2012

Weighted Adaptive Methods for Multivariate Response Models with an HIV/ Neurocognitive Application

Jennifer Ann Geis

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

WEIGHTED ADAPTIVE METHODS FOR MULTIVARIATE RESPONSE MODELS WITH AN HIV/NEUROCOGNITIVE APPLICATION

By

JENNIFER ANN GEIS

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 2012

Jennifer Ann Geis defended this dissertation on February 10, 2012.

The members of the supervisory committee were:

Yiyuan She
Professor Directing Thesis

Anke Meyer-Baese
University Representative

Adrian Barbu
Committee Member

Florentina Bunea
Committee Member

Xufeng Niu
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with the university requirements.

To: Mom and Dad

For always reminding me...
... it's not how fast you're going but where you're headed.

ACKNOWLEDGMENTS

I would like to acknowledge and give my most heartfelt thanks to my two advisors and mentors whom, with incredible patience and understanding, have seen me through this: Dr. Florentina Bunea and Dr. Yiyuan She. An additional thanks goes to the professors of the Department of Statistics at Florida State University, particularly Dr. Xufeng Niu, for professing their knowledge to me. A special thanks goes to Dr. Marten Wegkamp of Cornell University for all of his past support. I would also like to thank my other committee members Dr. Barbu and Dr. Meyer-Baese and the department staff, especially Pamela McGhee, for always keeping me on track.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF SYMBOLS AND NOTATIONS

Let $\mathbf{M}$ and $\mathbf{N}$ be generic matrices.

$(\mathbf{M})^-$	the Moore-Penrose generalized inverse of $\mathbf{M}$
$\mathbf{X}'$	the transpose of the matrix $\mathbf{X}$
$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$	the projection matrix
$r(\mathbf{M})$	the rank of the matrix $\mathbf{M}$
$(\mathbf{M})_{ij}$	the entry of $\mathbf{M}$ in the i-th row and j-th column
$\|\mathbf{M}\|_F = (\sum_i \sum_j (\mathbf{M})_{ij}^2)^{1/2}$	the Frobenius norm of the matrix $\mathbf{M}$
$d_i^2(\mathbf{M})$	the i-th largest singular value of $\mathbf{M}$
$\lambda_i(\mathbf{M})$	the i-th largest eigenvalue of $\mathbf{M}$
μ	the tuning parameter in the RSC
μ_{adap}	the adaptive tuning parameter in the RSC
μ_1	the tuning parameter in the WRSC
μ_{adap1}	the adaptive tuning parameter for WRSC
S^2	the unbiased estimator of the variance σ^2
$\hat{\sigma}^2$	the biased ML estimator of the variance σ^2
$\mathbf{M} \otimes \mathbf{N}$	Kronecker product between the matrices $\mathbf{M}$ and $\mathbf{N}$
$\text{vec}(\mathbf{M})$	the vectorized version of $\mathbf{M}$
$\|\|\mathbf{m}_j\|\|_2$	the Euclidean norm of the j-th row of $\mathbf{M}$
$\|\|\mathbf{M}\|\|_{2,1} = \sum_{j=1}^{p} \|\|\mathbf{m}_j\|\|_2$	the sum of the Euclidean norms $\|\|\mathbf{m}_j\|\|_2$ of rows $\mathbf{m}_j$ of $\mathbf{M}$.
$\mathbb{I}_k$	the $k \times k$ identity matrix
$\mathbb{O}$	the matrix of zeros

ABBREVIATIONS

i.i.d.	independently and identically distributed
p.d.	positive-definite
p.s.d.	positive-semidefinite
s.t.	subject to
w.r.t.	with respect to
ACCA	Adaptive Canonical Correlation Analysis
AR(1)	Autoregressive 1
BV	Brain Volumetric
CCA	Canonical Correlation Analysis
CV	Cross-Validation
DTI	Diffusion Tensor Imaging
FA	Fractional Anisotropy
GLASSO	Group Lasso
H-AR(1)	Heterogeneous Autoregressive 1
HIV	Human Immunodeficiency Virus
MD	Mean Diffusivity
MSE	Mean Squared Error
RCCA	Regularized Canonical Correlation Analysis
RRR	Reduced-Rank Regression
RSC	Rank Selection Criterion
SNR	signal to noise ratio
SVD	singular value decomposition
UN	Unstructured
WRSC	Weighted Rank Selection Criterion

ABSTRACT

Multivariate response models are being used increasingly more in almost all fields with the necessary employment of inferential methods such as Canonical Correlation Analysis (CCA). This requires the estimation of the number of uncorrelated canonical relationships between the two sets, or, equivalently so, determining the rank of the coefficient estimator in the multivariate response model.One way to do this is by the Rank Selection Criterion (RSC) by Bunea *et al.* with the assumption the error matrix has independent constant variance entries [8]. While this assumption is necessary to show their strong theoretical results, in practical application, some flexibility is required. That is, such assumption cannot always be safely made.

What is developed here are the theoretics that parallel Bunea *et al.*'s work with the addition of a "decorrelator" weight matrix. One choice for the weight matrix is the residual covariance, but this introduces many issues in practice. A computationally more convenient weight matrix is the sample response covariance. When such a weight matrix is chosen, CCA is directly accessible by this weighted version of RSC giving rise to an Adaptive CCA (ACCA) with principal proofs for the large sample setting.

However, particular considerations are required for the high-dimensional setting, where similar theoretics do not hold. What is offered instead are extensive empirical simulations that reveal that using the sample response covariance still provides good rank recovery and estimation of the coefficient matrix, and hence, also provides good estimation of the number of canonical relationships and variates. It is argued precisely why other versions of the residual covariance, including a regularized version, are poor choices in the high-dimensional setting. Another approach to avoid these issues is to employ some type of variable selection methodology first before applying ACCA. Truly, any group selection method may be applied prior to ACCA as variable selection in the multivariate response model is the same as group selection in the univariate response model and thus completely eliminates these high-dimensional concerns.

To offer a practical application of these ideas, ACCA is applied to a "large sample" neurocognitive dataset. Then, a high-dimensional dataset is generated to which Group LASSO will be first utilized before ACCA. This provides a unique perspective into the relationships between cognitive deficiencies in HIV-positive patients and the extensive, available neuroimaging measures.

CHAPTER 1

INTRODUCTION

While multivariate modeling and analysis techniques are common, they are far from simple. Models with multiple predictor and response variables are ideal in many data analysis settings in nearly any field of interest. But extracting pertinent information from each variable while also considering their within set and opposite set relationships is hardly trivial.

Consider, for instance, a cohort of HIV-positive patients. HIV is known to infect nerve cells and the brain itself, not long after contraction, potentially causing systemic issues in information processing speed and attention, psychomotor abilities, executive functions, verbal fluency, and in learning and memory [45]. Defects in these various neurocognitive "domains" may be measured by a variety of psychological assessments and indices. However, how can the physical ramifications be assessed? Changes in the brain can be identified by numerous diffusion tensor imaging (DTI) and brain volumetric (BV) measures. These quantify the restriction of water diffusion and measure the size of various parts of the brain, respectively. So, can these be linked to known cognitive problems while accounting for standard clinical variables?

It is desirable to withdraw the maximal amount of information from the correlation within the neuroimaging predictor set without forgetting the potential correlation in the response set of cognitive measures. The *multiple linear regression model* has both multiple predictor and response variables, but standard least-squares estimators regress on the response variables separately and independently, failing to utilize the within set correlation of the responses. In addition to this, with both multiple predictors and multiple response variables, the number of parameters that need to be estimated can easily be quite large. Estimators with reduced-ranks were developed to remedy this problem. Such estimation techniques were introduced by Anderson in the 1950s [1], and Reduced-Rank Regression (RRR) was later formally coined and developed by Izenman in 1975 [23]. Similar works were also produced by Robinson [34] [35] and Rao [31] around that time period. While innovative in their own right, the limitations and assumption requirements were less than ideal: These estimators were for fixed, given ranks only and asymptotic in nature, mostly derived in a likelihood framework with a Gaussian assumption on the errors. In 1999, Anderson was able to bypass the Gaussian assumption on the errors [2]. However, this still assumed that the true rank of the coefficient matrix was known and fixed. Later in 2002, he attempted to fix this problem by creating asymptotic rank selection tests, yet these were

only valid when the number of predictors was small and fixed [3].

In addition to these issues, modeling, by itself, provides limited inferential insight. With such a multivariate model, where correlation exists within both sets, how can multiple, unique relationships be identified? Standard statistical inferences utilize p-values and confidence intervals. However, Hotelling's Canonical Correlation Analysis (CCA) offers unique inferential methodology for identifying and interpreting such relationships [22]. Not surprising is that it is a unique version of RRR. While CCA provides a number of useful inference techniques, it has been shown to work poorly in the high-dimensional setting or even when the number of observations is close to the dimension size. Specifically, that is, as the number of variables increases, the canonical correlations become almost 1 and provide no meaningful interpretations. See Eaton and Perlman [11] for a full account. Also, the required sample covariances needed to perform CCA are sometimes ill-conditioned and often, singular and non-invertible.

While a regularized version of CCA has been offered as a solution by Vinod [41] and Leurgans *et al.* [27], Regularized Canonical Correlation Analysis (RCCA) is computationally expensive and time consuming as a 2-dimensional grid of regularization parameters needs to be cross-validated. In addition to this, RCCA examines the largest canonical correlation corresponding to only the first relationship between sets while disregarding the rest. So even if the number of variables may be copious, such as the numerous available DTI and BV measures, if the cohort is relatively small, RCCA is difficult to apply. Even disregarding this issue, as a first step, CCA requires the number of significant, canonical relationships to be known. This is equivalent to determining the rank of the coefficient matrix in RRR but with the previously listed problems. Only after this is found can the canonical weights be derived from the appropriate decomposition of the coefficient estimator to exercise the full power of CCA.

It was just in 2011 that an innovative data-adaptive solution to the rank estimation problem was provided by Bunea, She, and Wegkamp in their paper "Optimal selection of reduced rank estimators of high-dimensional matrices" [8]. While other methods have been constrained by rank and distributional assumptions, Bunea *et al* have introduced a penalized criterion to find a consistent estimator of the effective rank of the coefficient matrix while concurrently estimating the coefficient matrix itself. This is referred to as the Rank Selection Criterion (RSC) and opens a wide range of potential applications, including dataset reduction, as it is valid for any sample size and any number of response and predictor variables, even in the high-dimensional setting. Bunea *et al.* show that the estimated rank is characterized by a tuning parameter. While this tuning parameter may be cross-validated, the addition of an adaptive tuning parameter in a closed form makes RSC extremely convenient and easy to use. However, there is a constraint: the error terms are assumed to be independently and identically distributed (i.i.d.), an assumption that is not necessarily practical in application.

The generalization of RSC without this assumption on the error terms will be carefully examined here in the rank estimation of the coefficient matrix. Examination of how to do so will proceed carefully in two important settings: *large sample* and *high-dimensional*. First, relevant theoretical backing will be provided under the assumption that the correct decorrelator is known. This is essentially the inverse of the error covariance. These theoretics parallel the fundamental theorems by Bunea *et al.*. A simple model transformation will be

made so that RSC may be applied without issue to estimate the rank in the transformed model. Recovery of the final coefficient estimator will be only one additional step, a complete process referred to here as the Weighted Rank Selection Criterion (WRSC). A choice of the tuning parameter will also be provided for such a setup. In practice, however, the weight matrix needs to be estimated, a not so simple task. Careful inspection of this will show that in fact a particular selection of the weight matrix yields an adaptive version of CCA, the so-called Adaptive Canonical Correlation Analysis (ACCA). This relationship is only maintained through a convenient relationship between the population version of the error covariance and the response covariance. With careful tuning parameter selection, ACCA then provides *simultaneous* estimation of the number of significant canonical relationships and the canonical weights themselves.

Treatment of the sample versions of these two weight matrices is not throughly examined in literature. This becomes particularly relevant in the computational aspects of the high-dimensional setting. The connection between ACCA and WRSC only holds if the weight matrix can act as a true "decorrelator" of the model, a difficult requirement in the high-dimensional setting. In addition to this, singularity issues arise as the weight matrix needs to be invertible. A naive alternative would be to simply use the Moore-Penrose pseudoinverse. However, this will show to be at a computational disadvantage; the sample response covariance is much more reliable in practice. A second, popular alternative is to use shrinkage methods to regularize the weight matrix. However, the theoretics for the relationship between the two regularized covariances, so that the connection between ACCA and WRSC is maintained, are non-existent. These theoretics will be established here for the sample regularized versions of the residual and response covariance. Again, through such investigations, it will be readily apparent that the regularized residual covariance is at a disadvantage to the regularized sample response covariance, supporting the argument that the latter is a far more practical choice than the former. These concepts, both in the large sample and high-dimensional setting, will be supported through extensive simulations and demonstrate that WRSC often provides better rank recovery and coefficient estimation than RSC.

A third option for handling high-dimensionality is through a two step process of first applying a variable selection technique prior to ACCA or WRSC application. Variable selection in multivariate response models is equivalent with group variable selection in the univariate response models. Hence, theoretical results in the univariate setting carry over to the multivariate one. Bunea, She, and Wegkamp were among the first to employee such two-step techniques in a variety of settings using RSC [7]. If variable selection can be performed to reduce the predictor space enough, then WRSC and ACCA may also be applied without issue.

To substantiate the findings here, the application of ACCA will be illustrated using a sample of HIV-positive patients to establish unique relationships between the two sets of cognitive variables and neuroimaging measures along with standard clinical variables. [1] The original dataset will be analyzed first as an example of application in a large sample setting. From the original predictor set, a larger predictor set will be generated so that the high-dimensional application may also be shown. This setting will be handled using the

[1] Provided by Dr. Hernando Ombao, Brown University, Research Center for Statistical Sciences.

two step process mentioned above as a combination of a popular group selection technique, Group LASSO (GLASSO), and ACCA. These applications be preceded by first, careful examination of CCA as a unique variation of RRR in Chapter 2, followed by RSC, and a discussion about the choice of the weight matrix. This provides the framework for Chapter 3 where ACCA is developed, and WRSC is formed as consequence. Special considerations will be discussed for treatment of the high-dimensional setting. Following directly will be the supporting simulations for both the large sample and high-dimensional setting in Chapter 4. A detailed analysis using the large sample version of ACCA on the neurocognitive dataset of HIV-positive patients will be provided in Chapter 5, and Chapter 6 will introduce the two step process with variable selection before applying it to the generated high-dimensional neurocognitive dataset. This will be concluded with a summary of the new, relevant findings here of both theoretical and empirical nature and a discussion of continuing work in this area.

CHAPTER 2

BACKGROUND METHODOLOGY

The basic methodology of Canonical Correlation Analysis (CCA) is introduced here as understanding its relationship with Reduced-Rank Regression (RRR) is crucial in the development of Adaptive Canonical Correlation Analysis (ACCA) and thus, also of Weighted Rank Selection Criterion (WRSC). In particular, treatment of the weight matrix will be examined very closely as it is the crux upon which the rest of the theory will be built. First, some basic notation is introduced:

Let $\mathbf{x}$ and $\mathbf{y}$ be two generic vectors of variables of sizes $p \times 1$ and $n \times 1$, respectively, with joint means

$$\mathbb{E}\left[\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right] = \begin{pmatrix}\boldsymbol{\mu}_X\\ \boldsymbol{\mu}_Y\end{pmatrix} \tag{2.1}$$

and covariances matrix

$$\mathbb{E}\left[\begin{pmatrix}\mathbf{x}-\boldsymbol{\mu}_X\\ \mathbf{y}-\boldsymbol{\mu}_Y\end{pmatrix}\begin{pmatrix}\mathbf{x}-\boldsymbol{\mu}_X\\ \mathbf{y}-\boldsymbol{\mu}_Y\end{pmatrix}'\right] = \begin{pmatrix}\boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY}\\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}YY\end{pmatrix}. \tag{2.2}$$

Assume they are related via

$$\mathbf{y}' = \mathbf{x}'\mathbf{A} + \mathbf{e}' \tag{2.3}$$

under the following assumptions:

A.1. Without loss of generality that $\boldsymbol{\mu}_X = 0$ and $\boldsymbol{\mu}_Y = 0$.

A.2. The covariance matrices $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{YY}$ are assumed to be nonsingular unless otherwise noted.

A.2. The generic error vector $\mathbf{e}$ has mean $\mathbb{E}(\mathbf{e}) = 0$ and covariance $Cov(\mathbf{e}) = \boldsymbol{\Sigma}_e$.

Suppose that for $i = 1, ..., m$ observations there are corresponding $\mathbf{x}_i$ predictor and $\mathbf{y}_i$ response vectors that are the i-th realization of the generic vectors $\mathbf{x}$ and $\mathbf{y}$. Let $\mathbf{e}_i$ be the i-th realization of $\mathbf{e}$ and assume that the $\mathbf{e}_i$'s are independently distributed for $i = 1, \cdots, m$.

The model may be written in the standard matrix format for a multivariate response model using the data matrices formed from the $i = 1, \cdots, m$ observations:

$$\mathbf{X} = \begin{pmatrix}\mathbf{x}_1'\\ \mathbf{x}_2'\\ \vdots\\ \mathbf{x}_m'\end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix}\mathbf{y}_1'\\ \mathbf{y}_2'\\ \vdots\\ \mathbf{y}_m'\end{pmatrix}$$

so that the model that relates them may be written as

$$\mathbf{Y} = \mathbf{X}\mathbf{A} + \mathbf{E} \tag{2.4}$$

where $\mathbf{Y}$ is a $m \times n$ response matrix, $\mathbf{X}$ is a $m \times p$ predictor matrix, and $\mathbf{E}$ is the $m \times n$ matrix of error terms. The coefficient matrix $\mathbf{A}$ is an unknown matrix of size $p \times n$.

Assume that $\mathbf{X}$ and $\mathbf{Y}$ have been column-centered by their sample means. The corresponding maximum likelihood (ML) sample covariances are calculated as

$$\widehat{\mathbf{\Sigma}}_{XX} = \frac{1}{m}\mathbf{X}'\mathbf{X}, \tag{2.5}$$

$$\widehat{\mathbf{\Sigma}}_{XY} = \frac{1}{m}\mathbf{X}'\mathbf{Y} = \widehat{\mathbf{\Sigma}}'_{YX}, \tag{2.6}$$

$$\widehat{\mathbf{\Sigma}}_{YY} = \frac{1}{m}\mathbf{Y}'\mathbf{Y} \tag{2.7}$$

where no assumptions are made on the invertibility of the sample covariance matrices $\widehat{\mathbf{\Sigma}}_{XX}$ and $\widehat{\mathbf{\Sigma}}_{YY}$ unless otherwise noted.

2.1 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is used to examine to relationship between two sets of variables. It is a particularly powerful inferential technique used in aiding in interpretation as it does not disregard the potential within-set dependencies within either of the sets. Hence, the shared influences of the two sets upon the common sample may be examined carefully. While it is typically not used to create a direct model, as one set predicting the other, it maintains close ties to Reduced-Rank Regression (RRR). Many sources provide documentation of CCA including Hair *et al.* [18], Izenman [24], Johnson and Wichern [25], and Reinsel and Velu [32]. For the purposes of this section, the terms "predictor" and "response" will be slightly abused to label the two sets $\mathbf{x}$ and $\mathbf{y}$, respectively. In traditional CCA, there is no distinction between the two as treatment is symmetric.

While the previously given references provide the details of the population version of CCA, the sample version of CCA is less carefully documented. In particular, the details of the high-dimensional setting are unclear. The population version is first introduced from two perspectives. The first of these is Hotelling's traditional, sequential way [22]. The second of these is a multivariate approach, particularly well documented by Reinsel and Velu [32]. These two approaches are first given from the population point of view. The sample version of CCA follows, along with a brief description of a particular type of CCA that is applicable in the high-dimensional setting, Regularized Canonical Correlation Analysis (RCCA). However, it will be obvious that RCCA is computationally expensive and neglects to consider any additional relationships that may exist between the two sets aside from the first. CCA's close relationship to Reduced-Rank Regression (RRR) as a *weighted constrained problem* will then be examined, along with some useful additional descriptive measures that will be applied later.

2.1.1 Population Canonical Correlations and Variates

Following the notation from the beginning of Chapter 2, let (ξ_k, ω_k) denote a pair of new variables for $k = 1, \ldots, t \leq min(n, p)$ referred to as *canonical variates* where t is the number of non-zero canonical correlations, also thought of as the number of significant, uncorrelated relationships. For now, assume that t is given and that $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$ are nonsingular and invertible. The canonical variates ξ_k and ω_k are formed by

$$\xi_k = \mathbf{x}'\mathbf{f}_k \tag{2.8}$$

$$\omega_k = \mathbf{y}'\mathbf{h}_k \tag{2.9}$$

where $\mathbf{f}_k$ and $\mathbf{h}_k$ are the *canonical coefficients* or *canonical weights* found such that the k-th largest canonical correlation

$$\rho_k = Corr(\xi_k, \omega_k) = \frac{\mathbf{f}_k'\mathbf{\Sigma}_{XY}\mathbf{h}_k}{(\mathbf{f}_k'\mathbf{\Sigma}_{XX}\mathbf{f}_k)^{1/2}(\mathbf{h}_k'\mathbf{\Sigma}_{YY}\mathbf{h}_k)^{1/2}} \tag{2.10}$$

is maximized with the properties

(i) $Cov\,(\xi_k, \xi_l) = \mathbf{f}_k'\mathbf{\Sigma}_{XX}\mathbf{f}_l = 0, \quad \text{for} \quad l \neq k$

(ii) $Cov\,(\omega_k, \omega_l) = \mathbf{h}_k'\mathbf{\Sigma}_{YY}\mathbf{h}_l = 0, \quad \text{for} \quad l \neq k.$

That is, ξ_l is uncorrelated with all the other ξ_k's, and ω_l is uncorrelated with all the other ω_k's. The new pairs of variables (ξ_k, ω_k) are ordered based upon their corresponding canonical correlations so that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_t$. The appropriate canonical weights $\mathbf{f}_k$ and $\mathbf{h}_k$ may be derived in one of two ways:

Way 1- Sequentially Hotelling's sequential, correlation maximization approach to CCA is derived as follows [22] . Let $\xi = \mathbf{x}'\mathbf{f}$ and $\omega = \mathbf{y}'\mathbf{h}$ be the generic linear projections. First, continue to assume without loss of generality that $\mathbb{E}(\mathbf{x}) = 0$ and $\mathbb{E}(\mathbf{y}) = 0$ so that $\mathbb{E}(\xi) = 0$ and $\mathbb{E}(\omega) = 0$. Furthermore, assume that $\mathbf{f}'\mathbf{\Sigma}_{XX}\mathbf{f} = 1$ and $\mathbf{h}'\mathbf{\Sigma}_{YY}\mathbf{h} = 1$.

