approach*. John-Wiley.
- [15] Efron, B. (1979) Bootstrap methods: another look at the jackknife *Ann. Statist.* 7: pp 1-26.
- [16] Efron, B. and Stein, C. (1981) The jackknife estimate of variance *Ann. Statist.* 9: pp 586-596.
- [17] Galton, F. (1889) *Natural inheritance*. London-Macmillan.
- [18] Hoeffding, W.(1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* 19: pp 293-325.
- [19] Horst, P.(1961). Generalized canonical correlations and their applications to experimental data. *J. Clin. Psychol.* 14: pp 331-347.
- [20] Hotelling, H.(1935) The most predictable criterion. *J. Educ. Psych.* 26: pp 139-142
- [21] Hotelling, H.(1936) Relation between two sets of variates. *Biometrika* 28: pp 321-377.

- [22] Jewell, N.P. and Bloomfield, P.(1983). Canonical correlations of past and future for time series: definitions and theory. *Ann. Statist.* 11: pp 837-847.
- [23] Habib, M.K. and Sen, P.K. (1985). Non-stationary stochastic point process models in neurophysiology with applications to learning. *Biostatistics: Statistics in biomedical, public health, and environmental sciences.* pp 481:509.
- [24] Kandel, E.R. and Schwartz, J.H. (1982). *Principles of Neural Science.* Elsevier/North Holland.
- [25] Kettenring, J.R. (1971). Canonical analysis of several sets of variables. *Biometrika* 58, pp 433-450.
- [26] Khatri, C.G. (1976). A note on multiple and canonical correlation for a singular covariance matrix. *Psychometrika* 41: pp 465-470
- [27] Kshirsagar, A. (1972). *Multivariate Analysis.* Marcel-Dekker.
- [28] Kuperstein, M. and Whittington, D.A. (1981) A practical 24 channel microelectrode for neural recording in vivo. *IEEE transactions on biomedical engineering* BME-28: pp 288-293.
- [29] Miller, R.G.,Jr (1974). The jackknife: a review. *Biometrika* 61: pp 1-15.
- [30] Muirhead, R.J. and Waternaux, C.M.(1980). Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations. *Biometrika* 67: pp 31-43.
- [31] Nagao, H.(1985). On the limiting distribution of the jack-knife statistics for eigenvalues of sample covariance matrix . *Commun. Statist. theor. meth.* 14: pp 1547-1567.
- [32] Pearce, S.C. and Holland, D.A. (1960) Some applications of multivariate methods in botany *Applied statistics* 9: pp 1-7.

- [33] Pearson, K. (1996). Mathematical contributions to the theory of evolution III. Regresion, heredity and panmixia. *Philosophical transactions of the Royal Society of London, Series A* 187: pp 253-318.
- [34] Pearson, K. (1998) Mathematical contributions to the theory of evolution V. On the reconstruction of the stature of prehistoric races. *Philosophical transactions of the Royal Society of London, Series A* 192: pp 169-244.
- [35] Perkel, D.H., Gerstein, G.L., Moore, G.P. (1967). Neuronal Spike trains and stochastic point processes: the Single Spike Train, simultaneous spike trains. *Biophys. J.* 7: pp 391-440.
- [36] Quenouille, M. (1949) Approximate tests of correlation in time series. *J. Roy. Stat. Soc. B* 11: pp 68-84.
- [37] Rao, B.R. (1969). Partial canonical correlations. *Trabajos de estadistica y de investigacion operativa* 20: pp 211-219.
- [38] Rao, C.R. (1973). *Linear Statistical Inference and its applications*. (2nd ed.) John Wiley.
- [39] Rao, C.R. (1981). A lemma on g-inverse of a matrix and computation of correlation coefficients in the singular case. *Comm. Statist. Theory Methods* 10: pp 1-10.
- [40] Sen, P.K. (1977). Some invariance principles relating to jackknifing and their role in sequential analysis. *Ann. Statist.* 5: pp 316-329.
- [41] Sengupta, A.(1983). Generalised canonical variables. in *Kotz-Johnson: Encyclopedia of statistical sciences, vol. 3* John Wiley.
- [42] Sengupta, A.(1991). Generalized correlations in the singular case *J. of statistical planning and inference* 28: pp 241-245.
- [43] Stewart, D.K. and Love, W.A. (1968). A general canonical correlation index. *Psychological Bulletin* 70: pp 160-163.

- [44] Teicher, H.(1954). On the multivariate Poisson distribution. *Skandinavisk Aktuarietidskrift* 37: pp 1-9.
- [45] Timm, N.H. (1975). *Multivariate analysis with applications in education and psychology*. Belmont: Brooks/Cole.
- [46] Timm, N.H. and Carlson, J.E.(1976). Part and bipartial canonical correlation analysis. *Psychometrika* 41: pp 159-176.
- [47] Tintner, G. (1952). *Econometrics* John Wiley and sons, New York.
- [48] Tuckwell, H.C. (1989). *Stochastic processes in the neurosciences*. Siam.
- [49] Wollenberg, A.L. van den (1977) Redundancy analysis: an alternative for canonical correlation analysis. *Psychometrika* 42: pp 207-219.
- [50] Yanai, H. and Takane Y.(1992) Canonical correlation analysis in linear constraints. *Linear Algebra and its applications* 176: pp 75-89