The problem reduces to finding the vectors of $\mathbf{f}$ and $\mathbf{h}$ such that

$$Corr(\xi, \omega) = \mathbf{f}'\mathbf{\Sigma}_{XY}\mathbf{h} \tag{2.11}$$

has maximal correlation among all linear functions of $\mathbf{x}$ and $\mathbf{y}$. Set

$$f(\mathbf{f}, \mathbf{h}) = \mathbf{f}'\mathbf{\Sigma}_{XY}\mathbf{h} - \frac{1}{2}\sqrt{\phi}(\mathbf{f}'\mathbf{\Sigma}_{XX}\mathbf{f} - 1) - \frac{1}{2}\sqrt{\theta}(\mathbf{h}'\mathbf{\Sigma}_{YY}\mathbf{h} - 1) \tag{2.12}$$

where $\sqrt{\phi}$ and $\sqrt{\theta}$ are Lagrangian multipliers. Taking the partial derivatives with respect to (w.r.t.) $\mathbf{f}$ and $\mathbf{h}$ gives

$$\frac{\partial f(\mathbf{f}, \mathbf{h})}{\partial \mathbf{f}} = \mathbf{\Sigma}_{XY}\mathbf{h} - \sqrt{\phi}\mathbf{\Sigma}_{XX}\mathbf{f} \tag{2.13}$$

$$\frac{\partial f(\mathbf{f}, \mathbf{h})}{\partial \mathbf{h}} = \mathbf{\Sigma}_{YX}\mathbf{f} - \sqrt{\theta}\mathbf{\Sigma}_{YY}\mathbf{h}. \tag{2.14}$$

Setting these partials to zero and multiplying (2.13) and (2.14) by $\mathbf{f}'$ and $\mathbf{h}'$, respectively gives now

$$\begin{aligned} \mathbf{f}'\boldsymbol{\Sigma}_{XY}\mathbf{h} - \sqrt{\phi}\mathbf{f}'\boldsymbol{\Sigma}_{XX}\mathbf{f} &= 0 \quad (2.15) \\ \mathbf{h}'\boldsymbol{\Sigma}_{YX}\mathbf{f} - \sqrt{\theta}\mathbf{h}'\boldsymbol{\Sigma}_{YY}\mathbf{h} &= 0 \quad (2.16) \end{aligned}$$

so that the correlation satisfies

$$\mathbf{f}'\boldsymbol{\Sigma}_{XY}\mathbf{h} = \sqrt{\phi} = \sqrt{\theta}.$$

Using this to substitute $\sqrt{\phi}$ for $\sqrt{\theta}$ and with rearrangement of terms gives

$$\begin{aligned} -\sqrt{\phi}\boldsymbol{\Sigma}_{XX}\mathbf{f} + \boldsymbol{\Sigma}_{XY}\mathbf{h} &= 0 \quad (2.17) \\ \boldsymbol{\Sigma}_{YX}\mathbf{f} - \sqrt{\phi}\boldsymbol{\Sigma}_{YY}\mathbf{h} &= 0. \quad (2.18) \end{aligned}$$

Multiplying the former of these by $\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}$ and substituting it into the latter gives

$$(\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} - \phi\boldsymbol{\Sigma}_{YY})\mathbf{h} = 0$$

or equivalently

$$(\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2} - \phi\mathbb{I}_n)\mathbf{h} = 0. \quad (2.19)$$

This requires the determinant

$$|\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2} - \phi\mathbb{I}_n| = 0 \quad (2.20)$$

for there to be a nontrivial solution. This is a polynomial of ϕ with n real roots. These are the ordered eigenvalues $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_n \geq 0$ of

$$\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2} \quad (2.21)$$

with corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n$. Hence, the maximal correlation between ξ and ω is given by setting $\sqrt{\phi} = \sqrt{\phi}_1$. This then provides that the choice of the canonical coefficients $\mathbf{f}$ and $\mathbf{h}$ are given by the vectors

$$\begin{aligned} \mathbf{f}_1 &= \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{u}_1 \quad (2.22) \\ \mathbf{h}_1 &= \boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{u}_1. \quad (2.23) \end{aligned}$$

Thus, the first pair of canonical variates is (ξ_1, ω_1) where $\xi_1 = \mathbf{x}'\mathbf{f}_1$ and $\omega_1 = \mathbf{y}'\mathbf{h}_1$ with correlation $Corr(\xi_1, \omega_1) = \mathbf{f}_1'\boldsymbol{\Sigma}_{XY}\mathbf{h}_1 = \sqrt{\phi}_1$.

Now, given the first pair of canonical variates (ξ_1, ω_1), let $\xi = \mathbf{x}'\mathbf{f}$ and $\omega = \mathbf{y}'\mathbf{h}$ denote a second arbitrary pair of variates with unit variances. The second pair of canonical variates have maximal correlation amongst all possible linear combinations of $\mathbf{x}$ and $\mathbf{y}$ and are uncorrelated with the first pair of canonical variates (ξ_1, ω_1). That is,

$$\begin{aligned} \mathbf{f}'\boldsymbol{\Sigma}_{XX}\mathbf{f}_1 &= 0 \quad (2.24) \\ \mathbf{h}'\boldsymbol{\Sigma}_{YY}\mathbf{h}_1 &= 0. \quad (2.25) \end{aligned}$$

Following from (2.17) and (2.18),

$$Corr(\xi, \omega_1) = \mathbf{f}'\boldsymbol{\Sigma}_{XY}\mathbf{h}_1 = \sqrt{\phi}_1\mathbf{f}'\boldsymbol{\Sigma}_{XX}\mathbf{f}_1 = 0 \tag{2.26}$$

$$Corr(\omega, \xi_1) = \mathbf{h}'\boldsymbol{\Sigma}_{YX}\mathbf{f}_1 = \sqrt{\phi}_1\mathbf{h}'\boldsymbol{\Sigma}_{YY}\mathbf{h}_1 = 0. \tag{2.27}$$

Again, $\mathbf{f}$ and $\mathbf{h}$ are chosen to maximize $f(\mathbf{f}, \mathbf{h})$ so that

$$f(\mathbf{f}, \mathbf{h}) = \mathbf{f}'\boldsymbol{\Sigma}_{XY}\mathbf{h} - \frac{1}{2}\sqrt{\phi}(\mathbf{f}'\boldsymbol{\Sigma}_{XX}\mathbf{f} - 1) - \frac{1}{2}\sqrt{\theta}(\mathbf{h}'\boldsymbol{\Sigma}_{YY}\mathbf{h} - 1) + \sqrt{\eta}\mathbf{f}'\boldsymbol{\Sigma}_{XX}\mathbf{f}_1 + \sqrt{\nu}\mathbf{h}'\boldsymbol{\Sigma}_{YY}\mathbf{h}_1 \tag{2.28}$$

where $\sqrt{\phi}$, $\sqrt{\theta}$, $\sqrt{\eta}$, and $\sqrt{\nu}$ are Lagrangian multipliers. Taking the partial derivations of $f(\mathbf{f}, \mathbf{h})$ with respect to $\mathbf{f}$ and $\mathbf{h}$ and setting them equal to zero gives:

$$\frac{\partial f}{\partial \mathbf{f}} = \boldsymbol{\Sigma}_{XY}\mathbf{h} - \sqrt{\phi}\boldsymbol{\Sigma}_{XX}\mathbf{f} + \sqrt{\eta}\boldsymbol{\Sigma}_{XX}\mathbf{f}_1 = 0 \tag{2.29}$$

$$\frac{\partial f}{\partial \mathbf{h}} = \boldsymbol{\Sigma}_{YX}\mathbf{f} - \sqrt{\theta}\boldsymbol{\Sigma}_{YY}\mathbf{h} + \sqrt{\nu}\boldsymbol{\Sigma}_{YY}\mathbf{h}_1 = 0. \tag{2.30}$$

Similar to before, multiplying (2.29) and (2.30) by $\mathbf{f}'$ and $\mathbf{h}'$ respectively and then noting (2.26) and (2.27), these equations reduce to (2.17) and (2.18). Therefore, the second pair of canonical variates (ξ_2, ω_2) are calculated using the canonical coefficients

$$\mathbf{f}_2 = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{u}_2 \tag{2.31}$$

$$\mathbf{h}_2 = \boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{u}_2 \tag{2.32}$$

with correlation $Corr(\xi_2, \omega_2) = \mathbf{f}_2'\boldsymbol{\Sigma}_{XY}\mathbf{h} = \sqrt{\phi}_2$.

The remaining canonical variates and correlations are derived in a similar sequential manner until no further solutions can be found.

Way 2- Multivariately CCA may be approached from a multivariate setting as in [32] and [24] by looking to achieve a minimum in some least-squares sense. Let $\mathbf{F}_t = \begin{pmatrix}\mathbf{f}_1, & \mathbf{f}_2, & \ldots, & \mathbf{f}_t\end{pmatrix}$ and $\mathbf{H}_t = \begin{pmatrix}\mathbf{h}_1, & \mathbf{h}_2, & \ldots, & \mathbf{h}_t\end{pmatrix}$ so that $\boldsymbol{\xi}$ and $\boldsymbol{\omega}$ are now vector variates

$$\boldsymbol{\xi} = \mathbf{x}'\mathbf{F}_t \quad \boldsymbol{\omega} = \mathbf{y}'\mathbf{H}_t, \tag{2.33}$$

again assuming that $\mathbf{x}$ and $\mathbf{y}$ have zero means. The canonical weight matrices may be found by minimizing

$$\mathbb{E}\left[(\mathbf{Y}\mathbf{H}_t - \mathbf{X}\mathbf{F}_t)(\mathbf{Y}\mathbf{H}_t - \mathbf{X}\mathbf{F}_t)'\right] \tag{2.34}$$

for a given t where $Cov(\boldsymbol{\omega}) = \boldsymbol{\Sigma}_{\omega\omega} = \mathbf{H}'\boldsymbol{\Sigma}_{YY}\mathbf{H} = \mathbb{I}_t$.

To see this, first fix the matrix $\mathbf{H}$ and minimize the criterion (2.34):

$$\begin{aligned}
\min_{\mathbf{F}} \left\{ \mathbb{E}\left[(\mathbf{Y}\mathbf{H}_t - \mathbf{X}\mathbf{F}_t)(\mathbf{Y}\mathbf{H}_t - \mathbf{X}\mathbf{F}_t)' \right] \right\} &= tr\left[\mathbf{\Sigma}_{\omega\omega} - \mathbf{\Sigma}_{\omega X}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{X\omega} \right] \\
&\quad + \min_{\mathbf{F}} \left\{ tr\left[(\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{X\omega} - \mathbf{\Sigma}_{XX}^{1/2}\mathbf{F})(\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{X\omega} - \mathbf{\Sigma}_{XX}^{1/2}\mathbf{F})' \right] \right\} \\
&\geq tr\left[\mathbf{\Sigma}_{\omega\omega} - \mathbf{\Sigma}_{\omega X}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{X\omega} \right] \\
&= tr\left[\mathbf{H}'\mathbf{\Sigma}_{YY}\mathbf{H} - \mathbf{H}'\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{H} \right] \\
&= t - \sum_{j=1} \tilde{\rho}_j \qquad (2.35)
\end{aligned}$$

where $\tilde{\rho}_j$ is the j-th largest eigenvalue of $\mathbf{H}'\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{H} = \mathbf{H}'\mathbf{\Sigma}_{YY}^{1/2}\mathbf{R}\mathbf{\Sigma}_{YY}^{1/2}\mathbf{H}$ where

$$\mathbf{R} = \mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2} \qquad (2.36)$$

and $\mathbf{H}'\mathbf{\Sigma}_{YY}^{1/2}\mathbf{\Sigma}_{YY}^{1/2}\mathbf{H} = \mathbf{H}'\mathbf{\Sigma}_{YY}\mathbf{H} = \mathbb{I}_t$. Thus, by Theorem A.0.2, the Poincaré Separation Theorem:

$$t - \sum_{j=1} \tilde{\rho}_j \geq t - \sum_{j=1}^{t} \rho_j \qquad (2.37)$$

where ρ_j is the j-th largest eigenvalue of $\mathbf{R}$.

Using Theorem A.0.1,

$$\begin{aligned}
\mathbf{F}_t &= \mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{U}_t \qquad (2.38) \\
\mathbf{H}_t &= \mathbf{\Sigma}_{YY}^{-1/2}\mathbf{U}_t \qquad (2.39)
\end{aligned}$$

where $\mathbf{U}_t = \begin{pmatrix} \mathbf{u}_1, & \mathbf{u}_2, & \ldots, & \mathbf{u}_t \end{pmatrix}$ and $\mathbf{u}_k$ is the eigenvector corresponding to the k-th largest eigenvalue of

$$\mathbf{R} = \mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}. \qquad (2.40)$$

Notice that both the sequential method and the multivariate method yield identical formulas for finding the canonical correlations and the canonical variates provided that t is known. While CCA was initially established in the sequential way, the multivariate way is a more novel approach. It is far more convenient and quicker for matrix-based programming. However, the sample canonical correlations and variates have computational issues in the high-dimensional setting. To examine these, first the sample canonical correlations and variates are given for the large sample setting.

2.1.2 Sample Canonical Correlations and Variates

The sample canonical correlations and variates may be found using the data matrices $\mathbf{X}$ and $\mathbf{Y}$ which are $m \times p$ and $m \times n$ matrices, respectively. The canonical coefficient matrices $\mathbf{F}$ and $\mathbf{H}$ are easily estimated by $\mathbf{X}$ and $\mathbf{Y}$ using the ML sample covariances given at the beginning of the chapter in place of their theoretical versions, assuming $\widehat{\mathbf{\Sigma}}_{XX}$ and $\widehat{\mathbf{\Sigma}}_{YY}$ are nonsingular and invertible.

The sample estimates of the canonical weight matrices $\widehat{\mathbf{F}}_t$ and $\widehat{\mathbf{H}}_t$ are then given by

$$\widehat{\mathbf{F}}_t = \widehat{\mathbf{\Sigma}}_{XX}^{-1}\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_{YY}^{-1/2}\widehat{\mathbf{U}}_t \tag{2.41}$$

$$\widehat{\mathbf{H}}_t = \widehat{\mathbf{\Sigma}}_{YY}^{-1/2}\widehat{\mathbf{U}}_t \tag{2.42}$$

where $\widehat{\mathbf{U}}_t = \begin{pmatrix}\hat{\mathbf{u}}_1, & \hat{\mathbf{u}}_2, & \ldots, & \hat{\mathbf{u}}_t\end{pmatrix}$ and $\hat{\mathbf{u}}_k$ is the eigenvector corresponding to the k-th largest eigenvalue of

$$\widehat{\mathbf{R}} = \widehat{\mathbf{\Sigma}}_{YY}^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-1}\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_{YY}^{-1/2}. \tag{2.43}$$

The corresponding sample canonical variates are formed by

$$\widehat{\mathbf{\Xi}} = \mathbf{X}\widehat{\mathbf{F}} \tag{2.44}$$

$$\widehat{\mathbf{\Omega}} = \mathbf{Y}\widehat{\mathbf{H}} \tag{2.45}$$

where $\widehat{\mathbf{F}} = \begin{pmatrix}\widehat{\mathbf{f}}_1, & \widehat{\mathbf{f}}_2, & \ldots, & \widehat{\mathbf{f}}_t\end{pmatrix}$ and $\widehat{\mathbf{H}} = \begin{pmatrix}\widehat{\mathbf{h}}_1, & \widehat{\mathbf{h}}_2, & \ldots, & \widehat{\mathbf{h}}_t\end{pmatrix}$. Now, $\widehat{\mathbf{\Xi}} = \begin{pmatrix}\widehat{\boldsymbol{\xi}}_1, & \widehat{\boldsymbol{\xi}}_2, & \ldots, & \widehat{\boldsymbol{\xi}}_t\end{pmatrix}$ and $\widehat{\mathbf{\Omega}} = \begin{pmatrix}\widehat{\boldsymbol{\omega}}_1, & \widehat{\boldsymbol{\omega}}_2, & \ldots, & \widehat{\boldsymbol{\omega}}_t\end{pmatrix}$ are matrices both of size $m \times t$ where the columns make up the canonical variates. Note that each estimated canonical variate pair $(\widehat{\boldsymbol{\xi}}_k, \widehat{\boldsymbol{\omega}}_k)$ has m rows corresponding to the m observations.

The corresponding canonical correlations may be calculated for each canonical variate pair as

$$\hat{\rho}_k = \frac{\hat{\mathbf{f}}_k'\widehat{\mathbf{\Sigma}}_{XY}\hat{\mathbf{h}}_k}{(\hat{\mathbf{f}}_k'\widehat{\mathbf{\Sigma}}_{XX}\hat{\mathbf{f}}_k)^{1/2}(\hat{\mathbf{h}}_k'\widehat{\mathbf{\Sigma}}_{YY}\hat{\mathbf{h}}_k)^{1/2}} \tag{2.46}$$

for $k = 1, \cdots, t$.

Notice that the nonsingularity requirement on $\widehat{\mathbf{\Sigma}}_{XX}$ and $\widehat{\mathbf{\Sigma}}_{YY}$ is needed so that these matrices are invertible. In the settings where they are not invertible (e.g. high-dimensional), Regularized Canonical Correlation Analysis (RCCA) is often employed. However, there exist certain computational limitations making it less than ideal. It is detailed next.

2.1.3 Regularized Canonical Correlation Analysis

The theory in CCA holds if t is given and when the sample is large compared to the number of predictors and responses. However, when the sample size is less than the maximum of the number of predictors and responses or is only a bit larger, the inverses of $\widehat{\mathbf{\Sigma}}_{XX}$ and $\widehat{\mathbf{\Sigma}}_{YY}$ do not exist or are ill conditioned. In addition to this, the weighted criterion in (2.34) will be close to 0, and the canonical correlations will be nearly 1 as the number of variables increases. Due to this, Eaton and Perlman advised that the CCA should only be performed if $m \geq n + p + 1$ [11].

Regularized Canonical Correlation Analysis (RCCA) offers one solution to these problems. It introduces regularization parameters that essentially add just enough to the diagonal entries of the estimators to make them invertible. This technique was proposed first by Vinod [41] and developed by Leurgans *et al.* [27] and is similar to the application of regularization in ridge regression.

Estimators of $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$ are required and are typically the sample covariance matrices $\widehat{\mathbf{\Sigma}}_{XX}$ and $\widehat{\mathbf{\Sigma}}_{YY}$. These in RCCA are replaced by

$$\widetilde{\mathbf{\Sigma}}_{XX} = \widehat{\mathbf{\Sigma}}_{XX} + \delta_x \mathbb{I}_p \tag{2.47}$$

$$\widetilde{\mathbf{\Sigma}}_{YY} = \widehat{\mathbf{\Sigma}}_{YY} + \delta_y \mathbb{I}_n \tag{2.48}$$

where δ_x and δ_y are regularization parameters.

These regularization parameters may be determined using cross-validation as was proposed by González *et al.* [16] by the following steps.

1. Randomly divide the m observations into $i = 1, \ldots, \nu$ folds. Let $\boldsymbol{\delta} = \begin{pmatrix} \delta_x & \delta_y \end{pmatrix}$.

2. Compute the first canonical correlation $\rho_\delta^{(-i)}$ for a particular value $\boldsymbol{\delta}$ without the observations from the i-th fold along with the corresponding first canonical coefficients $\mathbf{f}_\delta^{(-i)}$ and $\mathbf{h}_\delta^{(-i)}$.

3. Find the cross-validation score (CV) for $\boldsymbol{\delta}$ as

$$CV(\delta_x, \delta_y) = Corr\left(\left\{ \mathbf{x}_i' \mathbf{f}_\delta^{(-i)} \right\}_{i=1}^{\nu}, \left\{ \mathbf{y}_i' \mathbf{h}_\delta^{(-i)} \right\}_{i=1}^{\nu} \right). \tag{2.49}$$

4. The final estimates of $\boldsymbol{\delta} = \begin{pmatrix} \delta_x & \delta_y \end{pmatrix}$ are chosen from a grid of values that maximize (2.49) so that

$$\widehat{\boldsymbol{\delta}} = \begin{pmatrix} \widehat{\delta}_1 & \widehat{\delta}_2 \end{pmatrix} = \underset{\delta_x, \delta_y}{\operatorname{argmax}}\ CV(\delta_x, \delta_y). \tag{2.50}$$

□

While this does provide one option to handle the high-dimensional setting, there are some serious setbacks. First of all, cross-validating a 2-dimensional grid of regularization parameters can be computationally expensive and time consuming. Second of all, the cross-validation score only considers the first canonical variate pair and completely ignores the others. This is less than ideal and an alternative to this will be offered in the coming chapters based upon the Reduced-Rank Regression (RRR) structure through the connection provided directly.

2.1.4 Relationship to Reduced-Rank Regression

Reduced-Rank Regression (RRR) has a direct relationship with CCA. This fact will be used later when connecting CCA to WRSC. To establish this, consider the model in (2.4) with the same given assumptions. RRR produces estimators of $\mathbf{A}$ of reduced-rank. That is, suppose $r(\mathbf{A}) \leq min(n, p)$. Then, $\mathbf{A}$ may be expressed as the decomposition of two matrices

$$\mathbf{A} = \mathbf{CD}, \tag{2.51}$$

and the model in (2.4) may be written as the RRR model

$$\mathbf{Y} = \mathbf{XCD} + \mathbf{E}. \tag{2.52}$$

In their most basic form, estimators of the coefficient matrix $\mathbf{A}$ minimize the weighted sum of squares with a given rank, say k, with a positive-definite matrix of weights $\mathbf{\Gamma}$. For clarity, these reduced-rank estimators of given rank k are denoted here as $\widehat{\mathbf{B}}_k$. For a positive-definite weight matrix $\mathbf{\Gamma}$, the estimator of $\mathbf{A}$ of given rank k, denoted as $\widehat{\mathbf{B}}_k$, may be found by

$$\begin{aligned} &\min_{\mathbf{B}} \left|\left|(\mathbf{Y}-\mathbf{XB})\mathbf{\Gamma}^{1/2}\right|\right|_F^2 \\ &\text{s.t. } r(\mathbf{B}) \leq k \end{aligned} \tag{2.53}$$

as a *weighted constrained problem.*

To compute $\widehat{\mathbf{B}}_k$, a computationally efficient procedure suggested by Reinsel and Velu [32] may be used with the ML sample covariances $\widehat{\mathbf{\Sigma}}_{XX}$, $\widehat{\mathbf{\Sigma}}_{XY} = \widehat{\mathbf{\Sigma}}'_{YX}$, and $\widehat{\mathbf{\Sigma}}_{YY}$. For now, assume that $\widehat{\mathbf{\Sigma}}_{XX}$ is nonsingular. The steps for computing $\widehat{\mathbf{B}}_k$ of given rank k are then:

1. Find the (normalized) eigenvectors $\widehat{\mathbf{V}}_k = \left(\hat{\mathbf{v}}_1, \quad \hat{\mathbf{v}}_2, \quad \ldots, \hat{\mathbf{v}}_k\right)$, where $\hat{\mathbf{v}}_j$ is the eigenvector corresponding to the j-th largest eigenvalue of the symmetric matrix

$$\widehat{\mathbf{R}} = \mathbf{\Gamma}^{1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-1}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Gamma}^{1/2}. \tag{2.54}$$

2. Calculate the (full-rank) least-squares estimator $\widehat{\mathbf{B}} = \widehat{\mathbf{\Sigma}}_{XX}^{-1}\widehat{\mathbf{\Sigma}}_{XY}$. Then, form $\widehat{\mathbf{C}} = \widehat{\mathbf{B}}\mathbf{\Gamma}^{1/2}\widehat{\mathbf{V}}_k$ and $\widehat{\mathbf{D}} = \widehat{\mathbf{V}}'_k\mathbf{\Gamma}^{-1/2}$.

3. Compute the estimator $\widehat{\mathbf{B}}_k = \widehat{\mathbf{C}}_k\widehat{\mathbf{D}}_k$ where $\widehat{\mathbf{C}}_k = \widehat{\mathbf{C}}[, 1:k]$ denotes the matrix $\widehat{\mathbf{C}}$ retaining only its first k columns and $\widehat{\mathbf{D}}_k = \widehat{\mathbf{D}}[1:k,]$ denotes the matrix $\widehat{\mathbf{D}}$ retaining only its first k rows. □

These steps give the estimated coefficient matrix $\widehat{\mathbf{B}}_k$ of rank k as

$$\widehat{\mathbf{B}}_k = \widehat{\mathbf{C}}_k\widehat{\mathbf{D}}_k = \widehat{\mathbf{\Sigma}}_{XX}^{-1}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Gamma}^{1/2}\left(\sum_{j=1}^{k}\hat{\mathbf{v}}_j\hat{\mathbf{v}'}_j\right)\mathbf{\Gamma}^{-1/2} \tag{2.55}$$

where

$$\widehat{\mathbf{C}}_k = \widehat{\mathbf{\Sigma}}_{XX}^{-1}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Gamma}^{1/2}\widehat{\mathbf{V}}_k \tag{2.56}$$

$$\widehat{\mathbf{D}}_k = \widehat{\mathbf{V}}'_k\mathbf{\Gamma}^{-1/2}. \tag{2.57}$$

The two matrices $\widehat{\mathbf{C}}_k$ and $\widehat{\mathbf{D}}_k$ constructed above yield the unique decomposition of $\widehat{\mathbf{B}}_k$ with the following properties:

(i) $\widehat{\mathbf{D}}_k\mathbf{\Gamma}\widehat{\mathbf{D}}'_k = \mathbb{I}_k$,

(ii) $\widehat{\mathbf{C}}'_k(\mathbf{X}'\mathbf{X})\widehat{\mathbf{C}}_k$ is a diagonal matrix

provided that $\mathbf{\Gamma}$ is positive-definite.

Then, following, e.g. Izenman [24], finding the first $t = k$ canonical variates reduces to first minimizing the criterion in (2.53) with $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$ (assuming for now that $\widehat{\mathbf{\Sigma}}_{YY}$ is

nonsingular). The two canonical weight matrices needed to construct the canonical variates are given by the decomposition of $\widehat{\mathbf{B}}_k$. That is, by Reinsel and Velu [32], the canonical variates $\widehat{\Xi}$ and $\widehat{\mathbf{\Omega}}$ may be constructed via the matrices $\widehat{\mathbf{C}}_k$ and $\widehat{\mathbf{D}}_k$ as

$$\widehat{\Xi} = \mathbf{X}\widehat{\mathbf{C}}_k \tag{2.58}$$

$$\widehat{\mathbf{\Omega}} = \mathbf{Y}\left(\widehat{\mathbf{D}}_k\right)^{-}. \tag{2.59}$$

That is,

$$\widehat{\mathbf{F}}_t = \widehat{\mathbf{C}}_k \tag{2.60}$$

and

$$\begin{aligned} \widehat{\mathbf{H}}_t\widehat{\mathbf{D}}_k\widehat{\mathbf{H}}_t &= \widehat{\mathbf{H}}_t \\ \widehat{\mathbf{D}}_k\widehat{\mathbf{H}}_t\widehat{\mathbf{D}}_k &= \widehat{\mathbf{D}}_k \end{aligned}$$

so that,

$$\widehat{\mathbf{H}}_t = \left(\widehat{\mathbf{D}}_k\right)^{-} \tag{2.61}$$

where the number of canonical relationships $t = k$, the rank of the coefficient matrix.

Now that the two approaches of CCA have been detailed, a brief note will be made about the additional descriptive measures that are typically found for inferential purposes in CCA. These are included here as they will be utilized in the HIV/neurocognitive application.

2.1.5 Additional Descriptive Measures

While the previous sections have addressed the computational aspects of CCA, it is left to be said how interpretations may be derived. Interpretation of canonical correlations and canonical variates is highly subjective. However, many additional descriptive measures aid in identifying relationships. Basic interpretation includes examination of the magnitude and sign of the estimated canonical correlations and the estimated canonical coefficients. However, the analysis may also include various additional calculations. These calculations include canonical loadings, canonical cross-loadings, and the canonical correlations. Their squares provide percentage measures of shared variance, contributed variance, and explained variance. Additional measures also include the redundancy indexes. An extensive discussion of these is found in [17], and the most relevant definitions are provided directly.

Definition 2.1.1. *Let $L_x(k,j)$ and $L_y(k,j)$ denote the k-th **canonical loadings** between the k-th canonical variates $\hat{\boldsymbol{\xi}}_k$ and $\hat{\boldsymbol{\omega}}_k$ and the j-th variable of the predictor and response set, respectively. Then,*

$$L_x(k,j) = Corr(\hat{\boldsymbol{\xi}}_k, \mathbf{x}_j) \tag{2.62}$$

is the simple linear correlation between the k-th predictor canonical variate $\widehat{\boldsymbol{\xi}}_k$ and the j-th predictor variable $\mathbf{x}_j$. Similarly,

$$L_y(k,j) = Corr(\hat{\boldsymbol{\omega}}_k, \mathbf{y}_j) \tag{2.63}$$

is the simple linear correlation between the k-th response canonical variate $\widehat{\boldsymbol{\omega}}_k$ and the j-th response variable $\mathbf{y}_j$.

The *percentage of shared variance* for each variable with its canonical variate is simply the square of the canonical loadings. This may be thought of as the amount of variation in each variable explained by its respective canonical variate. Here, the relationship of each variable with its own canonical variate is isolated.

Definition 2.1.2. *Let E_k denote the* ***percentage of explained variance*** *of the k-th canonical variate. Then,*

$$E_k = \hat{\rho}_k^2, \tag{2.64}$$

the k-th canonical correlation squared.

The *amount of explained variance* is the percentage of variance in one canonical variate that can be explained by the other canonical variate, treating the two canonical variates as a whole.

Definition 2.1.3. *The* ***redundancy index*** *is*

$$k\text{-th Redundancy Index of the Predictor Set} = \frac{1}{p}\sum_{j=1}^{p} L_x^2(k,j) \times E_k \tag{2.65}$$

$$k\text{-th Redundancy Index of the Response Set} = \frac{1}{n}\sum_{j=1}^{n} L_y^2(k,j) \times E_k, \tag{2.66}$$

the product of the average amount of shared variance and the amount of explained variance.

This was developed by Stewart and Love as an overall measure between each pair of canonical variates [39]. In order to have a high redundancy, there must exist a high canonical correlation *and* a high degree of shared variance. Typically, a researcher is only interested in the variance extracted from the response set as a measure of the predictive ability of the estimated canonical function, but the variance extracted from the predictor set may also provide useful insight.

One caution in using these types of descriptive measures is that they are subject to a great deal of variability from one sample to another, the canonical coefficients even more so than these measures. Problems such as suppression and multicollinearity may influence these types of statistics and make the interpretations invalid. Thus, the *canonical cross-loadings* are also useful in interpretation as they are the most resistant to these problems.

Definition 2.1.4. *Let $C_x(k,j)$ and $C_y(k,j)$ denote the* ***canonical cross-loading*** *of a predictor variable and a response variable, respectively. Then,*

$$C_x(k,j) = Corr(\hat{\boldsymbol{\omega}}_k, \mathbf{x}_j), \tag{2.67}$$

the simple linear correlation between the k-th response canonical variate $\hat{\boldsymbol{\omega}}_k$ and the j-th predictor variable $\mathbf{x}_j$ and

$$C_y(k,j) = Corr(\hat{\boldsymbol{\xi}}_k, \mathbf{y}_j), \tag{2.68}$$

the simple linear correlation between the k-th predictor canonical variate $\hat{\boldsymbol{\xi}}_k$ and the j-th response variable $\mathbf{y}_j$.

The squares of these give the *percentages of contributed variance* a variable gives to its opposite or cross canonical variate. This provides a better measure of the relationship between the predictor and response sets as the percentage of contributed variance isolates the influence of one variable across to the other canonical variate and helps in individually examining the predictor-response relationship.

2.2 Rank Selection Criterion (RSC)

The next major topic to introduce is the Rank Selection Criterion (RSC), a novel approach by Bunea *et al.* for determining reduced-rank estimators in a data-adaptive way [8]. It has shown to be a computationally elegant solution to the long-standing rank selection problem and treats it in a penalized fashion. Because it is valid for any number of p predictors, n responses, and m observations, there is a wide range of applications. RSC is outlined here, but see [8] for the theoretical details.

Following from model (2.4), make the additional assumption that the covariance of the error terms is $\mathbf{\Sigma}_e = \sigma^2 \mathbb{I}_n$. The $p \times n$ estimator of $\mathbf{A}$ is found by minimizing the Frobenius norm of the model fit with a penalty term. That is, the *penalized criterion* is

$$\widehat{\mathbf{A}} = \underset{\mathbf{B}}{\operatorname{argmin}} \left\{ \Big|\Big| \mathbf{Y} - \mathbf{XB} \Big|\Big|_F^2 + \mu r(\mathbf{B}) \right\} \tag{2.69}$$

where $\mu > 0$ is the tuning parameter. Note that the assumption on the error matrix is required in order to obtain precise numerical constants for the penalty term in the criterion. The following two-step procedure may be used to compute $\widehat{\mathbf{A}}$:

Step 1: Let $\widehat{\mathbf{B}}_k$ denote the minimizer of $\|\mathbf{Y} - \mathbf{XB}\|_F^2$ over all matrices $\mathbf{B}$ of fixed rank k. This may be done by the following steps:

1. Find the (normalized) eigenvectors $\widehat{\mathbf{V}}_k = \big(\hat{\mathbf{v}}_1, \ \hat{\mathbf{v}}_2, \ \ldots, \hat{\mathbf{v}}_k\big)$, where $\hat{\mathbf{v}}_j$ is the eigenvector corresponding to the j-th largest eigenvalue of the symmetric matrix

$$\widehat{\mathbf{R}} = \widehat{\mathbf{\Sigma}}_{YX} \widehat{\mathbf{\Sigma}}_{XX}^{-} \widehat{\mathbf{\Sigma}}_{XY}. \tag{2.70}$$

2. Calculate the (full-rank) least-squares estimator $\widehat{\mathbf{B}} = \widehat{\mathbf{\Sigma}}_{XX}^{-} \widehat{\mathbf{\Sigma}}_{XY}$. Then, form $\widehat{\mathbf{C}} = \widehat{\mathbf{B}} \widehat{\mathbf{V}}_k$ and $\widehat{\mathbf{D}} = \widehat{\mathbf{V}}_k'$.
3. Compute the estimator $\widehat{\mathbf{B}}_k = \widehat{\mathbf{C}}_k \widehat{\mathbf{D}}_k$ where $\widehat{\mathbf{C}}_k = \widehat{\mathbf{C}}[, 1:k]$ denotes the matrix $\widehat{\mathbf{C}}$ retaining only its first k columns and $\widehat{\mathbf{D}}_k = \widehat{\mathbf{D}}[1:k,]$ denotes the matrix $\widehat{\mathbf{D}}$ retaining only its first k rows.

Step 2: The final estimator is $\widehat{\mathbf{A}} = \widehat{\mathbf{B}}_{\hat{k}}$ where $\hat{k}$ is given by the following proposition:

Proposition 2.2.1. *Let $\lambda_1 \geq \lambda_2 \geq ...$ denote the ordered eigenvalues of $\widehat{\mathbf{\Sigma}}_{YX} \widehat{\mathbf{\Sigma}}_{XX}^{-} \widehat{\mathbf{\Sigma}}_{XY}$. Then,*

$$\hat{k} = max\{k : \lambda_k \geq \frac{\mu}{m}\} \tag{2.71}$$

where μ is the tuning parameter.

Proof. See Proposition 1 of [8]. □

Thus by definition, $\hat{k}$ is the number of eigenvalues of $\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}$ that exceeds $\frac{\mu}{m}$. □

Note that *Step 1* is essentially the same computing the steps in Section 2.1.4 with $\mathbf{\Gamma} = \mathbb{I}_n$. *Bunea et al.* demonstrate, both empirically and theoretically, that even if $\widehat{\mathbf{\Sigma}}_{XX}$ is singular and the Moore-Penrose generalized inverse is used, as is written above, consistent rank estimation is still possible. That is, $\widehat{\mathbf{B}}_{\hat{k}}$, as found above, is the globally optimal solution to this problem regardless of the sample size.

2.2.1 Tuning Parameter

If an additional assumption of Gaussian errors is employed so that $\mathbf{E}$ has independent $N(0, \sigma^2)$ entries, the penalty term $pen(\mathbf{B}) = \mu r(\mathbf{B})$ may be further characterized by the following corollary.

Corollary 2.2.2. *Assume that* $\mathbf{E}$ *has independent* $N(0, \sigma^2)$ *entries. For any* $\theta > 0$*, the penalty term may be defined up to constants as*

$$pen(\mathbf{B}) = (1+\theta)(1+\xi)^2(\sqrt{n}+\sqrt{q})^2\sigma^2 r(\mathbf{B}) \tag{2.72}$$

with $\theta, \xi > 0$ *arbitrary and* $q = r(\mathbf{X})$*.*

Proof. See Corollary 4 of [8]. □

This ensures good rank selection and prediction performance provided that the tuning parameter μ is just a bit larger than $(\sqrt{n}+\sqrt{q})^2$.

However, the practical case is when the variance σ^2 is unknown. Then, the penalty term may be defined using an estimated variance so that for any $\theta, \xi > 0$, and $0 < \delta < 1$,

$$pen(\mathbf{B}) = \frac{(1+\theta)}{1-\delta}(1+\xi)^2(\sqrt{n}+\sqrt{q})^2 S^2 r(\mathbf{B}) \tag{2.73}$$

where S^2 is the unbiased estimator

$$S^2 = \frac{1}{(m-q)n}\left|\left|\mathbf{Y} - \mathbf{P}\mathbf{Y}\right|\right|_F^2 \tag{2.74}$$

under the assumption that $q < m$.

The tuning parameter μ may be determined in one of two ways: (1) utilizing cross-validation (typically 5-fold cross-validation is sufficient) or (2) using the data-dependent closed form as suggested by Bunea *et al.* The final closed form of the so-called adaptive tuning parameter is

$$\mu_{adap} = 2S^2(n+q). \tag{2.75}$$

Bunea *et al.* have shown through extensive simulation experiments that constants slightly larger or smaller than 2 give similar results and that μ_{adap} has demonstrated excellent performance in both cases of large sample and high-dimension. In practice, the choice

between cross-validation and the adaptive tuning parameter may be based upon the one that yields the best performance through validation. That is, the one that gives the smallest mean squared error (MSE) when calculated on an independent part of the data that is set aside prior to estimation ought to be used.

2.2.2 Dimension Reduction

RSC may also be applied as a dimension reduction technique of the predictor space. The two properties given in Section 2.1.4 provide that new, uncorrelated predictors may be found. Recall, the final estimator of the coefficient matrix is $\widehat{\mathbf{A}} = \widehat{\mathbf{B}}_{\hat{k}} = \widehat{\mathbf{C}}_{\hat{k}}\widehat{\mathbf{D}}_{\hat{k}}$ with $\mathbf{\Gamma} = \mathbb{I}_n$. To see this, consider the predicted response matrix rewritten as

$$\begin{aligned} \widehat{\mathbf{Y}} &= \mathbf{X}\widehat{\mathbf{A}} \\ &= \mathbf{X}\widehat{\mathbf{C}}_{\hat{k}}\widehat{\mathbf{D}}_{\hat{k}} \\ &= \mathbf{Q}\widehat{\mathbf{D}}_{\hat{k}}. \end{aligned} \tag{2.76}$$

The $\hat{k}$ columns of $\mathbf{Q} = \mathbf{X}\widehat{\mathbf{C}}_{\hat{k}}$ are orthogonal by construction and may be regarded as new predictors. Since typically $\hat{k}$ is much smaller than p, this procedure offers significant dimension reduction.

The previous considerations and optimality results of the RSC estimator are well established under the assumption that $\mathbf{\Sigma}_e = \sigma^2\mathbb{I}_n$, treating the weight matrix as $\mathbf{\Gamma} = \mathbb{I}_n$. In Chapter 3, this will be generalized for any $\mathbf{\Sigma}_e$, forming the Weighted Rank Selection Criterion (WRSC). This will be developed to connect WRSC to RSC to form ACCA. Yet, the final piece of background that will do so is based upon the choice of the weight matrix $\mathbf{\Gamma}$. The population version is detailed next.

2.3 Choice of Weight Matrix

The choice of the weight matrix $\mathbf{\Gamma}$ is very important in Reduced-Rank Regression (RRR). This is carefully examined from the population point of view as documented by [32]. In the Rank Selection Criterion (RSC), treatment of such is just $\mathbf{\Gamma} = \mathbb{I}_n$. However, in Canonical Correlation Analysis (CCA), CCA is recovered through RRR by setting $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$. Here, two very popular choices of the weight matrix will be discussed: $\mathbf{\Sigma}_e^{-1} = (\mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY})^{-1}$ and $\mathbf{\Sigma}_{YY}^{-1}$. Note that the sample version of $\mathbf{\Sigma}_e$ is simply the residual covariance, denoted as $\widehat{\mathbf{\Sigma}}_r$. Treatment of the sample estimators $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$ and $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_r^{-1} = \widehat{\mathbf{\Sigma}}_{YY} - \widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}$ in the high-dimensional setting are lacking however. These are examined carefully in the next chapter as they become extremely important in the establishment of ACCA and WRSC.

First, assume that $\mathbf{\Gamma}$ is positive-definite, i.e. $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ is positive-definite and $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$ is positive-definite. Then, the relationship between the the two weight matrices is given by the following lemmas, the first for the relationship between eigenvalues and the second for the relationship between eigenvectors.

Lemma 2.3.1. *Let ρ_j be the j-th largest eigenvalue of $\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}$ and let λ_j be the j-th largest eigenvalue of $\mathbf{\Sigma}_e^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_e^{-1/2}$. Then, ρ_j and λ_j are related*

by

$$\lambda_j = \frac{\rho_j}{1-\rho_j} \tag{2.77}$$

and

$$\rho_j = \frac{\lambda_j}{1+\lambda_j}. \tag{2.78}$$

Proof. See [32]. □

Lemma 2.3.2. *Let ρ_j be the j-th largest eigenvalue of $\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}$ with $\mathbf{v}_j$ being the corresponding eigenvector. Let λ_j be the j-th largest eigenvalue of $\boldsymbol{\Sigma}_e^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_e^{-1/2}$ with the corresponding eigenvector $\mathbf{v}_j^*$. Denote the eigenvectors as $\mathbf{V}_k = \begin{pmatrix}\mathbf{v}_1, & \mathbf{v}_2, & \ldots, & \mathbf{v}_k\end{pmatrix}$ and $\mathbf{V}_k^* = \begin{pmatrix}\mathbf{v}_1^*, & \mathbf{v}_2^*, & \ldots, & \mathbf{v}_k^*\end{pmatrix}$, respectively. Then, the eigenvectors are related via*

$$\mathbf{V}_k = \boldsymbol{\Sigma}_{YY}^{1/2}\boldsymbol{\Sigma}_e^{-1/2}\mathbf{V}_k^*(1-\boldsymbol{\Lambda})^{1/2} \tag{2.79}$$

or

$$\mathbf{V}_k^* = \boldsymbol{\Sigma}_e^{1/2}\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{V}_k(1-\boldsymbol{\Lambda})^{-1/2} \tag{2.80}$$

where $\boldsymbol{\Lambda} = diag(\rho_1, \rho_2, \ldots, \rho_k)$.

Proof. See [32]. □

These lemmas are used to establish the following relationship.

Lemma 2.3.3. *For $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_e^{-1}$, let the estimator of the coefficient matrix be denoted as $\mathbf{B}_k^* = \mathbf{C}_k^*\mathbf{D}_k^*$ of rank k. For $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{YY}^{-1}$, let the estimator of the coefficient matrix be denoted as $\mathbf{B}_k = \mathbf{C}_k\mathbf{D}_k$. Then,*

$$\mathbf{B}_k^* = \mathbf{C}_k^*\mathbf{D}_k^* \equiv \mathbf{C}_k\mathbf{D}_k = \mathbf{B}_k \tag{2.81}$$

even though

$$\mathbf{C}_k^* \neq \mathbf{C}_k \quad \textit{and} \quad \mathbf{D}_k^* \neq \mathbf{D}_k. \tag{2.82}$$

Proof. This proof simply requires the use of the relationship established in Lemma 2.3.2 so that

$$\begin{aligned}
\mathbf{C}_k^* &= \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_e^{-1/2}\boldsymbol{\Sigma}_e^{1/2}\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{V}_k(\mathbb{I}-\boldsymbol{\Lambda}^2)^{-1/2} \\
&= \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{V}_k(\mathbb{I}-\boldsymbol{\Lambda}^2)^{-1/2} \\
&= \mathbf{C}_k(\mathbb{I}-\boldsymbol{\Lambda}^2)^{-1/2}
\end{aligned} \tag{2.83}$$

and

$$\begin{aligned}
\mathbf{D}_k^* &= \left(\boldsymbol{\Sigma}_r^{1/2}\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{V}_k(\mathbb{I}-\boldsymbol{\Lambda}^2)^{-1/2}\right)'\boldsymbol{\Sigma}_r^{1/2} \\
&= (\mathbb{I}-\boldsymbol{\Lambda}^2)^{-1/2}\mathbf{V}_k'\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_r \\
&= (\mathbb{I}-\boldsymbol{\Lambda}^2)^{-1/2}\mathbf{V}_k'\boldsymbol{\Sigma}_{YY}^{-1/2}(\boldsymbol{\Sigma}_{YY}-\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}) \\
&= (\mathbb{I}-\boldsymbol{\Lambda}^2)^{-1/2}\mathbf{V}_k'(\boldsymbol{\Sigma}_{YY}^{1/2}-\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}) \\
&= (\mathbb{I}-\boldsymbol{\Lambda}^2)^{-1/2}\mathbf{V}_k'(\mathbb{I}-\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2})\boldsymbol{\Sigma}_{YY}^{1/2} \\
&= (\mathbb{I}-\boldsymbol{\Lambda}^2)^{1/2}\mathbf{V}_k'\boldsymbol{\Sigma}_{YY}^{1/2} \\
&= (\mathbb{I}-\boldsymbol{\Lambda}^2)^{1/2}\mathbf{D}_k.
\end{aligned} \tag{2.84}$$

So even though

$$\mathbf{C}_k \neq \mathbf{C}_k^* \qquad\qquad \mathbf{D}_k \neq \mathbf{D}_k^*, \tag{2.85}$$

the following still holds.

$$\begin{aligned} \mathbf{C}_k^*\mathbf{D}_k^* &= \mathbf{C}_k(\mathbb{I} - \mathbf{\Lambda}^2)^{-1/2}(\mathbb{I} - \mathbf{\Lambda}^2)^{1/2}\mathbf{D}_k \\ &= \mathbf{C}_k\mathbf{D}_k. \end{aligned} \tag{2.86}$$

□

CHAPTER 3

WEIGHTED METHODOLOGY

Adaptive Canonical Correlation Analysis (ACCA) is developed here from the weighted version of the Rank Selection Criterion (RSC), Weighted Rank Selection Criterion (WRSC). ACCA adds the additional step to standard CCA in which the number of significant canonical correlations (relationships) t is data-adaptively estimated.

To develop these ideas, first recall that CCA is simply RRR with $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$. What was noted in Section 2.3 is that the eigenvalues from

$$\mathbf{\Gamma}^{1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Gamma}^{1/2} \tag{3.1}$$

for $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ and $\mathbf{\Gamma} = \mathbf{\Sigma}_{e}^{-1} = \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}$ are related directly through Lemma 2.3.1. On the other hand, RSC is simply RRR working under the assumption that $\mathbf{\Sigma}_e = \sigma^2\mathbb{I}_n$ and that $\mathbf{\Gamma} = \mathbb{I}_n$, and the estimated rank is characterized by the number of eigenvalues of the above with $\mathbf{\Gamma} = \mathbb{I}_n$.

However, if no assumption is made on $\mathbf{\Sigma}_e$ and the correlation in the model could be removed by a "decorrelator" or weight matrix, then RSC could still be applied to estimate the rank. If $\mathbf{\Sigma}_e$ were known, then the appropriate weight matrix would simply be $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$. This will be clarified next. The relationship to CCA is then direct by the concepts given in Section 2.3, and estimating the rank and the coefficient estimator in WRSC is the same as estimating the number of significant canonical relationships $\hat{t}$ and the canonical weights themselves in CCA.

This, however, invites a variety of issues in practice, including whether the continued use of an adaptive tuning parameter may be justified and the invertibility issues that arise from the high-dimensional setting. The typical estimator of $\mathbf{\Sigma}_e$ is the sample residual covariance $\widehat{\mathbf{\Sigma}}_r$, but as noted before, $\widehat{\mathbf{\Sigma}}_r$ is only a good estimator in the large sample setting and is always singular. If only considering the large sample setting, $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$ may be used in place of $\widehat{\mathbf{\Sigma}}_r^{-}$ as the relationship established in Section 2.3 holds from the population version to the sample version.

When the weight matrix is singular, more delicate handling is required as $\widehat{\mathbf{\Sigma}}_r$ is no longer a good estimator of $\mathbf{\Sigma}_e$ and thus, the connection with $\widehat{\mathbf{\Sigma}}_{YY}$ no longer holds. While theoretical properties are not defined here for such a setting, there is a number of good reasons why immediate alternatives of $\widehat{\mathbf{\Sigma}}_r$ are not computationally ideal (such as the Moore-Penrose pseudoinverse or a regularized version of the residual covariance). An alternative proposed is $\widehat{\mathbf{\Sigma}}_{YY}^{-}$ or a regularized version thereof. In fact, it will be demonstrated that the

relationship established in Lemma 2.3.1 holds for the regularized sample versions of the two weight matrices. In Chapter 5, it will be empirically shown that the choice of $\mathbf{\Gamma}$ as some version of $\widehat{\mathbf{\Sigma}}_{YY}$ still provides good rank recovery and coefficient estimation. A more direct way to bypass these issues will be advocated in Chapter 6 with the addition of a variables selection step.

3.1 Weighted Rank Selection Criterion

Recall that the general multivariate response model is

$$\mathbf{Y} = \mathbf{XA} + \mathbf{E} \tag{3.2}$$

where $\mathbf{Y}$ is $m \times n$, $\mathbf{X}$ is $m \times p$, and $\mathbf{A}$ is $p \times n$ for m observations, p predictor variables, and n response variables. Assume for now that

A.1. The matrices $\mathbf{X}$ and $\mathbf{Y}$ have been centered by their column means.

A.2. The error terms are distributed with mean zero and covariance $\mathbf{\Sigma}_e$ with independent rows.

A.3. The weight matrix $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$ is positive-definite (p.d.) unless otherwise noted.

If $\mathbf{\Sigma}_e$ were known, then the original model (3.2) could be immediately transformed into a truly decorrelated model

$$\mathbf{Z} = \mathbf{XA}_1 + \mathbf{T} \tag{3.3}$$

where $\mathbf{Z} = \mathbf{Y}\mathbf{\Sigma}_e^{-1/2}$, $\mathbf{A}_1 = \mathbf{A}\mathbf{\Sigma}_e^{-1/2}$, and $\mathbf{T} = \mathbf{E}\mathbf{\Sigma}_e^{-1/2}$. The new error matrix $\mathbf{T}$ is a matrix of m independent realizations of the generic vector $\mathbf{t}$ where $Cov(\mathbf{t}) = \sigma^2 \mathbb{I}_n$ has uncorrelated errors.

If RSC is applied to (3.3), the penalized criterion may be written as

$$\min_{\mathbf{M}} \left\{ \left|\left| \mathbf{Z} - \mathbf{XM} \right|\right|_F^2 + \mu_1 r(\mathbf{M}) \right\} \tag{3.4}$$

where $\mu_1 > 0$ is the tuning parameter. Denote the estimator of $\mathbf{A}_1$ of rank $\hat{k}$ recovered from this RSC application as $\widehat{\mathbf{M}}_{\hat{k}}$. Given any $\mathbf{B}$, note $\mathbf{M} = \mathbf{B}\mathbf{\Sigma}_e^{-1/2}$, then

$$\begin{aligned} \left|\left| (\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}_e^{-1/2} \right|\right|_F^2 + \mu_1 r(\mathbf{B}) &\geq \left|\left| \mathbf{Z} - \mathbf{XM} \right|\right|_F^2 + \mu_1 r(\mathbf{M}) \\ &\geq \left|\left| \mathbf{Z} - \mathbf{X}\widehat{\mathbf{M}}_{\hat{k}} \right|\right|_F^2 + \mu_1 r(\widehat{\mathbf{M}}_{\hat{k}}) \end{aligned} \tag{3.5}$$

where $r(\mathbf{M}) \leq r(\mathbf{B})$. The final global minimizer is recovered from $\widehat{\mathbf{M}}_{\hat{k}}$ as

$$\widehat{\mathbf{B}}_{\hat{k}} = \widehat{\mathbf{M}}_{\hat{k}} \mathbf{\Sigma}_e^{1/2}. \tag{3.6}$$

Remark 3.1.1. *The estimator* $\widehat{\mathbf{B}}_{\hat{k}}$ *is also the globally optimal solution to the weighted constrained problem:*

$$\min_{r(\mathbf{B})\leq\hat{k}} \left|\left|(\mathbf{Y}-\mathbf{X}\mathbf{B})\boldsymbol{\Sigma}_e^{-1/2}\right|\right|_F^2. \tag{3.7}$$

In fact, if there exists a $\widetilde{\mathbf{B}}$ *such that* $r(\widetilde{\mathbf{B}})\leq\hat{k}$ *and*

$$\left|\left|(\mathbf{Y}-\mathbf{X}\widetilde{\mathbf{B}})\boldsymbol{\Sigma}_e^{-1/2}\right|\right|_F^2 < \left|\left|(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_{\hat{k}})\boldsymbol{\Sigma}_e^{-1/2}\right|\right|_F^2,$$

then

$$\left|\left|(\mathbf{Y}-\mathbf{X}\widetilde{\mathbf{B}})\boldsymbol{\Sigma}_e^{-1/2}\right|\right|_F^2 + \mu_1 r(\widetilde{\mathbf{B}}) < \left|\left|(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_{\hat{k}})\boldsymbol{\Sigma}_e^{-1/2}\right|\right|_F^2 + \mu_1 r(\widehat{\mathbf{B}}_{\hat{k}}) \tag{3.8}$$

and thus, $\widetilde{\mathbf{M}}=\widetilde{\mathbf{B}}\boldsymbol{\Sigma}_e^{-1/2}$ *achieves a lower value than* $\widehat{\mathbf{M}}_{\hat{k}}$ *which is a contradiction.*

Now, consider the possibility that $\boldsymbol{\Gamma}$ is positive-semidefinite and (3.5) still holds. But, there then exists infinitely many $\mathbf{B}$'s such that $\mathbf{B}\boldsymbol{\Gamma}^{1/2}=\widehat{\mathbf{M}}_{\hat{k}}$. However, of these $\mathbf{B}$'s, the one with the smallest Frobenius norm may always be chosen so that

$$\widehat{\mathbf{B}}_{\hat{k}} \triangleq \widehat{\mathbf{M}}_{\hat{k}}\left(\boldsymbol{\Gamma}^{1/2}\right)^{-}. \tag{3.9}$$

Hence, even if the weight matrix is positive-semidefinite, the global minimizer of the weighted penalized problem may always be found through a weighted version of RSC, referred here as Weighted Rank Selection Criterion (WRSC).

The direct calculation of the estimator of $\mathbf{A}$ is

$$\widehat{\mathbf{B}}_{\hat{k}} = \widehat{\mathbf{C}}_{\hat{k}}\widehat{\mathbf{D}}_{\hat{k}} \tag{3.10}$$

where

$$\widehat{\mathbf{C}}_{\hat{k}} = \widehat{\boldsymbol{\Sigma}}_{XX}^{-}\widehat{\boldsymbol{\Sigma}}_{XY}\boldsymbol{\Sigma}_e^{-1/2}\widehat{\mathbf{V}}_{\hat{k}} \tag{3.11}$$

$$\widehat{\mathbf{D}}_{\hat{k}} = \widehat{\mathbf{V}}_{\hat{k}}\boldsymbol{\Sigma}_e^{1/2} \tag{3.12}$$

where $\widehat{\mathbf{V}}_{\hat{k}} = \begin{pmatrix}\hat{\mathbf{v}}_1 & \dots & \hat{\mathbf{v}}_{\hat{k}}\end{pmatrix}$ and $\hat{\mathbf{v}}_j$ is the eigenvector corresponding to the j-th largest eigenvalue of

$$\widehat{\mathbf{R}} = \boldsymbol{\Sigma}_e^{-1/2}\widehat{\boldsymbol{\Sigma}}_{YX}\widehat{\boldsymbol{\Sigma}}_{XX}^{-}\widehat{\boldsymbol{\Sigma}}_{XY}\boldsymbol{\Sigma}_e^{-1/2}. \tag{3.13}$$

3.1.1 Tuning Parameter Selection

Tuning parameter selection needs to be done carefully. In RSC, a closed-form adaptive tuning parameter could be used. If $\boldsymbol{\Gamma}=\boldsymbol{\Sigma}_e^{-1}$ or a good estimator thereof is used in WRSC, the adaptive tuning parameter from RSC could still be used with $\mathbf{X}$ and $\mathbf{Z}$. To generalize the tuning parameter to WRSC in terms of $\mathbf{X}$ and $\mathbf{Y}$, a necessary lemma is first stated.

Lemma 3.1.2. *Let* $\widehat{\mathbf{M}}=\widehat{\mathbf{B}}\boldsymbol{\Sigma}_e^{-1/2}$ *be the full-rank coefficient estimator of* $\mathbf{M}$ *where* $\widehat{\mathbf{B}}=\widehat{\boldsymbol{\Sigma}}_{XX}^{-}\widehat{\boldsymbol{\Sigma}}_{XY}$ *and let* $\widehat{\mathbf{M}}_k=\widehat{\mathbf{B}}_k\boldsymbol{\Sigma}_e^{-1/2}$ *be the reduced-rank estimator of* $\mathbf{M}_k$ *of rank* k *for the positive-definite matrix* $\boldsymbol{\Sigma}_e$ *with* $\widehat{\mathbf{B}}_k$ *defined in (2.55). Then,*

$$\frac{1}{m}\left|\left|\mathbf{X}\widehat{\mathbf{M}}-\mathbf{X}\widehat{\mathbf{M}}_k\right|\right|_F^2 = \frac{1}{m}\left|\left|(\mathbf{X}\widehat{\mathbf{B}}-\mathbf{X}\widehat{\mathbf{B}}_k)\boldsymbol{\Sigma}_e^{-1/2}\right|\right|_F^2 = \sum_{j>k}\lambda_j \tag{3.14}$$

where λ_j *denotes the* j*-th largest eigenvalue of* $\boldsymbol{\Sigma}_e^{-1/2}\widehat{\boldsymbol{\Sigma}}_{YX}\widehat{\boldsymbol{\Sigma}}_{XX}^{-}\widehat{\boldsymbol{\Sigma}}_{XY}\boldsymbol{\Sigma}_e^{-1/2}$.

Proof. By Izenman [24],

$$\begin{aligned}
\frac{1}{m}\|(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_k)\mathbf{\Sigma}_e^{-1/2}\|_F^2 &= \frac{1}{m}tr\left\{(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_k)'(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_k)\mathbf{\Sigma}_e^{-1}\right\} \\
&= tr\left\{\left(\widehat{\mathbf{\Sigma}}_{YY}-\mathbf{\Sigma}_e^{-1/2}\left(\sum_{j=1}^{k}\lambda_k\mathbf{v}_j\mathbf{v}_j'\right)\mathbf{\Sigma}_e^{-1/2}\right)\mathbf{\Sigma}_e^{-1}\right\} \\
&= tr\left\{(\widehat{\mathbf{\Sigma}}_{YY}-\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY})\mathbf{\Sigma}_e^{-1}+\sum_{j>k}\lambda_j\mathbf{v}_j\mathbf{v}_j'\right\} \\
&= tr\left\{(\widehat{\mathbf{\Sigma}}_{YY}-\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY})\mathbf{\Sigma}_e^{-1}\right\}+\sum_{j>k}\lambda_j
\end{aligned}$$

where λ_j is the j-th largest eigenvalue of $\mathbf{\Sigma}_e^{1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_e^{1/2}$ and $\mathbf{v}_j$ is its corresponding normalized eigenvector. With a rearrangement of terms,

$$\begin{aligned}
\sum_{j>k}\lambda_j &= \frac{1}{m}tr\left\{(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_k)'(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_k)\mathbf{\Sigma}_e^{-1}\right\}-tr\left\{(\widehat{\mathbf{\Sigma}}_{YY}-\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY})\mathbf{\Sigma}_e^{-1}\right\} \\
&= \frac{1}{m}tr\left\{\left((\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_k)'(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_k)-(\mathbf{Y}'\mathbf{Y}-\mathbf{Y}'\mathbf{P}\mathbf{Y})\right)\mathbf{\Sigma}_e^{-1}\right\} \\
&= \frac{1}{m}tr\left\{(\mathbf{X}\widehat{\mathbf{B}}-\mathbf{X}\widehat{\mathbf{B}}_k)'(\mathbf{X}\widehat{\mathbf{B}}-\mathbf{X}\widehat{\mathbf{B}}_k)\mathbf{\Sigma}_e^{-1}\right\} \\
&= \frac{1}{m}\|(\mathbf{X}\widehat{\mathbf{B}}-\mathbf{X}\widehat{\mathbf{B}}_k)\mathbf{\Sigma}_e^{-1/2}\|_F^2.
\end{aligned}$$

□

This lemma is now used in the following proposition that defines the estimated rank $\hat{k}$.

Proposition 3.1.3. *Let $\lambda_1 \geq \lambda_2 \geq \ldots$ be the ordered eigenvalues of $\mathbf{\Sigma}_e^{-1/2}\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_{XX}^{-1}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_e^{-1/2}$ with $\mathbf{\Gamma}=\mathbf{\Sigma}_e^{-1}$. Then,*

$$\hat{k}=\max\left\{k:\lambda_k\geq\frac{\mu_1}{m}\right\} \tag{3.15}$$

where μ_1 is the tuning parameter from (3.4).

Proof. Assume first that the rank k is fixed. For $\widehat{\mathbf{B}}_k$,

$$\|(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_k)\mathbf{\Sigma}_e^{-1/2}\|_F^2=\|(\mathbf{Y}-\mathbf{P}\mathbf{Y})\mathbf{\Sigma}_e^{-1/2}\|_F^2+\|(\mathbf{P}\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}_k)\mathbf{\Sigma}_e^{-1/2}\|_F^2$$

by the Pythagorean theorem. Note that $\mathbf{X}\widehat{\mathbf{B}}=\mathbf{P}\mathbf{Y}$. By Lemma 3.1.2 with $\mathbf{\Gamma}=\mathbf{\Sigma}_e^{-1}$, the second term may be written as

$$\|(\mathbf{X}\widehat{\mathbf{B}}-\mathbf{X}\widehat{\mathbf{B}}_k)\mathbf{\Sigma}_e^{-1/2}\|_F^2=m\sum_{j>k}\lambda_j.$$

Then, the weighted penalized least squares criterion reduces to

$$\|(\mathbf{Y}-\mathbf{P}\mathbf{Y})\mathbf{\Sigma}_e^{-1/2}\|_F^2+\left\{m\sum_{j>k}\lambda_j+\mu_1 k\right\}.$$

Now it is given that

$$\min_B \left\{ \|(\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}_e^{-1/2}\|_F^2 + \mu_1 r(\mathbf{B}) \right\} = \\ \|(\mathbf{Y} - \mathbf{PY})\mathbf{\Sigma}_e^{-1/2}\|_F^2 - \mu_1 n + \min_k \sum_{j>k} \{m\lambda_j - \mu_1\}.$$

By taking k as the largest index j for which $m \cdot \lambda_j - \mu_1 \geq 0$ or $\lambda_j \geq \frac{\mu_1}{m}$, $\sum_{j>k} \{m\lambda_j - \mu_1\}$ is minimized. □

Therefore, $\hat{k}$ is the the number of eigenvalues of $\mathbf{\Sigma}_e^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_e^{-1/2}$ that is greater than $\frac{\mu_1}{m}$.

In computation, a successful, practical application of this will require employing an estimator of $\mathbf{\Sigma}_e$ in the above. If the number of observations m is large enough, the estimator of the covariance of residual $\widehat{\mathbf{\Sigma}}_r$ is such an estimator where

$$\widehat{\mathbf{\Sigma}}_r = \widehat{\mathbf{\Sigma}}_{YY} - \widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}. \tag{3.16}$$

If $\mathbf{\Sigma}_e$ were known, then ideally the adaptive tuning parameter μ_{adap1} may be used where

$$\mu_{adap1} = 2S_1^2(n+q) \tag{3.17}$$

and

$$S_1^2 = \frac{\|(\mathbf{Y} - \mathbf{PY})\mathbf{\Sigma}_e^{-1/2}\|_F^2}{mn - qn} \tag{3.18}$$

where S_1^2 is just the unbiased estimator of σ^2. However, it is only hypothesized here that when

$$\mu_1 = \frac{(1+\theta)}{1-\delta}(1+\xi)^2 S_1^2(\sqrt{n} + \sqrt{q}) \tag{3.19}$$

for any $\theta, \xi > 0$ and $0 < \delta < 1$, that the rank consistency as from RSC still holds as in (2.73).

Given $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$, the transformed model is perfectly decorrelated. That is, the ML estimator of the variance $\hat{\sigma}^2 = 1$. Therefore, one choice for an adaptive style tuning parameter $\mu_{adap1} = \mu_{adap1}^{(r)}$ is

$$\mu_{adap1}^{(r)} = 2(n+q). \tag{3.20}$$

Hence, theoretics were developed to support the use of a known decorrelator weight matrix $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$. In the large sample setting, $\widehat{\mathbf{\Sigma}}_r$ provides a good estimator and examination of the tuning parameter and its adaptive form was provided. While this provides modeling methodology, inferences still need to be drawn using CCA. In the connection of WRSC to CCA, the Adaptive CCA will be formed in the following section. In the process, it will be demonstrated that WRSC can be performed with a version of $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ instead, with the same rank recovery and coefficient estimation (provided that the previous connection to $\widehat{\mathbf{\Sigma}}_r$ holds).

3.2 Adaptive Canonical Correlation Analysis

ACCA complements the construction of WRSC by adding a step to CCA in which the number of significant canonical correlations t is chosen adaptively from the data in an optimal fashion. Since estimating the rank $\hat{k}$ of the coefficient matrix is equivalent to estimating the number of significant canonical correlations $\hat{t}$, an estimate of t may be found using WRSC, with some requirements. It was established in Section 2.3 that using $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$ and $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ results in the same estimated coefficient matrix in the population version. Thus, either choice of weight matrix results in the same estimate of the rank $\hat{k}$ and thus $\hat{t}$.

However, one of the difficulties lies in the proper scaling of the tuning parameter. Assume the rank is k. Even if the relationship in Lemma 2.3.3 exists and assuming that $\mathbf{\Sigma}_e$ and $\mathbf{\Sigma}_{YY}$ are positive-definite and nonsingular, it is still true that

$$\min_{\mathbf{B}, r(\mathbf{B})=k} \left|\left|(\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}_e^{-1/2}\right|\right|_F^2 \neq \min_{\mathbf{B}, r(\mathbf{B})=k} \left|\left|(\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}_{YY}^{-1/2}\right|\right|_F^2.$$

That is, the minimum value that is attained by the criterion is not the same for $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$ and $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$. Hence, scaling of the tuning parameter is required in the WRSC criterion also.

First assume $\mu_1 = \mu_1^{(y)}$ is the tuning parameter. The canonical variates may be recovered by setting $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ in place of $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$. Then the criterion is written as

$$\widehat{\mathbf{A}} = \widehat{\mathbf{C}}_{\hat{k}}\widehat{\mathbf{D}}_{\hat{k}} = \underset{\mathbf{B}=\mathbf{CD}}{\operatorname{argmin}} \left\{ \left|\left|(\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}_{YY}^{-1}\right|\right| + \mu_1^{(y)} r(\mathbf{B}) \right\}. \tag{3.21}$$

Provided that the number of observations m is large enough, a good estimator of $\mathbf{\Sigma}_{YY}$ is simply the sample response matrix $\widehat{\mathbf{\Sigma}}_{YY}$ which may be substituted in (3.21).

The canonical variates may then be recovered by

$$\widehat{\mathbf{\Xi}} = \mathbf{X}\widehat{\mathbf{C}}_{\hat{k}} \tag{3.22}$$

$$\widehat{\mathbf{\Omega}} = \mathbf{Y}\left(\widehat{\mathbf{D}}_{\hat{k}}\right)^{-} \tag{3.23}$$

as in Section 2.1.4 where $\hat{k} = \hat{t}$ the number of significant canonical correlations.

Now for a closer examination of the tuning parameter.

3.2.1 Tuning Parameter Selection

Good rank recovery is dependent upon the tuning parameter μ_1 selection in WRSC. If μ_1 is a known tuning parameter (adaptive or otherwise) when $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$, then a tuning parameter, denoted as $\mu_1^{(y)}$ for clarity, for $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ may be easily derived though the relationship established in Lemma 2.3.1. This is shown in the following theorem (continuing to assume that both weight matrix choices are positive-definite and nonsingular).

Theorem 3.2.1. *For any given tuning parameter μ_1 in*

$$\min_{\mathbf{B}} \left\{ \left|\left|(\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}_e^{-1/2}\right|\right|_F^2 + \mu_1 r(\mathbf{B}) \right\},$$

if

$$\frac{\mu_1^{(y)}}{m} = \frac{\mu_1/m}{1+\frac{\mu_1}{m}} \tag{3.24}$$

in (3.21), then the rank recovered from setting the weight matrix to $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$ *using the tuning parameter* μ_1 *is identical to the rank recovered from setting the weight matrix to* $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ *using the tuning parameter* $\mu_1^{(y)}$ *as defined above.*

Proof. By definition,

$$\lambda_{\hat{k}} \geq \frac{\mu_1}{m} > \lambda_{\hat{k}+1}$$

where λ_k is the k-th largest eigenvalue of $\mathbf{\Sigma}_e^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_e^{-1/2}$. Let ρ_k denote the k-th largest eigenvalue of $\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}$. By Lemma 2.3.1, λ_k and ρ_k are related by

$$\rho_k = \frac{\lambda_k}{1+\lambda_k} \qquad \lambda_k = \frac{\rho_k}{1-\rho_k}.$$

So,

$$\frac{\rho_{\hat{k}}}{1-\rho_{\hat{k}}} \geq \frac{\mu_1}{m} > \frac{\rho_{\hat{k}+1}}{1-\rho_{\hat{k}+1}}$$

implying

$$\rho_{\hat{k}} \geq \frac{\mu_1}{m+\mu_1} > \rho_{\hat{k}+1}.$$

Therefore, the tuning parameter $\mu_1^{(y)}$ for $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ that yields the same estimated rank $\hat{k}$ as from using $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$ is

$$\mu_1^{(y)} = \frac{\mu_1}{1+\frac{\mu_1}{m}}. \tag{3.25}$$

□

Hence, given a tuning parameter μ_1 and that $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$ in WRSC, a choice for the tuning parameter $\mu_1^{(y)}$ is as defined in (3.25) with $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$. In practice, the sample response covariance $\widehat{\mathbf{\Sigma}}_{YY}$ would be used instead.

Since the adaptive tuning parameter when $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$ is $\mu_{adap1} = 2(n+q)$, using Theorem 3.2.1 provides a closed-form of the adaptive tuning parameter for $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ as

$$\mu_{adap1}^{(y)} = \frac{2(n+q)}{1+\frac{2(n+q)}{m}}. \tag{3.26}$$

Note that up until now, in this chapter, only the population versions and large sample estimators of the weight matrices have been discussed. A careful discussion of their sample versions in the high-dimensional setting is now required.

3.3 High-Dimensional Considerations

WRSC has established that a globally optimal solution to this weighted constrained problem exists. The properties and theorems of RSC hold only if the weight matrix $\mathbf{\Gamma}$ can act as a true model decorrelator, such as $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$. If $\mathbf{\Sigma}_e$ were known, no further issues need to be addressed as Bunea *et al.* have established properties for high-dimensional rank recovery. Through the population version established by Reinsel and Velu, the relationship to $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ is direct, provided that $\mathbf{\Sigma}_{YY}$ is positive-definite [32]. This is then directly CCA with the additional step of adaptive estimation of the number of significant canonical correlations. In the large sample setting, these concepts carry over to the sample version.

However, in the high-dimensional setting, the sample version introduces a plethora of problems. It is desirable to achieve two goals of interest: (1) estimating the rank $\hat{k}$ of the coefficient matrix and (2) recovering a good estimator of the unknown coefficient matrix $\mathbf{A}$. The idea here is to gently, and empirically, explore options for high-dimensional application. It is noted that in any case, $\widehat{\mathbf{\Sigma}}_r = \widehat{\mathbf{\Sigma}}_{YY} - \widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY} = \frac{1}{m}\mathbf{Y}'(\mathbb{I}_n - \mathbf{P})\mathbf{Y}$ is singular and not invertible. An alternative is to use Moore-Penrose pseudoinverse $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_r^{-}$. However, this will show to be unstable as finding this reduces to taking the generalized inverse of the singular values. If the singular values are small, these approach ∞. On the other hand, $\widehat{\mathbf{\Sigma}}_{YY}$ is still nonsingular even in $p > m$ as long as $n < m$. A second option is to employ regularization on the singular matrices. That is, the sample residual covariance could be regularized so that $\mathbf{\Gamma} = (\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{-1}$. But what can be directly established is that even for the regularized versions of the weight matrices $\widehat{\mathbf{\Sigma}}_r$ and $\widehat{\mathbf{\Sigma}}_{YY}$, the relationships in Lemma 2.3.1 still hold and hence would provide equivalent results. But the latter has a small advantage over the former as its corresponding eigenvalues have a stable limit while those for the regularized version of $\widehat{\mathbf{\Sigma}}_r$ may not.

To establish similar relationships as in Lemma 2.3.1 for the two sample regularized weight matrices $\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n$ and $\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n$, consider the following:

Lemma 3.3.1. *Let $\rho_j^{(\delta)}$ be the j-th largest eigenvalue of*

$$(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2} \tag{3.27}$$

and let $\lambda_j^{(\delta)}$ be the j-th largest eigenvalue of

$$(\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{-1/2}. \tag{3.28}$$

Then, $\rho_j^{(\delta)}$ and $\lambda_j^{(\delta)}$ are related by

$$\rho_j^{(\delta)} = \frac{\lambda_j^{(\delta)}}{1 + \lambda_j^{(\delta)}} \tag{3.29}$$

and

$$\lambda_j^{(\delta)} = \frac{\rho_j^{(\delta)}}{1 - \rho_j^{(\delta)}}. \tag{3.30}$$

Proof. By definition,

$$|\rho_j^{(\delta)} - (\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}| = 0 \quad (3.31)$$

$$|\lambda_j^{(\delta)} - (\widehat{\mathbf{\Sigma}}_{r} + \delta\mathbb{I}_n)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_{r} + \delta\mathbb{I}_n)^{-1/2}| = 0 \quad (3.32)$$

where $\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n$ and $\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n$ are nonsingular for $\delta > 0$. From (3.32),

$$\begin{aligned} |(\widehat{\mathbf{\Sigma}}_r &+ \delta\mathbb{I}_n)^{1/2}\lambda_j^{(\delta)}\mathbb{I}_n(\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{1/2} \\ &- (\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{1/2}(\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{-1/2}(\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{1/2}| = 0 \\ \Rightarrow\ &|\lambda_j^{(\delta)}(\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n) - \widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}| = 0. \end{aligned}$$

From this, it follows that

$$\begin{aligned} 0 &= |\lambda_j^{(\delta)}(\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n) - \widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}| \\ &= |\lambda_j^{(\delta)}(\widehat{\mathbf{\Sigma}}_{YY} - \widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY} + \delta\mathbb{I}_n) - \widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}| \\ &= |\lambda_j^{(\delta)}(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n) - (1 + \lambda_j^{(\delta)})\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}| \end{aligned}$$

Thus,

$$\Big|\frac{\lambda_j^{(\delta)}}{1 + \lambda_j^{(\delta)}} - (\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}\Big| = 0.$$

Because a function of the form $h(t) = \frac{t}{1+t}$, $t > 0$ is strictly monotonic, $\frac{\lambda_j^{(\delta)}}{1+\lambda_j^{(\delta)}}$ yields all eigenvalues of

$$(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}.$$

That is,

$$\rho_j^{(\delta)} = \frac{\lambda_j^{(\delta)}}{1 + \lambda_j^{(\delta)}}$$

and it is immediate that

$$\lambda_j^{(\delta)} = \frac{\rho_j^{(\delta)}}{1 - \rho_j^{(\delta)}}.$$

□

Note on Convergence Lemma 3.3.1 assumes that $\delta > 0$ and that $\rho_j^{(\delta)}$ and $\lambda_j^{(\delta)}$ are dependent on δ. When $\delta \Rightarrow 0+$, $\rho_j^{(\delta)}$ has a stable limit while $\lambda_j^{(\delta)}$ may not. This is noted using the following lemma.

Lemma 3.3.2. *For $\delta > 0$,*

$$\lim_{\delta \to 0+} (\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}\mathbf{Y}' = \left((\frac{1}{m}\mathbf{Y}'\mathbf{Y})^{1/2}\right)^{-}\mathbf{Y}' \quad (3.33)$$

Proof. Assume that $\mathbf{Y}$ has been centered by its column means. Let $\mathbf{Y} = \mathbf{UDV}'$ be the reduced singular value decomposition of $\mathbf{Y}$ where $r = r(\mathbf{Y})$. Hence, $(\mathbf{D})_{ii} > 0$ for all i and $\mathbf{D}$ is a $r \times r$ matrix. Also, $\mathbf{U}$ is a $m \times r$ matrix and $\mathbf{V}$ is a $n \times r$ matrix with $\mathbf{U}'\mathbf{U} = \mathbb{I}_m$ and $\mathbf{V}'\mathbf{V} = \mathbb{I}_r$. Thus, $\widehat{\mathbf{\Sigma}}_{YY}$ may be written as

$$\widehat{\mathbf{\Sigma}}_{YY} = \frac{1}{m}\mathbf{Y}'\mathbf{Y} = \frac{1}{m}\mathbf{V}\mathbf{D}^2\mathbf{V}'.$$

Let $\widetilde{\mathbf{V}} = \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix} \in \mathbb{R}^{n\times n}$ be an orthonormal matrix so that $\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}' = \widetilde{\mathbf{V}}'\widetilde{\mathbf{V}} = \mathbb{I}_n$. Thus,

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n &= \frac{1}{m}\widetilde{\mathbf{V}}\begin{pmatrix}\mathbf{D}^2 & \mathbb{O}\\ \mathbb{O} & \mathbb{O}\end{pmatrix}\widetilde{\mathbf{V}}' + \delta\mathbb{I}_n\\ &= \widetilde{\mathbf{V}}\begin{pmatrix}\frac{1}{m}\mathbf{D}^2 + \delta\mathbb{I}_r & \mathbb{O}\\ \mathbb{O} & \delta\mathbb{I}_{n-r}\end{pmatrix}\widetilde{\mathbf{V}}'\end{aligned}$$

$$\begin{aligned}\Rightarrow (\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}\mathbf{Y}' &= \widetilde{\mathbf{V}}\begin{pmatrix}(\frac{1}{m}\mathbf{D}^2 + \delta\mathbb{I}_r)^{-1/2} & \mathbb{O}\\ \mathbb{O} & \delta^{-1/2}\mathbb{I}_{n-r}\end{pmatrix}\widetilde{\mathbf{V}}'\left(\mathbf{VDU}'\right)\\ &= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}(\frac{1}{m}\mathbf{D}^2 + \delta\mathbb{I}_r)^{-1/2} & \mathbb{O}\\ \mathbb{O} & \delta^{-1/2}\mathbb{I}_{n-r}\end{pmatrix}\begin{pmatrix}\mathbf{V}\\ \mathbf{V}_\perp\end{pmatrix}\mathbf{VDU}'\\ &= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}(\frac{1}{m}\mathbf{D}^2 + \delta\mathbb{I}_r)^{-1/2} & \mathbb{O}\\ \mathbb{O} & \delta^{-1/2}\mathbb{I}_{n-r}\end{pmatrix}\begin{pmatrix}\mathbb{I}_r\mathbf{DU}'\\ \mathbb{O}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}(\frac{1}{m}\mathbf{D}^2 + \delta\mathbb{I}_r)^{-1/2}\mathbf{DU}'\\ \mathbb{O}\end{pmatrix}\\ &= \mathbf{V}(\frac{1}{m}\mathbf{D}^2 + \delta\mathbb{I}_r)^{-1/2}\mathbf{DU}'.\end{aligned}$$

From the reduced singular value decomposition, $\mathbf{D}$ has positive diagonal entries. Hence,

$$\lim_{\delta\to 0+}(\frac{1}{m}\mathbf{D}^2 + \delta\mathbb{I}_r)^{-1/2}\mathbf{D} = \sqrt{m}\mathbf{I}_r.$$

So,

$$\lim_{\delta\to 0+}(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2} = \sqrt{m}\mathbf{VU}'$$

where

$$\begin{aligned}\sqrt{m}\mathbf{VU}' &= \sqrt{m}\mathbf{VD}^{-1}\mathbf{DU}' = \sqrt{m}\mathbf{VD}^{-1}\mathbf{V}'\mathbf{VDU}' = \left((\frac{1}{m}\mathbf{VD}^2\mathbf{V}')^{1/2}\right)^{-}\mathbf{VDU}'\\ &= \left((\frac{1}{m}\mathbf{Y}'\mathbf{Y})^{1/2}\right)^{-}\mathbf{Y}'\end{aligned}$$

□

Remark 3.3.3. *Lemma 3.3.2 implies that*

$$\begin{aligned}(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1/2} &\to (\widehat{\mathbf{\Sigma}}_{YY}^{1/2})^{-}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_{YY}^{1/2})^{-}\\ &\quad as\ \delta \to 0+.\end{aligned}$$

Thus $\rho_j^{(\delta)}$ converges to the eigenvalues of $(\widehat{\mathbf{\Sigma}}_{YY}^{1/2})^{-}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_{YY}^{1/2})^{-}$.

The eigenvalues $\rho_j^{(\delta)}$ have some advantage over the eigenvalues $\lambda_j^{(\delta)}$ since $\lambda_j^{(\delta)}$ may not be stable. To see this, consider a small example:

Example 3.3.4. *Let*

$$\mathbf{Y} = \begin{pmatrix} \mathbb{I} & & \\ & \mathbb{I} & \\ & & \mathbb{O} \end{pmatrix} \text{ and } \mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' = \begin{pmatrix} \mathbb{O} & & \\ & \mathbb{I} & \\ & & \mathbb{O} \end{pmatrix}.$$

Since

$$(\mathbb{I} - \mathbf{P})\mathbf{Y} = \begin{pmatrix} \mathbb{I} & & \\ & \mathbb{O} & \\ & & \mathbb{O} \end{pmatrix},$$

then

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_r &= \widehat{\mathbf{\Sigma}}_{YY} - \widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY} \\ &\propto \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{P}\mathbf{Y} \\ &= \mathbf{Y}'(\mathbb{I}_n - \mathbf{P})\mathbf{Y} \\ &= \mathbf{Y}'(\mathbb{I}_n - \mathbf{P})(\mathbb{I}_n - \mathbf{P})\mathbf{Y} \\ &= \begin{pmatrix} \mathbb{I} & & \\ & \mathbb{O} & \\ & & \mathbb{O}. \end{pmatrix} \end{aligned}$$

Hence,

$$\begin{aligned} (\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{-1/2} &\propto \begin{pmatrix} (1+\delta)\mathbb{I} & & \\ & \delta\mathbb{I} & \\ & & \delta\mathbb{I} \end{pmatrix}^{-1/2} \\ &= \begin{pmatrix} \frac{1}{1+\delta}\mathbb{I} & & \\ & \frac{1}{\delta}\mathbb{I} & \\ & & \frac{1}{\delta}\mathbb{I} \end{pmatrix}^{1/2} \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} (\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{-1/2}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_r + \delta\mathbb{I}_n)^{-1/2} &\propto \begin{pmatrix} \frac{1}{\sqrt{1+\delta}}\mathbb{I} & & \\ & \frac{1}{\sqrt{\delta}}\mathbb{I} & \\ & & \frac{1}{\sqrt{\delta}}\mathbb{I} \end{pmatrix} \begin{pmatrix} \mathbb{O} & & \\ & \mathbb{I} & \\ & & \mathbb{O} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{1+\delta}}\mathbb{I} & & \\ & \frac{1}{\sqrt{\delta}}\mathbb{I} & \\ & & \frac{1}{\sqrt{\delta}}\mathbb{I} \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{O} & & \\ & \frac{1}{\delta}\mathbb{I} & \\ & & \mathbb{O} \end{pmatrix} \\ &\to \lambda_j^{(\delta)} = \infty \text{ as } \delta \to 0+. \end{aligned}$$

Therefore $\lambda_j^{(\delta)}$ becomes more unstable as δ gets smaller.

Remark 3.3.5. *Also noted in this example is that*

$$\begin{aligned}\left(\widehat{\mathbf{\Sigma}}_r^{1/2}\right)^{-}\widehat{\mathbf{\Sigma}}_{YX}\widehat{\mathbf{\Sigma}}_{XX}^{-}\widehat{\mathbf{\Sigma}}_{XY}\left(\widehat{\mathbf{\Sigma}}_r^{1/2}\right)^{-} &\propto \begin{pmatrix}\mathbb{I} & & \\ & \mathbb{O} & \\ & & \mathbb{O}\end{pmatrix}\begin{pmatrix}\mathbb{O} & & \\ & \mathbb{I} & \\ & & \mathbb{O}\end{pmatrix}\begin{pmatrix}\mathbb{I} & & \\ & \mathbb{O} & \\ & & \mathbb{O}\end{pmatrix} \\ &= \begin{pmatrix}\mathbb{O} & & \\ & \mathbb{O} & \\ & & \mathbb{O}\end{pmatrix},\end{aligned}$$

so using $\widehat{\mathbf{\Sigma}}_r^{-}$ *does not give good results either.*

From these result, using $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_r^{-}$ is not recommended and instead, using either $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-}$ or $\mathbf{\Gamma} = (\widehat{\mathbf{\Sigma}}_{YY}+\delta\mathbb{I}_n)^{-1}$ is suggested. While the proofs of the theoretical properties are obscure, simulations in the following chapter offer strong support for these ideas.

CHAPTER 4

SIMULATION EXPLORATION

Data was simulated here to compare and contrast WRSC and RSC when the error covariance took on a variety of forms. Two types of settings or "experiments" were examined: *large sample* and *high-dimension*. These were designed to offer perspective in application and practicality of WRSC in computation in comparison to RSC. Details of the simulation setups are first presented, followed by a discussion of the results.

4.1 Experiment Setup

The various covariance structures included the i.i.d. structure, autoregressive 1 (AR(1)) structure, heterogeneous autoregressive 1 (H-AR(1)) structure, and an unstructured (UN) structure. Table 4.1 contains the calculation for the (i, j)-th entry of each covariance structure. For the i.i.d. structure, $\sigma^2 = 1$ so that it was simply the identity matrix. The AR(1) structure had $\rho = 0.9$ and $\sigma^2 = 3$. The H-AR(1) structure randomly generated integers between 1 and 5 for the variances on the diagonal, used their square roots to find $\sigma_i^{1/2}$ and $\sigma_j^{1/2}$, and set $\rho = 0.9$. The unstructured covariance matrix was randomly generated from a uniform distribution. This was done for each of the following experiments and setups.

Table 4.1: Covariance Error Structures

Name	(i,j)-th element
i.i.d.	$\sigma^2 \mathbb{I}_n$
AR(1)	$\sigma^2 \rho^{\lvert i-j \rvert}$
H-AR(1)	$\sigma_i \sigma_j \rho^{\lvert i-j \rvert}$
UN	σ_{ij}

Experiment 1: Large Sample Size Generate $i = 1, ..., m$ observations of $\mathbf{x}_i \sim N_p(\mathbf{0}, \mathbf{\Sigma}_{XX})$ i.i.d. with $(\mathbf{\Sigma}_{XX})_{jk} = \rho_x^{|j-k|}$ for $\rho_x > 0$ and $j, k = 1, ..., p$. The true (unknown) coefficient matrix is set to $\mathbf{A} = b \cdot \mathbf{B}_0 \mathbf{B}_1$ where $b > 0$, $\mathbf{B}_0$ is a $p \times r$ matrix, and $\mathbf{B}_1$ is a $r \times n$

matrix so that $rank(\mathbf{A}) = r$. Both $\mathbf{B}_0$ and $\mathbf{B}_1$ have i.i.d. $N(0,1)$ entries. The $m \times n$ noise matrix $\mathbf{E}$ is composed of $\mathbf{e}_i$ observations that are also generated as independent $N(0,1)$. Then observation $\mathbf{y}_i'$ of the $m \times n$ $\mathbf{Y}$ matrix is calculated as $\mathbf{y}_i = \mathbf{x}_i'\mathbf{A} = \mathbf{e}_i'$.

The control parameters are initialized as $m = 100$, $p = 25$, $n = 25$, and $r = 10$ for combinations of $\rho_x = 0.1, 0.5, 0.9$ and $b = 0.2, 0.4, 0.6, 0.8$. The following two setups are used:

Setup 1: Assume that $\mathbf{\Sigma}_e$ is known. Utilize WRSC with the tuning parameter μ_{adap1} and RSC with the tuning parameter μ_{adap}.

Setup 2: Use $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$. Utilize WRSC with tuning parameter $\mu_{adap1} = \mu_{adap1}^{(y)}$ and RSC with the tuning parameter μ_{adap}.

Experiment 2: High-Dimensionality The matrix $\mathbf{X}$ is generated as $\mathbf{X}_0\mathbf{\Sigma}_{XX}^{1/2}$ where $(\mathbf{\Sigma}_{XX})_{jk} = \rho_x^{|j-k|}$ for $j, k = 1, \ldots, p$ and $\mathbf{X}_0 = \mathbf{X}_1\mathbf{X}_2$. Both $\mathbf{X}_1$ and $\mathbf{X}_2$ are simulated separately with independent $N(0,1)$ entries of sizes $m \times q$ and $q \times p$, respectively so that $rank(\mathbf{X}) = q$. The coefficient matrix is generated the same way as in in Experiment 1, and q is chosen so that $q < m$.

Again, this is done for combinations of $\rho_x = 0.1, 0.5, 0.9$ and $b = 0.2, 0.4, 0.6, 0.8$. The setups used here are:

Setup 1: Large p, Large n Assume that $\mathbf{\Sigma}_e$ is known. Utilize WRSC with the tuning parameter μ_{adap1} and RSC with the tuning parameter μ_{adap}. Here, the control parameters are initialized as $m = 20$, $p = 100$, $n = 25$, $q = 10$, and $r = 5$.

Setup 2: Large p, Small n Use $\mathbf{\Gamma} = (\widehat{\mathbf{\Sigma}}_{YY})^{-1}$. Utilize WRSC with tuning parameter $\mu_{adap1} = \mu_{adap1}^{(y)}$ and RSC with the tuning parameter μ_{adap}. Here, the control parameters are initialized as $m = 75$, $p = 200$, $n = 15$, $q = 10$, and $r = 5$. This is so that $n < m$ but $p > m$.

Setup 3: Large p, Large n Use $\mathbf{\Gamma} = (\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1}$ where δ is tuned using 5-fold cross-validation. Utilize WRSC with tuning parameter $\mu_{adap1} = \mu_{adap1}^{(y)}$ and RSC with the tuning parameter μ_{adap}. Here, the control parameters are initialized as $m = 20$, $p = 100$, $n = 25$, $q = 10$, and $r = 5$ so that $m < n < p$.

The control parameter b is used here to control the signal-to-noise ratio

$$SNR = d_r^2(\mathbf{XA})/(\sqrt{q} + \sqrt{n}) \tag{4.1}$$

where $d_r^2(\mathbf{XA})$ denotes the r-th largest singular value of $\mathbf{XA}$. Bunea *et al.* proved that recovery of the correct rank is only possible if the SNR is large enough [8]. The parameters $b = 0.2, 0.4, 0.6, 0.8$ correspond to moderate to high SNR values for low to moderate correlation between the predictors ($\rho_x = 0.1, 0.5, 0.9$).

4.2 Results

The results were evaluated at two levels: (1) correct rank recovery and (2) accuracy of the estimator. Rank recovery is measured as the percentage of times the correct rank is recovered out of 100 iterations. Note that this simulation has a "fixed X design" as $\mathbf{X}$ and $\mathbf{A}$ are simulated once per b and ρ_x combinations for all 100 iterations. Therefore for each b and ρ_x combination, only $\mathbf{E}$ is re-generated for each 100 iterations (and $\mathbf{Y}$ is re-calculated). Accuracy of the coefficient estimator is measured by the scaled average mean squared error (MSE) where

$$MSE(\mathbf{XA}) = 100 \cdot \frac{\|\mathbf{XA} - \mathbf{X}\widehat{\mathbf{A}}\|_F^2}{mn}. \tag{4.2}$$

The most complex setup here is for a moderate signal strength $b = 0.2$ with high correlation within the predictor set $\rho_x = 0.9$. Results are expected to improve as the predictor correlation decreases and as the signal strength increases. See Tables B.1 and B.2 for rank recovery rates and average MSEs for Experiment 1, Setup 1. Tables B.3 and B.4 contain similar results for Experiment 1, Setup 2. The high-dimensional results from Experiment 2, Setup 1 are given in Tables C.1 and C.2, followed by Tables C.3 and C.4 for Setup 2, and by Tables C.5 and C.6 for Setup 3.

Experiment 1: Large Sample For **Setup 1**, rank recovery rates and accuracy measures were identical for RSC and WRSC. This was expected since WRSC treats $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$ as a known matrix and WRSC reduces to RSC when $\mathbf{\Gamma} = \mathbb{I}_n$. But, for the non-i.i.d. covariances, WRSC outperforms RSC in the high signal strengths and the low predictor correlations, particularly noted for signal strengths of greater than $b = 0.4$ and a corresponding high predictor correlation $\rho_x = 0.9$. However, even moderate signal strengths ($b = 0.2$) yielded good results for WRSC for predictor correlations of $\rho_x = 0.5, 0.1$. Differences in recovery rates were sometimes 100% better for WRSC than RSC. These results offer support of WRSC provided that $\mathbf{\Sigma}_e$ is known indicating the validity of WRSC's basic theory.

Similar results were seen in the average MSEs for this setup as the average MSEs of WRSC and RSC were identical for i.i.d. errors but were much smaller for WRSC for the non-i.i.d. errors when rank recovery was high. This indicates that good recovery of the coefficient matrix (and thus, the canonical weights) is possible. There is a noted increase in the overall average MSEs as the i.i.d. structure had the lowest overall average MSEs (in the 15-18 range), followed by heterogeneous autoregressive 1 (in the 27-30 range), then autoregressive 1 (in the 32-35 range), and finally the unstructured covariance (in the 67-73 range). Notice that the overall ranges of the MSEs increase with the complexity of the covariance structure. Aside from this, there seems to be no obvious pattern between average MSEs and signal strength or predictor correlation.

Setup 2 treated $\mathbf{\Sigma}_e$ as an unknown covariance and replaces it with $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$. The corresponding adaptive tuning parameter $\mu_1^{(y)}$ is also used here in WRSC while RSC remains the same as in Setup 1. Rank recovery rates for i.i.d. errors were better for RSC than WRSC. However, WRSC did competitively well provided that the signal

strength was moderate to high ($b = 0.2$ to $b = 0.8$) and the predictor correlation was weak to moderate ($\rho_x = 0.1$ to $\rho_x = 0.5$). The worst cases for WRSC were for signal strength $b = 0.2$ or for high predictor correlation $\rho_x = 0.9$. The rank recovery rates for the non-i.i.d. covariance structures were consistently better for WRSC than RSC for most signal strengths and predictor correlations, particularly after the most challenging setup of $b = 0.2$ and $\rho_x = 0.9$. While rank recovery rates for WRSC were not as strong as in Setup 1 overall, they were still very good for strong signals and low predictor correlation.

The respective ranges of the average MSEs for the various structures remain very similar to those in Setup 1. This validates that using $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$ as the weight matrix provides good rank recovery and estimation of the coefficient matrix in the large sample setting.

Experiment 2: High-Dimension In **Setup 1,** when the covariance of the errors is treated as known, rank recovery from RSC was, as expected in the i.i.d. case, near perfect. Identical results were given by WRSC, too. However, rank recovery by RSC in non-i.i.d. settings was less than ideal with rank recovery rates in the 38% to 65% range. But, WRSC performed excellently across all types of covariance errors. Rates were almost 100% for all signal strengths and predictor correlations.

Average MSEs were identical in the i.i.d. covariance. But in the non-i.i.d. covariances, WRSC had noticeably smaller average MSEs than RSC. Again, as in Experiment 1, the average MSEs increase overall with the complexity of the covariance structure. Thus, rank recovery and good coefficient estimation is possible for WRSC. This offers promising support for WRSC application to the high-dimensional setting.

Setup 2 is designed so that $m > n$ but $m < p$. Because $m > n$, $\widehat{\mathbf{\Sigma}}_{YY}$ should still be a reasonable estimator of $\mathbf{\Sigma}_{YY}$ and be nonsingular. RSC has perfect rank recovery when errors are i.i.d., but WRSC does competitively well. As seen previously, when errors are generated with non-i.i.d. errors, WRSC does much better than RSC with rank recovery rates in the high 90% range. Rates for RSC vary in the non-i.i.d. settings in the 60% to 80% range (with the exception in AR(1) when $b = 0.2$ and $\rho_x = 0.9$).

The average MSEs for RSC when errors are generated i.i.d. are slightly better than those of WRSC, corresponding also to its slightly better rank recovery rates. In the other situations when errors are non-i.i.d., WRSC has lower average MSEs than RSC. Those that are generated from the unstructured covariance matrix are the largest for WRSC, in the mid-40s range while RSC's are in the 70s range. The average MSEs that are from AR(1) errors are the second largest, followed by H-AR(1), still with WRSC outperforming RSC.

The simulations for this high-dimension version of WRSC were performed as the given in **Setup 3** with the weighting matrix now set to $\mathbf{\Gamma} = (\widehat{\mathbf{\Sigma}}_{YY} + \delta \mathbb{I}_n)^{-1}$. In the i.i.d.

setting, rank recovery rates of WRSC were competitive with RSC provided that the signal strength was strong enough. Interestingly, WRSC did very well when $b = 0.2$ (moderate signal strength) with a moderate predictor correlation of $\rho_x = 0.5$, but then did comparatively poorer for a low predictor correlation $\rho_x = 0.1$. For non-i.i.d. structures, WRSC performed well for a signal strength $b = 0.4$ and a low predictor correlation $\rho_x = 0.1$. For signal strength $b = 0.4$ and moderate predictor correlation $\rho_x = 0.5$, rank recovery rates did moderately well. However, in general, WRSC outperforms RSC in the non-i.i.d. setting with near perfect recovery rates for a strong signal $b = 0.4$. While RSC does overall better than WRSC in the i.i.d. setting, WRSC does competitively well in the noted signal and correlation strength.

The average MSEs were similar in the i.i.d. setting between RSC and WRSC, provided that WRSC performed as well as RSC. However, in the non-i.i.d. setting, WRSC had much lower average MSEs than RSC where rank recovery rates were better. That is, typically for predictor correlations $\rho_x = 0.9$ and $\rho_x = 0.1$. Again, interestingly, the moderate predictor correlation $\rho_x = 0.5$ gave higher average MSEs for some of the lower signals examined here. In general, if RSC had better rank recovery rates, the corresponding average MSEs were also poor. The overall average MSE errors were larger in this high-dimensional setting, but again had similar patterns in the ranges as in Experiment 1. This shows that with the regularization parameter, coefficient recovery for WRSC is also better than RSC in the non-i.i.d. setting given a strong signal strength. However, the accuracy of the coefficient estimator does seem to be dependent upon the correct rank recovery.

It is noted that when Experiment 2, Setups 2 and 3 had similar size and parameters settings that they yielded similar results. Hence, it is recommended that $\mathbf{\Gamma} = (\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1}$ used only when $m < p$ *and* $m < n$ as cross-validation of δ can be time consuming. These simulations provide the support for WRSC in both settings of large sample and high-dimensional with weighting matrices $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$ and $\mathbf{\Gamma} = (\widehat{\mathbf{\Sigma}}_{YY} + \delta\mathbb{I}_n)^{-1}$ as rank recovery and coefficient estimation are shown to outperform RSC when the error covariance is not i.i.d. and in the i.i.d. case are only marginally poorer.

CHAPTER 5

LOW-DIMENSIONAL NEUROIMAGING APPLICATION OF ACCA

CCA has shown to be a useful interpretation tool for examining multiple relationships between two multivariate sets. As an extension of such, the practical application of ACCA will be demonstrated here using a dataset complied of clinical, cognitive, and neuroimaging measures. This dataset, recently collected by the Neurobehavioral Laboratory at the Miriam Hospital, was provided by Dr. Hernando Ombao of Brown University. It is a unique compilation of data as neurocognitive assessments have often been examined with clinical variables, but the collection of neuroimaging data is still a fairly new discipline. The application examined here extends the work of [9] where the same dataset was utilized in a penalized least-squares regression application. ACCA methodology will be implemented along with the standard inferential techniques as described in Chapter 2 to offer a new perspective to the relationships between the two sets. A brief description of the data will be provided followed by the analysis techniques applied. Results will then be summarized and significant connections will be drawn between the two sets.

5.1 Data Description

Measurements from sixty-two HIV-positive participants were used to examine the relationship between brain volumetric (BV) measures and diffusion tensor imaging (DTI) derived measures with those of HIV associated cognitive deficits measured using various standardized psychological assessments. Variables of the predictor set included clinical descriptors, as well as brain volumetric and DTI-derived measures, totaling 31 original predictor variables.

Automated brain segmentation was performed as described in [9] to produce measures of cortical grey matter, white matter, caudate, putamen, pallidum, thalamus, hippocampus, amygdala, and corpus callosum. The DTI derived measures give fractional anisotropy (FA) and mean diffusivity (MD) measurements. These were taken on the internal capsule of the brain on 5 consecutive axial slices on the white matter segmentation images creating four regions of interest with the first being the anterior-most and the last being the posterior-most sections of the brain. This was also done for the corpus callosum which was divided into three subregions: genu (anterior subregion), body (middle subregion), and splenium

(posterior subregion).

Anisotropy is defined as the property of having different values when measured in different directions. Fractional anisotropy is simply a scalar value that characterizes the amount of anisotropy of the diffusion in areas of the brain, reflecting fiber density, axonal diameter, and myelination in white matter [4]. Mean diffusivity measures the water diffusion or flow in the brain. In addition to these were clinical measures of HIV stage (early or advanced stages), Hepatitis C status, age, education, alcohol use, and cocaine/opiate use. Table 5.2 contains a complete list of all predictor variables

The response set includes 13 variables collected from the following tests: WAIS-III Symbol Search, Digit Span, and Letter-Number Sequencing, Grooved Pegboard (Gpeg), Trail Making A, Trail Making B, Controlled Oral Word Association Test (COWAT), Hopkins Verbal Learning Test- Revised (HVLT), and Brief Visuospatial Memory Test- Revised. See Table 5.1 for a complete list of all response variables. These are widely used neuropsychological tests and make up the following response domains of interest: 1) verbal fluency, 2) pyschomotor speed, 3) information processing speed and attention, 4) executive function, 5) learning, and 6) memory. Table 5.3 gives the response variables by each domain. The raw data collected from these assessments has been transformed so that all values are positive with smaller values associated with debilitated cognitive functioning.

Table 5.1: Description of Cognitive Response Variables

Response variables	**Description**
`HVLT_sum`	Total free recall for verbal learning test
`HVLT_delay`	Delayed recall for verbal memory test
`BVMT_sum`	Total free recall for non-verbal learning test
`BVMT_delay`	Delayed recall for non-verbal memory test
`WAIS_LNS`	Measure of attention and working memory with letter number sequencing
`WAIS_DigSym`	Measures processing speed with digit symbol
`WAIS_SymSmSrch`	Measures processing speed with symbol search
`Gpeg_dom`	Measures motor speed of putting pegs into holes with dominant hand
`Gpeg_nondom`	Measures motor speed of putting pegs into holes with non-dominant hand
`Trail_A`	Measures motor speed through simple test of visuomotor sequencing
`Trail_B`	Measures complex visuomotor sequencing
`COWAT`	Measures verbal fluency through word generations within one minute
`Animal`	Measures of verbal fluency

5.2 Methodology

The primary goal here is to estimate $\hat{t}$ the number of significant relationships between the response set and predictor set and to identify which variables contribute the most to these relationships. While there are many ways to examine and interpret the results from canonical analysis, a unique approach will be taken here to better understand the independent-dependent relationship between the predictor and response sets. This is done

Table 5.2: Description of Clinical and Neuroimagaing Predictor Variables

Predictor Variable	**Type**	**Description**
HIV_stage	Clinical	Binary measure based upon severity of immunosuppression
hcv_current	Clinical	Binary measure Hepatitis C status
age	Clinical	Number of years
Education	Clinical	Number of years
kmsk_alc	Clinical	Binary measure of alcohol use
kmsk_cocopi	Clinical	Binary measure of cocaine and opiate use
fa_cc_genu	DTI derived	Fractional anisotropy in anterior subregion of corpus callosum
fa_cc_body	DTI derived	Fractional anisotropy in body subregion of corpus callosum
fa_cc_splenium	DTI derived	Fractional anisotropy in posterior subregion of corpus callosum
md_cc_genu	DTI derived	Mean diffusivity in anterior subregion of corpus callosum
md_cc_body	DTI derived	Mean diffusivity in body subregion of corpus callosum
md_cc_splenium	DTI derived	Mean diffusivity in posterior subregion of corpus callosum
fa_ic1	DTI derived	Fractional anisotropy in 1st (anterior) subregion of internal capsule
fa_ic2	DTI derived	Fractional anisotropy in 2nd subregion of internal capsule
fa_ic3	DTI derived	Fractional anisotropy in 3rd subregion of internal capsule
fa_ic4	DTI derived	Fractional anisotropy in 4th (posterior) subregion of internal capsule
md_ic1	DTI derived	Mean diffusivity in 1st (anterior) subregion of internal capsule
md_ic2	DTI derived	Mean diffusivity in 2nd subregion of internal capsule
md_ic3	DTI derived	Mean diffusivity in 3rd subregion of internal capsule
md_ic4	DTI derived	Mean diffusivity in 4th (posterior) subregion of internal capsule
whitematter	Brain Volumetric	Volume measurement
cortex	Brain Volumetric	Volume measurement
thalamus	Brain Volumetric	Volume measurement
caudate	Brain Volumetric	Volume measurement
putamen	Brain Volumetric	Volume measurement
pallidum	Brain Volumetric	Volume measurement
hippocampus	Brain Volumetric	Volume measurement
amygdala	Brain Volumetric	Volume measurement
cc_splenium	Brain Volumetric	Corpus callosum posterior subregion
cc_body	Brain Volumetric	Corpus callosum middle 3 subregions
cc_genu	Brain Volumetric	Corpus callosum anterior subregion

despite the ability of canonical analysis to examine the relationships from a symmetric point of view and to better understand the data from a modeling aspect.

At the variable level, only the predictor cross-loadings will be examined to quantify the relationship between each predictor variable and the response canonical variate as a whole. That is, for $k = 1, \ldots, \hat{t}$ and $j = 1, \ldots, p$, $C_x(k, j)$ as given in Section 2.1.5 will be examined for each k response canonical variates with each j predictor variable. Only the canonical loadings of the response variables will be looked at to determine which response variables contribute the most to its respective variate. So, for $k = 1, \ldots, \hat{t}$ and $j = 1, \ldots, n$, $L_y(k, j)$ will be examined as the relationship between the k-th canonical response variate with the j-th response variable. The canonical coefficients, predictor loadings, and response cross-loadings are also provided, but not examined here for reasons previously stated.

All variables will be first centered and scaled by their columns. This is because the canonical variates are considered to be "artificial." By scaling and centering, the canonical

Table 5.3: Response Variables by Domain

Domain	**Response Variables**
Verbal	`COWAT, Animal`
Motor	`GPeg_dom, GPeg_nondom, Trail_A`
Information/Attention	`WAIS_DigSym, WAIS_SymSrch, Trail_A`
Executive	`Trail_B, COWAT, WAIS_LNS`
Learning	`HVLT_sum, BVMT_sum`
Memory	`HVLT_delay, BVMT_delay`

coefficients may be examined as standardized variables [25]. Following this, the number of significant relationships will be determined adaptively using WRSC/ACCA methodology. This leads to the recovery of the canonical coefficients and canonical variates. Generally speaking, only the largest of the canonical loadings/cross-loadings will be examined as the largest contributors to the variable-to-variate correlation. A larger canonical loading/cross-loading indicates that the particular variable is a larger contributor to its variate/opposite variate. The canonical loadings and cross-loadings may be understood as stated previously: values closer to ± 1 indicate stronger relationships while values close to 0 indicating a weaker relationship, positive values indicate as the variable increases, the variate as a whole also increases giving a direct relationship while negative values indicate that as the variable increases the variate as a whole decreases giving an inverse relationship.

First, the variables will be examined with all original predictors and all response variables. Results from this will lead to the examination of subsets, grouping by various domains. Specifically, 1.) learning and memory domains, 2.) executive functioning and information processing speed/attention domains, and 3.) verbal fluency and motor speed.

5.3 Results

The results are all tabulated in Appendix D.

5.3.1 Original Predictors, All Responses

Utilizing the original 31 predictor variables and all 13 response variables, the estimated number of significant canonical correlations was $\hat{t} = 4$. Therefore, ACCA estimated that there are four significant, uncorrelated relationships between the the two sets. The canonical correlations are:

$$\begin{aligned} \hat{\rho}_1 &= 0.799 \\ \hat{\rho}_2 &= 0.754 \\ \hat{\rho}_3 &= 0.739 \\ \hat{\rho}_4 &= 0.719 \end{aligned}$$

The normalized canonical coefficients, loadings, and cross-loadings are given in Table D.1 and D.2. The canonical response variates can be thought of as a compilation of all neurocognitive assessments as an overall index with larger values in the response set reflecting better cognitive performance.

First Relationship The first relationship yields strong response loadings for `HVLT_sum`, `HVLT_delay`, and `WAIS_DigSym` (≥ 0.50), followed closely by `BVMT_sum` and `BVMT_delay` (0.43 and 0.42). The HVLT and BVMT variables correspond to learning and memory domains while `WAIS_DigSym` corresponds to the information processing and attention domain, indicating a certain amount of information processing speed/attention are required for learning and memory function.

The predictor cross-loadings that are the strongest are `hcv_current`, `Education`, `kmsk_alc`, `kmsk_cocopi`, and `md_cc_splenium`. The binary variables corresponding to Hepatitis C status, alcohol use, and cocaine/opiate use all have negative cross-loadings, giving an inverse relationship with the response variate. Hinkin *et al.* and Ryan *et al.* had similar findings as they found that the co-infection of HIV and Hepatitis C were almost three times more likely to have impairments in learning and memory than those infected by HIV alone [20] [36]. Alcohol and drug use have also been long established to affect learning and memory functions. The subject's education is the only positive cross-loading here, as those with greater amounts of education tend to have greater learning and memory function.

Second Relationship In the second relationship, the strongest loadings in the response variate are `COWAT` and `Animal`, both positive and ≥ 0.50. These correspond to the verbal fluency domain and influence the response variate the most in a positive, direct direction. Loadings in the 0.40-0.50 range were associated primarily with executive functioning and pychomotor speed.

The predictor cross-loadings in this relationship are greatest for `age` (0.47), followed by `HIV_stage` and `Education` (both ≥ 0.25) and then by `fa_cc_body` and `pallidum`. All of these predictor cross-loadings are positive in direction. The age of the subject is the strongest influencing factor in the response variate, followed by the amount of education the subject has. While these are logical influencing factors in verbal fluency, the positive cross-loading of the HIV stage is counter-intuitive. The variable `pallidum` corresponds to the size of the pallidum in the brain, which is typically associated with motor control [42]. Its positive sign is not surprising as diminishing brain volume is associated with decrease in neurocogitive performance [30]. A moderate, direct relationship also is shown in the fractional anisotropy of the middle regions of the corpus callosum.

Third Relationship The third relationship had negative predictor cross-loadings in the brain volumetric measures of `whitematter`, `cortex`, `thalamus`, `putamen`, and `pallidum`. Also, significant negative cross-loadings were `fa_ic1`, `fa_ic2`, and `fa_cc_genu`, all fractional anisotropy measures in the internal capsule and the corpus capsule. Here, `Education` also had a negative cross-loading where `age`, `hcv_current`, `kmsk_cocopi`,

and `md_cc_genu` had positive cross-loadings.

While this may seem initially surprising, examination of the response loadings will show that most are negative, notably `BVMT_sum`, `BVMT_delay`, `WAIS_DigSym`, `WAIS_SymSrch`, `Gpeg_dom`, `Gpeg_nondom`, `Trail_B`, and `Animal`. While these variables cover a variety of domains, they are primarily associated with the information processing/attention and motor domains.

Fourth Relationship Examining the canonical loadings of the response variables for the fourth relationship, the strongest loadings correspond to variables `WAIS_DigSym`, `Gpeg_dom`, `Gpeg_nondom`, and `Trail_A`, all with positive loadings ≥ 0.50. `WAIS_SymSrch` has the next largest loading of 0.35. All of these variables correspond to the information and attention and motor domains. Thus, the largest contributors to this response variate are associated with information processing/attention, and psychomotor domains.

The predictor cross-loadings for the fourth relationship are the strongest for `age`, `md_cc_genu`, `md_cc_body`, `md_cc_splenium`, `fa_ic2`, and `md_ic1` to `md_ic4`. Both the age of the subject and the fractional anisotropy of the second region of the internal capsule are positive cross-loadings indicating a direct relationship with the response variate. All of the mean diffusivity measures cover the complete areas of the corpus callosum and the internal capsule and are negative cross-loadings. Thus, they maintain an inverse relationship with the response variate.

This corresponds with the findings by Wu *et al.* [43] that high mean diffusivity and low fractional anisotropy values in portions of the corpus callosum are correlated with defects in motor speed in HIV-positive patients.

5.3.2 Learning and Memory Domains with Original Predictors

Next, a subset of the domains was analyzed using the 31 original predictors and all of the response variables associated with the learning and memory domains. This returned $\hat{t} = 1$ significant relationship with $\hat{\rho} = 0.753$. Learning and memory are two very closely related domains. In the previous analysis, it was seen that variables relating to these domains from the first relationship loaded significantly together. Hence, it is unsurprising to see all four response variables from these two domains with very large, positive loadings. Thus, as any of these response variables increase, the overall response variate increases. The fact that only one significant relationship was estimated indicates that these two domains are correlated with each other.

The strongest cross-loadings of the predictors were `hcv_current`, `Education`, and `kmsk_cocopi`. Of these, only the measurement of the subjects' education had a positive relationship with the response variate. Both Hepatitis C and cocaine/opiate use had strong, negative relationships with the response variate. Co-infection of Hepatitis C and HIV has shown greater learning and memory impairment than just HIV infection alone [36] [20]. Drug use is well-known to adversely affect both learning and memory abilities [28]. Thus the strong, inverse relationship is expected.

Following these in strength are `fa_cc_genu`, `thalamus`, `putamen`, `pallidum`, and `hippocampus`, all positive in the 0.20 to 0.22 range. The latter four are brain volumetric measures. Again, decrease in brain volume is associated with impaired neurocognitive function so the positive relationship with learning and memory functions is expected [30].

5.3.3 Executive Functioning and Information Processing/Attention Domains with Original Predictors

The 6 response variables corresponding to either the executive functioning or the information processing/attention domains and the 31 original predictor variables returned $\hat{t} = 2$ significant relationships with canonical correlations $\hat{\rho}_1 = 0.783$ and $\hat{\rho}_2 = 0.735$.

First Relationship The three response variables that load into the first relationship are `WAIS_DigSym`, `WAIS_SymSrch`, and `Trail_A`, all of which are related to the information processing/attention domain. They are much stronger than the other three response variates and are all positive so that as each one of them increases, the response variate as a whole also increases.

The predictor variables that cross-load with the greatest strength in a negative direction are `hcv_current`, `kmsk_alc`, `kmsk_cocopi`, `md_cc_genu`, `md_cc_body`, and `md_cc_splenium`. The inverse relationship between the information/attention domain and Hepatitis C status, alcohol use, and cocaine/opiate use indicate that as they increase, the response variate decreases as a whole. The remaining three variables are all associated with the mean diffusivity of the corpus callosum, also indicating that as the diffusivity in the corpus callosum increases, the response variate decreases.

The predictor variables that cross-load positively are `fa_cc_splenium`, `whitematter`, `putamen`, and `hippocampus`. The later three are all brain volumetric measures, indicating an increase in brain volume in these areas can be associated with the increase in the information/attention domain. The first of these is the fractional anisotropy in the posterior region corpus callosum.

Second Relationship Interestingly, the variables that most strongly load into this response variate are those that are associated with the executive functioning domain. These are `Trail_B`, `COWAT`, and `WAIS_LNS`, all with loadings ≥ 0.60.

The predictor variable that cross-loads with the most strength is `age`. This cross-loading is positive and far stronger (0.59) than the rest of the predictor variables. The next three predictor variables in the 0.20-0.30 range are `hcv_current`, `kmsk_cocopi`, and `cc_splenium`. These are all positive, which is surprising because one would not intuitively think that Hepatitis C and cocaine/opiate use would be positively correlated to executive function. However, the gap between `age` and the next largest cross-loading is vastly different, indicating that this is not as strong contributor.

It is noted here that these two domains can be considered to be uncorrelated as two relationships were established, with loadings of response variables associated with each domain.

5.3.4 Verbal Fluency and Motor Domains with Original Predictors

With the verbal fluency and motor domains combined, there were a total of 5 response variables that were analyzed with the 31 original predictor variables. Only $\hat{t} = 1$ significant relationship was returned.

The response loadings that were the strongest were `COWAT`, `Animal`, `Gpeg_dom`, and `Gpeg_nondom`. The first two of these correspond with verbal fluency while the latter two correspond with psychomotor abilities. The single relationship indicates a close correlated relationship between the two domains.

Of the predictor cross-loadings, `age`, `Education`, `fa_cc_body`, `fa_ic1`, `pallidum`, and `amygdala` were the strongest, with a positive relationship with the response variate. Increase in age and the number of years of education that subject has had corresponds with greater verbal fluency and psychomotor abilities. The two brain volumetric variables indicate that larger volumes in these areas correspond also with an overall greater response variate. Also, the increase in fractional anisotropy in the middle three sections of the corpus callosum and the anterior most subregion of the internal capsule also corresponds with an overall greater response variate.

5.4 Summary

The application of ACCA has been shown here in the large sample setting to illustrate how unique relationships may be identified. By using all response variables with all original predictors, four uncorrelated relationships were returned. These corresponded to 1.) learning and memory domains, 2.) verbal fluency, 3.) a subset of the response variables across all of the domains, and 4.) information processing/attention and motor domains. The corresponding strongest cross-loadings of the predictor variables were unsurprising and documented by others.

Further subsetting steps were taken to examine these relationships closer into the sets of 1.) learning and memory domains, 2.) executive functioning and information processing/attention domains, and 3.) verbal fluency and motor domains. While the first and last of these estimated only one significant relationship apiece, the executive functioning and information/processing/attention domains estimated two significant relationships. Hence, it is appropriate to group learning and memory together and verbal fluency and motor skills together.

Generally speaking, supporting evidence has been given here to associate increased mean diffusivity, alcohol use, and cocaine/opiate use with decreased neurocognitive performance. Also education, age, increased fractional anisotropy and brain volumetrics have been shown to correspond with increased neurocognitive functioning.

CHAPTER 6

ACCA WITH VARIABLE SELECTION

Often times, particularly when the data is of high-dimension, it is desirable to first select variables before performing any type of additional analysis. A multivariate response regression model may be sparse in the sense that not all variables may be needed in the model, that the model is row sparse. Methods that handle row sparsity are variable selection techniques. So one way to address the issues that arise from a high-dimensional dataset is through a combination of both variable selection and some other multivariate technique to offer a more complete methodology. If Adaptive Canonical Correlation Analysis (ACCA) is implemented in the second step, then since the predictor set is first reduced by variable selection, it may not be necessary to have any other additional high-dimensional considerations for ACCA. This two step type of methodology is similar to that of Bunea *et al.* in [7] as models that are both row sparse and rank sparse (i.e. rank deficient) are explored. They address the rank sparseness aspect through techniques like the Rank Selection Criterion (RSC), that produces low rank estimators. See [7] for a complete discussion of these models. While ACCA is not immediately a rank reduction technique, it is very closely related as discussed previously.

Thus, a two-step process of variable selection followed by ACCA, will be offered to manage high-dimensional issues. Variable selection in multivariate response models is equivalent to group selection in the univariate response models. That is, if the j-th predictor variable is not in (2.4), this is equivalent with assuming that the j-th row in the coefficient matrix $\mathbf{A}$ is zero. Since the rows of $\mathbf{A}$ may be treated in groups of coefficients, then any group selection method for univariate response models may be employed to variable selection in multivariate response models. This will be made clear in the following section. What is then a nice consequence is that results in the univariate setting carry over to the multivariate setting. A common group selection method for high-dimensional models is Group Lasso (GLASSO) [44]. This will be introduced and then applied to a high-dimensional dataset derived from the neurocognitive data from Chapter 5. Then ACCA will be utilized to identify and infer the canonical relationships in the predictors only selected by GLASSO.

6.1 Equivalence of Group Selection in Univariate Models

Variable selection in a multivariate response regression model as given in (2.4) may be shown to be equivalent with group selection of variables in the univariate regression model.

If variables may be clustered together in predefined groups, then group selection of variables is simply selecting whole groups of variables to be included in the model.

To note this, first denote the predictor matrix $\widetilde{\mathbf{X}}$ as the predictor matrix prior to variable selection. Recall the multivariate response model

$$\mathbf{Y} = \widetilde{\mathbf{X}}\mathbf{A} + \mathbf{E} \tag{6.1}$$

where the rows of $\widetilde{\mathbf{X}}$ and $\mathbf{Y}$ make-up the $i = 1, \ldots, m$ observations, each row of size $1 \times p$ and $1 \times n$, respectively. Continue to assume that $\widetilde{\mathbf{X}}$ and $\mathbf{Y}$ have been centered by their column means and that $\mathbf{E}$ is a random zero mean matrix. This may then be written in a vectorized version of the transposed model as

$$\text{vec}(\mathbf{Y}') = (\widetilde{\mathbf{X}} \otimes \mathbb{I}_n)\text{vec}(\mathbf{A}') + \text{vec}(\mathbf{E}') \tag{6.2}$$

where $\mathbf{M} \otimes \mathbf{N}$ denotes the Kronecker product of two generic matrices $\mathbf{M}$ and $\mathbf{N}$. Hence, since $\widetilde{\mathbf{X}}$ is a $m \times p$ matrix and $\mathbb{I}_n$ is a $n \times n$ matrix, $\widetilde{\mathbf{X}} \otimes \mathbb{I}_n$ is a $mn \times pn$ size matrix with block elements of $(\widetilde{\mathbf{X}})_{ij}\mathbb{I}_n$ where $(\widetilde{\mathbf{X}})_{ij}$ denotes the (i, j)-th element of $\widetilde{\mathbf{X}}$.

Let $\mathbf{A}' = \begin{pmatrix}\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_p\end{pmatrix}$ where $\boldsymbol{\beta}_j \in \mathbb{R}^n$ is the j-th row of $\mathbf{A}$. Define

$$\begin{aligned}
\boldsymbol{\gamma} &:= \text{vec}(\mathbf{A}') && (6.3)\\
\mathbf{y} &:= \text{vec}(\mathbf{Y}') && (6.4)\\
\mathbf{Z} &:= (\widetilde{\mathbf{X}} \otimes \mathbb{I}_n) && (6.5)\\
\boldsymbol{\varepsilon} &:= \text{vec}(\mathbf{E}') && (6.6)
\end{aligned}$$

where $\boldsymbol{\gamma}$ is a $pn \times 1$ vector and $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are a $mn \times 1$ vectors. Thus, (6.2) may be written as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{6.7}$$

where

$$\boldsymbol{\gamma} = \begin{pmatrix}\boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_p\end{pmatrix} = \begin{pmatrix} a_{11} \\ \vdots \\ a_{1n} \\ a_{21} \\ \vdots \\ a_{p1} \\ \vdots \\ a_{pn}\end{pmatrix}.$$

Denote the columns of $\widetilde{\mathbf{X}}$ as $\widetilde{\mathbf{X}}^{(1)}, \widetilde{\mathbf{X}}^{(2)}, \ldots, \widetilde{\mathbf{X}}^{(p)}$, each a vector of size $m \times 1$. Then (6.7) can be re-written as

$$\mathbf{y} = \sum_{k=1}^{p} (\widetilde{\mathbf{X}}^{(k)} \otimes \mathbb{I}_n)\boldsymbol{\beta}_k + \boldsymbol{\varepsilon} := \sum_{k=1}^{p} \mathbf{Z}_k\boldsymbol{\beta}_k + \boldsymbol{\varepsilon} \tag{6.8}$$

where

$$\mathbf{Z}_k := \begin{pmatrix} x_{1,k}\mathbb{I}_n \\ x_{2,k}\mathbb{I}_n \\ \vdots \\ x_{m,k}\mathbb{I}_n\end{pmatrix}.$$

Consistent group selection holds now on $\mathbf{Z}$, but using properties of the Kroenecker product, reduce to conditions holding on $\widetilde{\mathbf{X}}$. If $\widetilde{\mathbf{X}}^{(j)}$ is not in the model, then the j-th row of $\mathbf{A}$ is identically zero which is just $\boldsymbol{\beta}_j$ in the vectorized model. By treating the $\boldsymbol{\beta}_j$'s as groups of coefficients, any group selection method for univariate response regression models may now be employed.

6.2 Group Lasso

Group LASSO is one group selection technique for univariate models (and thus also a variable selection technique in the multivariate setting) that is popular for high-dimensional data [44]. Continuing in the above notation, suppose there are p groups, all of the same size m. The GLASSO minimizes the following convex criterion:

$$\min_{\mathbf{B}} \left\{ \frac{1}{2} \left|\left| \text{vec}(\mathbf{Y}') - (\widetilde{\mathbf{X}} \otimes \mathbb{I}_n)\text{vec}(\mathbf{B}') \right|\right|_F^2 + \lambda \left|\left| \mathbf{B} \right|\right|_{2,1} \right\} = \min_{\mathbf{B}} \left\{ \frac{1}{2} \left|\left| \mathbf{Y} - \widetilde{\mathbf{X}}\mathbf{B} \right|\right|_F^2 + \lambda \left|\left| \mathbf{B} \right|\right|_{2,1} \right\} \tag{6.9}$$

where $\widetilde{\mathbf{X}}$ denotes the predictor set prior to variable selection, $||\mathbf{B}||_{2,1} = \sum_{j=1}^{p} ||\mathbf{b}||_2$, and λ is the tuning parameter.

While this is a convex problem, it may be computationally expensive to find a global minimum, particularly for a large dataset. This may be remedied by using thresholding operations instead such as soft-thresholding.

Let Θ be the soft-thresholding operator. Define $\overrightarrow{\Theta}$ as the multivariate version for any vector $\boldsymbol{\alpha}$ as

$$\overrightarrow{\Theta}(\boldsymbol{\alpha}; \lambda) = \boldsymbol{\alpha}^{\circ} \Theta(\|\boldsymbol{\alpha}\|_f; \lambda) \tag{6.10}$$

where $\boldsymbol{\alpha}^{\circ} = \begin{cases} \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|_f}, & \text{if } \boldsymbol{\alpha} \neq 0 \\ 0, & \text{if } \boldsymbol{\alpha} = 0 \end{cases}$. Then, define

$$T \circ \mathbf{B} = \overrightarrow{\Theta}\left(\frac{1}{K}\widetilde{\mathbf{X}}'\mathbf{Y} + (\mathbb{I} - \frac{1}{K}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})\mathbf{B}; \frac{\lambda}{K} \right) \tag{6.11}$$

for all $\mathbf{B}$ matrices of size $p \times k$, where K is a constant satisfying $K > \|\mathbf{X}\|_2^2/2$. Refer to [37] and [7] for further details.

6.3 High-Dimensional Neuroimaging Application

Consider, once again, the neurocognitive dataset from Chapter 5. Recall that the predictor set contains 31 original predictors of clinical, DTI-imaging, and brain volumetric measures. The response set contains 13 variables from neurocognitive assessments that looked at the following domains: 1) verbal fluency, 2) pyschomotor speed, 3) information processing and attention, 4) executive functioning, 5) learning, and 6) memory. These measurements were for a total of 62 HIV-positive patients.

To create a predictor set of higher dimension, quadratic DTI and quadratic brain volumetric measures were added, along with interaction terms between `HIV_stage`, the DTI measures, and the brain volumetric measures. This resulted in a total of 235 predictors

of which 31 were the original predictors. Hence, this new dataset has $p = 235$ predictors, $n = 13$ responses, and $m = 62$ observations.

Group LASSO was first used with five-fold cross-validation. Both $\widetilde{\mathbf{X}}$ and $\mathbf{Y}$ are standardized prior to analysis where $\widetilde{\mathbf{X}}$ denotes the original predictor set as it is prior to the GLASSO step. After predictor variables are selected in the first step by GLASSO, then ACCA is applied to identify significant relationships. The additional descriptive measures that are examined here are the same as those in Chapter 5, for the same listed reasons. These include the canonical loadings of the response set and the canonical cross-loadings of the predictor set.

6.3.1 Results

The analysis is performed here with the high-dimensional predictor set along with, first of all, all response variables, then response variables associated with the learning and memory domains, followed by the response variables associated with the executive functioning and information processing/attention domains, and finally, the verbal and motor domains. Results are tabulated in Appendix E.

All Response Variables When all response variables were retained in the response set, only four predictors of the 235 were selected by GLASSO. ACCA was then performed and returned two canonical, uncorrelated relationships with canonical correlations of $\hat{\rho}_1 = 0.706$ and $\hat{\rho}_2 = 0.560$. The four variables that were selected by GLASSO were the predictor variables `age`, `Education`, `kmsk_cocopi`, and `fa_ic2_quad`. That is, the subject's age, number of years of education, their cocaine/opiate usage, and the quadratic term of the fractional anisotropy in the second subregion of the internal capsule, respectively.

First Relationship When the loadings of the response variate are examined in the first relationship, `HVLT_sum` and `HVLT_delay` and `BVMT_sum` and `BVMT_delay` are all negative in direction while the remaining nine are positive in direction. They are mid-range in strength in -0.20's for the HVLT assessments and -0.40's for the BVMT assessments. These variables with negative loadings correspond to the learning and memory domains. The strongest of the positive loadings are `GPeg_nondom`, `Trail_B`, `Trail_A` followed by `COWAT` and `GPeg_dom`, all in excess of 0.50.

Of the cross-loadings in the predictor set, the subject's age and cocaine/opiate usage are the strongest, both positive in direction indicating that as they increase, the response variate as a whole increases. The latter of the two is counter-intuitive. The only cross-loading that is negative is number of years of education but is fairly weak with a cross-loading of -0.21.

Second Relationship In the second relationship, all of the response loadings are positive. The strongest ones here are `WAIS_SymSrch`, `BVMT_sum`, and `BVMT_delay`, followed by `WAIS_DigSym`. These were roughly in the 0.70's range. The predictor set has only one negative cross-loading, which corresponds to the subject's cocaine/opiate use. While only -0.34, it is still the second strongest cross-loading in the predictor set. This inverse relationship with the response variate is unsurprising as cocaine/opiate use

is known to negatively affect cognitive function. The other three response variables have positive cross-loadings, of which education is the strongest, indicating a direct relationship with the response variate.

Learning and Memory Domains With the learning and memory domains combined, GLASSO selected 9 predictor variables. After ACCA was performed, only one significant canonical relationship was returned with a canonical correlation of $\hat{\rho}_1 = 0.730$. Of the response loadings, the largest contributors are those from the BVMT assessment. These are all positive and fairly strong, all being ≥ 0.55.

There were three cross-loadings in the predictor set that were negative, of which only Hepatitis C status and cocaine/opiate use were of notable strength. The inverse relationship with the learning and memory domains as a whole is noted again. The positive cross-loading of the education variable is also expected, as it is of positive strength as much as cocaine/opiate use is of negative strength (0.50). There are two interaction terms that are of moderate, positive strength that involve the brain volumetric measure of the putamen. While there is one interaction term that involves `HIV_stage`, this is positive of relatively low strength (0.19).

Executive Functioning and Information Processing/Attention Domains These domains combined returned only three predictor variables after GLASSO: age, cocaine/opiate use, and the quadratic mean diffusivity in the body of the corpus callosum. Interestingly, these three predictors returned two significant canonical relationships with canonical correlations of $\hat{\rho}_1 = 0.699$ and $\hat{\rho}_2 = 0.531$.

First Relationship The strongest response loadings were for `COWAT`, `Trail_B`, and `Trail_A`, and the other being much smaller. The three variables from the WAIS assessments are the smallest contributors to this response variate. In the predictor set, both age and past cocaine/opiate use are fairly strong and positive while the one quadratic term is negative but relatively small.

Second Relationship In the second relationship, all response variables had positive loadings but now with `WAIS_DigSym` and `WAIS_SymSrch` being the strongest. They are larger contributors in the second relationship than in the first with loadings in the high 0.80's. Both of these are, in particular, in the information processing/attention domain. The two predictor variables with the strongest cross-loadings are both negative and give an interesting relationship with this response variate and past cocaine/opiate use and the quadratic term.

Verbal and Motor Domains The verbal and motor domains combined returned 7 predictors using GLASSO. Then, only one significant relationship was returned by ACCA with a correlation of $\hat{\rho}_1 = 0.651$. The response loadings are all fairly large and positive with `COWAT` being the weakest at 0.47, which is associated with the verbal domain. The cross-loadings of the predictors are all positive, aside from the quadratic mean diffusivity in the body of the corpus callosum. However, this is also the weakest of the cross-loadings. The stronger cross-loadings were in the 0.30's range with age and three interaction terms involving the brain volumetric of the amygdala. This is followed in strength by two interaction

variables involving `HIV_stage` with two brain volumetric measures of, again, amygdala and whitematter.

6.3.2 Summary

To summarize these results, GLASSO significantly reduced the predictor space, making ACCA possible to apply without any further high-dimensional considerations. When all response variables are retained, it is difficult to see and understand the relationships with the domains, as they are combined together into the response variate. However, breaking down the response variables into domains that are grouped in pairs (as was done in Chapter 5) helps identify the more important predictor variables. Also shown are cases where there may exist more than one relationship within a pair of domains. In general, some of the more important predictors returned were age, education, and cocaine/opiate usage. Interestingly, none of the lower-order DTI and brain volumetric measures were ever retained by GLASSO. They were only included in ACCA as quadratic terms or interaction terms.

CHAPTER 7

CONCLUSION

What has been offered here is a new, data-adaptive method for Canonical Correlation Analysis (CCA) that estimates the number of significant canonical relationships while at the same time estimating the canonical weights themselves. The novelty in this approach is from a multivariate point of view in the Reduced-Rank Regression (RRR) setting, rather than the traditional sequential way. While plenty of literature documents CCA from a population version or in a large sample setting, only Regularized CCA (RCCA) has ever been offered as solution for dealing with high-dimensional issues. RCCA has computation limits, as it is expensive to cross-validate across a 2-dimensional grid, and completely neglects to consider the other canonical relationship aside from the strongest one. ACCA offers empirical support for alternative options to RCCA that do not have these issues.

This new Adaptive Canonical Correlation Analysis (ACCA) is built upon the theoretical findings of Bunea *et al.* from their formulation of the Rank Selection Criterion (RSC) [8]. RSC estimates the rank in a penalized fashion through a tuning parameter with the nice addition of the closed-form adaptive version of the tuning parameter. This eliminates any need to perform additional cross-validation. In order to draw the connection between the two, a weighted version of RSC has been developed here, the Weighted Rank Selection Criterion (WRSC), which requires a weight matrix and careful treatment of the tuning parameter. This was done in Chapter 3.

The choice of the weight matrix has been carefully examined from a large sample and high-dimensional point of view. The theoretics of RSC carry over to WRSC only through a true decorrelator weight matrix. The obvious choice is the sample residual covariance matrix, but this may not be practical in application. An alternative weight matrix is the sample response matrix which, while lacking some of the theoretical support, has demonstrated empirically to recover rank and estimate the coefficient matrix well, hence also well estimating the number of significant canonical relationships and the canonical weights. This is arguably a more suitable choice of weight matrix in the high-dimensional setting, as alternatives such as the pseudoinverse of the sample residual covariance or the regularized version of the sample residual covariance have been shown here to be, in computation, less feasible and may prove to be unstable in some settings.

Another approach to the high-dimensional setting is by adding a variable selection technique to reduce the dimension of the predictors prior to performing ACCA. Variable selection in multivariate response models is the same as group selection in univariate response

regression models. With the dimension reduced using a technique such as Group Lasso, ACCA may be then applied without further issue.

These concepts have been supported through simulation findings that indicate the applicability of ACCA in both the large sample and high-dimensional setting. Furthermore, true application has been shown through a neurocognitive dataset from HIV-positive patients. Treatment of this dataset has been both through a large sample view using only the original predictors, and through a high-dimensional view, using generated predictors. It is further analyzed with additional techniques from standard CCA, broken down by neurocognitive domains of interest. In the high-dimensional treatment, GLASSO is first applied to reduce the dimension before ACCA is used. This offers a complete picture of the application of these ideas.

Research pertaining to this composition is ongoing. First of all, while reasoning is provided here, rigorous proofs are required, particularly for the adaptive style tuning parameter for WRSC and ACCA. Rank consistency needs to be proven when the variance is estimated in the transformed model. Additional proofs are also required for the concepts applied to the high-dimensional setting. Exploration of other estimators of the error covariance may offer other alternatives with good rank and coefficient estimation. Additionally so, other two-step processes similar to [7] may give rise to more complete approaches that are better suited to the high dimensional setting.

APPENDIX A

THEOREMS

Theorem A.0.1. *Let* $\mathbf{S} \in \mathbb{R}^{m \times n}$ *of rank* $r(\mathbf{S}) = m$*. Then the minimum of the Euclidean norm*

$$tr\left[(\mathbf{S} - \mathbf{P})(\mathbf{S} - \mathbf{P})'\right]$$

over all matrices P *of size* $m \times n$ *with rank* $r(\mathbf{P}) = r \leq m$ *is acheieved when* $\mathbf{P} = \mathbf{M}\mathbf{M}'\mathbf{S}$ *where* $\mathbf{M} \in \mathbb{R}^{m \times r}$ *and has columns of* r *eigenvectors corresponding to the first* r *largest eigenvalues of* $\mathbf{S}\mathbf{S}'$*.*

Theorem A.0.2 (Poinecaré Separation Theorem)**.** *Let* $\mathbf{A}$ *be a* $m \times m$ *matrix and let* $\mathbf{U}$ *be a* $m \times k$ *matrix with* $k \leq m$ *such that* $\mathbf{U}'\mathbf{U} = \mathbb{I}_k$*. Then,*

$$\tilde{\lambda}_j \leq \lambda_j$$

where $\tilde{\lambda}_j$ *is the* j*-th largest eigenvalue of* $\mathbf{U}'\mathbf{A}\mathbf{U}$ *and* λ_j *is the* j*-th largest eigenvalue of* $\mathbf{A}$*.*

APPENDIX B

LARGE SAMPLE SIMULATION RESULTS

Table B.1: Experiment 1, Setup 1: Rank Recovery Rates with $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	0	0	0	0.01	0	0.25	0	0
	0.5	0.62	0.62	0.06	1	0.07	1	0	1
	0.1	1	1	0.56	1	0.49	1	0.02	1
0.4	0.9	0.08	0.08	0.16	0.96	0.11	1	0	0.93
	0.5	1	1	0.81	1	0.76	1	0.8	1
	0.1	1	1	0.58	1	0.75	1	0.81	1
0.6	0.9	0.99	0.99	0.35	1	0.48	1	0.03	1
	0.5	1	1	0.58	1	0.61	1	0.49	1
	0.1	1	1	0.48	1	0.7	1	0.35	1
0.8	0.9	1	1	0.69	1	0.61	1	0.58	1
	0.5	1	1	0.48	1	0.59	1	0.31	1
	0.1	1	1	0.44	1	0.57	1	0.2	1

Table B.2: Experiment 1, Setup 1: Average Mean Square Error with $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	17.2	17.2	72.0	33.4	62.5	29.3	179.4	72.3
	0.5	17.2	17.2	70.7	34.0	61.1	28.5	166.6	70.5
	0.1	16.2	16.2	66.7	32.6	58.6	27.3	176.1	72.5
0.4	0.9	16.7	16.7	70.1	34.5	60.0	29.7	167.3	71.3
	0.5	16.1	16.1	65.1	34.3	54.7	28.3	162.0	70.8
	0.1	15.8	15.8	61.6	33.3	52.3	28.7	160.9	70.9
0.6	0.9	16.3	16.3	65.0	33.2	57.3	28.8	164.3	68.7
	0.5	16.0	16.0	59.7	32.5	52.8	28.1	148.7	69.6
	0.1	15.9	15.9	62.6	33.3	50.8	28.1	146.4	68.0
0.8	0.9	16.3	16.3	67.8	33.3	56.7	27.5	163.4	72.5
	0.5	15.9	15.9	61.3	32.2	52.6	28.3	150.2	69.5
	0.1	15.9	15.9	60.9	32.4	50.4	28.0	147.8	67.6

Table B.3: Experiment 1, Setup 2: Rank Recovery Rates with $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	0	0	0	0	0	0.03	0	0
	0.5	0.62	0.15	0.06	0.94	0.07	1	0	0.83
	0.1	1	0.68	0.56	0.99	0.49	1	0.02	1
0.4	0.9	0.08	0	0.16	0.58	0.11	0.98	0	0.52
	0.5	1	1	0.81	1	0.76	1	0.8	0.99
	0.1	1	1	0.58	1	0.75	0.99	0.81	1
0.6	0.9	0.99	0.61	0.35	1	0.48	1	0.03	1
	0.5	1	1	0.58	1	0.61	1	0.49	1
	0.1	1	1	0.48	1	0.7	1	0.35	1
0.8	0.9	1	0.98	0.69	1	0.61	1	0.58	1
	0.5	1	1	0.48	1	0.59	1	0.31	1
	0.1	1	1	0.44	1	0.57	0.99	0.2	0.98

Table B.4: Experiment 1, Setup 2: Average Mean Square Error with $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	17.2	21.4	72.0	33.9	62.5	29.6	179.4	66.8
	0.5	17.2	19.8	70.7	35.1	61.1	29.0	166.6	71.4
	0.1	16.2	19.1	66.7	33.6	58.6	27.9	176.1	74.0
0.4	0.9	16.7	18.8	70.1	35.2	60.0	30.4	167.3	70.1
	0.5	16.1	17.6	65.1	35.3	54.7	28.8	162.0	72.0
	0.1	15.8	17.4	61.6	34.3	52.3	29.2	160.9	72.0
0.6	0.9	16.3	19.4	65.0	34.1	57.3	29.4	164.3	69.9
	0.5	16.0	17.5	59.7	33.4	52.8	28.6	148.7	70.5
	0.1	15.9	17.4	62.6	34.1	50.8	28.6	146.4	68.9
0.8	0.9	16.3	18.1	67.8	34.1	56.7	28.1	163.4	73.4
	0.5	15.9	17.4	61.3	32.9	52.6	28.8	150.2	70.5
	0.1	15.9	17.4	60.9	33.3	50.4	28.6	147.8	69.3

APPENDIX C

HIGH-DIMENSIONAL SIMULATION RESULTS

Table C.1: Experiment 2, Setup 1: Rank Recovery Rates with $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	1	1	0.63	0.99	0.62	0.99	0.54	0.99
	0.5	1	1	0.57	1	0.59	1	0.53	0.99
	0.1	1	1	0.49	0.99	0.52	0.98	0.63	0.99
0.4	0.9	1	1	0.53	1	0.6	1	0.51	1
	0.5	1	1	0.65	1	0.61	1	0.53	0.99
	0.1	0.99	0.99	0.5	1	0.63	0.99	0.42	1
0.6	0.9	0.99	0.99	0.47	0.98	0.59	0.99	0.46	1
	0.5	1	1	0.59	1	0.5	1	0.41	0.99
	0.1	1	1	0.52	0.99	0.57	1	0.39	1
0.8	0.9	1	1	0.56	1	0.53	1	0.38	1
	0.5	1	1	0.56	1	0.61	1	0.51	0.99
	0.1	1	1	0.65	1	0.59	0.99	0.45	0.99

Table C.2: Experiment 2, Setup 1: Average Mean Square Error with $\mathbf{\Gamma} = \mathbf{\Sigma}_e^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	29.7	29.7	110.7	74.4	98.5	66.7	281.4	170.5
	0.5	29.8	29.8	117.1	75.8	100.1	67.1	260.9	168.8
	0.1	30.4	30.4	127.3	82.5	104.9	67.9	259.3	164.7
0.4	0.9	29.7	29.7	118.7	79.1	102.5	69.2	273.1	165.8
	0.5	30.0	30.0	107.7	76.7	99.9	67.5	253.9	162.8
	0.1	30.2	30.2	121.2	78.1	96.1	66.0	265.2	162.0
0.6	0.9	29.8	29.8	121.1	77.3	99.2	67.9	259.8	162.4
	0.5	30.0	30.0	113.4	75.5	103.6	67.7	275.3	165.9
	0.1	30.0	30.0	124.1	81.0	98.4	65.3	275.1	162.8
0.8	0.9	29.9	29.9	116.0	77.3	105.6	70.6	286.1	166.3
	0.5	30.0	30.0	112.8	75.3	95.5	64.5	277.9	186.8
	0.1	29.4	29.4	111.9	79.1	92.0	61.6	261.5	160.2

Table C.3: Experiment 2, Setup 2: Rank Recovery Rates with $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	1	0.98	0.84	0.97	0.7	0.99	0.69	0.97
	0.5	1	0.99	0.77	0.97	0.61	0.98	0.67	0.97
	0.1	1	0.98	0.78	0.97	0.63	0.99	0.72	0.97
0.4	0.9	1	0.97	0.73	0.98	0.69	0.98	0.71	0.97
	0.5	1	0.99	0.64	0.96	0.65	1	0.65	0.98
	0.1	1	0.97	0.75	0.95	0.73	1	0.67	0.98
0.6	0.9	1	1	0.78	1	0.69	0.98	0.67	0.98
	0.5	1	0.95	0.71	0.99	0.59	0.98	0.74	0.99
	0.1	1	0.98	0.71	0.98	0.73	0.99	0.69	0.95
0.8	0.9	1	0.97	0.69	0.95	0.75	0.98	0.73	0.99
	0.5	1	0.98	0.73	0.98	0.7	0.97	0.79	0.99
	0.1	1	0.94	0.72	0.99	0.63	0.97	0.7	1

Table C.4: Experiment 2, Setup 2: Average Mean Square Error with $\mathbf{\Gamma} = \widehat{\mathbf{\Sigma}}_{YY}^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	9.0	9.5	28.7	20.7	30.1	20.3	69.9	47.2
	0.5	8.7	9.1	30.3	20.5	29.1	17.8	68.3	44.1
	0.1	9.0	9.5	30.1	21.6	29.6	18.7	64.9	44.1
0.4	0.9	9.0	9.4	30.6	21.1	28.0	19.1	66.3	45.4
	0.5	8.7	9.1	33.9	21.7	28.6	18.6	68.8	45.7
	0.1	8.7	9.2	31.7	22.4	27.3	18.1	68.4	46.2
0.6	0.9	8.9	9.4	30.8	21.8	27.4	18.1	70.6	47.2
	0.5	8.9	9.4	31.4	21.3	31.4	19.7	68.0	45.6
	0.1	9.0	9.5	31.1	21.0	28.3	19.3	66.6	45.0
0.8	0.9	8.9	9.5	32.9	22.2	26.4	18.7	61.9	42.0
	0.5	9.0	9.5	30.4	20.3	27.3	18.5	62.7	45.6
	0.1	9.1	9.6	31.9	22.3	30.2	19.5	66.0	44.2

Table C.5: Experiment 2, Setup 3: Rank Recovery Rates with $\mathbf{\Gamma} = (\widehat{\mathbf{\Sigma}}_{YY} + \delta \mathbb{I}_n)^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	1	0.87	0.63	0.66	0.62	0.67	0.54	0.51
	0.5	1	1	0.57	1	0.59	1	0.53	1
	0.1	1	0.88	0.49	0.51	0.52	0.54	0.63	0.31
0.4	0.9	1	1	0.53	0.76	0.6	0.87	0.51	0.59
	0.5	1	1	0.65	0.68	0.61	0.71	0.53	0.35
	0.1	0.99	1	0.5	1	0.63	1	0.42	0.98
0.6	0.9	0.99	1	0.47	1	0.59	1	0.46	0.98
	0.5	1	1	0.59	0.79	0.5	0.87	0.41	0.63
	0.1	1	1	0.52	1	0.57	1	0.39	1
0.8	0.9	1	1	0.56	1	0.53	1	0.38	1
	0.5	1	1	0.56	0.98	0.61	1	0.51	0.99
	0.1	1	1	0.65	1	0.59	1	0.45	1

Table C.6: Experiment 2, Setup 3: Average Mean Square Error with $\mathbf{\Gamma} = (\widehat{\mathbf{\Sigma}}_{YY} + \hat{\delta} \mathbb{I}_n)^{-1}$.

		IID		**AR(1)**		**H-AR(1)**		**UN**	
b	ρ_x	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**	**RSC**	**WRSC**
0.2	0.9	29.7	57.9	110.7	152.7	98.5	149.2	281.4	280.8
	0.5	29.8	29.8	117.1	84.8	100.1	75.5	260.9	179.2
	0.1	30.4	60.4	127.3	210.3	104.9	188.2	259.3	336.3
0.4	0.9	29.7	29.7	118.7	236.8	102.5	158.3	273.1	426.0
	0.5	30.0	30.0	107.7	262.7	99.9	237.8	253.9	525.1
	0.1	30.2	30.1	121.2	87.2	96.1	74.0	265.2	218.8
0.6	0.9	29.8	29.7	121.1	87.7	99.2	77.2	259.8	199.1
	0.5	30.0	30.1	113.4	345.6	103.6	210.1	275.3	672.6
	0.1	30.0	30.0	124.1	91.6	98.4	74.7	275.1	179.6
0.8	0.9	29.9	29.9	116.0	89.6	105.6	82.2	286.1	184.9
	0.5	30.0	30.0	112.8	211.2	95.5	77.6	277.9	271.2
	0.1	29.4	29.4	111.9	92.3	92.0	73.8	261.5	182.8

APPENDIX D

LARGE SAMPLE NEUROIMAGING DATA ANALYSIS RESULTS

Table D.1: Large Sample Neurocognitive ACCA Application with All Response Variables: Predictor Set Results

	Coeffs.				**Load.**				**Cross-Load.**			
Predictor Variable	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$	$\hat{\mathbf{f}}_3$	$\hat{\mathbf{f}}_4$	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$	$\hat{\mathbf{f}}_3$	$\hat{\mathbf{f}}_4$	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$	$\hat{\mathbf{f}}_3$	$\hat{\mathbf{f}}_4$
HIV_stage	0.10	0.22	-0.04	-0.41	0.23	0.36	0.11	-0.13	0.23	0.27	0.11	-0.06
hcv_current	-0.04	0.57	0.10	0.21	-0.22	0.38	0.39	0.20	-0.28	0.15	0.41	0.11
age	-0.12	0.07	-0.25	0.29	-0.25	0.45	0.21	0.32	-0.10	0.47	0.09	0.23
Education	0.06	0.66	-0.40	-0.07	0.24	0.21	-0.28	-0.15	0.28	0.28	-0.35	-0.15
kmsk_alc	-0.05	0.00	-0.08	-0.02	-0.23	-0.02	-0.07	0.06	-0.26	-0.01	-0.06	-0.13
kmsk_cocopi	-0.33	0.31	0.09	0.02	-0.36	0.16	0.39	0.31	-0.34	0.05	0.41	0.18
fa_cc_genu	-0.04	-0.10	-0.16	0.08	-0.07	0.03	-0.40	-0.02	-0.01	0.06	-0.42	0.13
fa_cc_body	-0.10	0.50	0.21	-0.19	0.08	0.16	-0.03	-0.12	0.08	0.20	-0.15	0.07
fa_cc_splenium	-0.04	-0.31	-0.26	-0.10	0.25	-0.26	-0.16	-0.03	0.23	-0.15	-0.18	0.13
md_cc_genu	0.48	-0.04	0.76	-0.09	0.00	0.11	0.25	-0.05	-0.09	-0.02	0.20	-0.20
md_cc_body	-0.50	-0.07	-0.60	-0.42	-0.19	0.06	-0.18	-0.17	-0.24	0.00	-0.08	-0.33
md_cc_splenium	-0.64	-0.14	-0.39	-0.34	-0.39	0.19	0.18	-0.08	-0.31	0.11	0.14	-0.23
fa_ic1	-0.13	0.40	-0.01	-0.24	0.03	0.11	-0.39	0.07	0.05	0.15	-0.35	0.15
fa_ic2	-0.22	0.03	-0.22	0.79	-0.19	0.21	-0.54	0.23	-0.17	0.18	-0.44	0.27
fa_ic3	-0.16	0.08	0.19	-0.74	0.00	-0.05	-0.17	-0.07	-0.03	-0.09	-0.04	0.08
fa_ic4	0.55	0.33	-0.32	0.91	0.07	-0.04	-0.24	0.10	0.06	-0.06	-0.07	0.17
md_ic1	-0.58	0.39	0.09	-0.50	-0.02	0.03	0.13	-0.17	-0.03	-0.02	0.12	-0.25
md_ic2	0.55	-0.26	0.26	0.50	0.18	-0.01	0.22	-0.14	0.16	0.03	0.17	-0.22
md_ic3	0.88	1.18	0.57	-0.72	0.17	0.18	-0.01	-0.15	0.08	0.11	-0.05	-0.26
md_ic4	-0.33	-1.01	-0.73	1.00	0.13	0.06	0.04	-0.17	0.05	0.02	0.03	-0.28
whitematter	0.36	0.01	0.18	1.24	0.05	0.01	-0.22	0.24	0.06	0.10	-0.29	0.21
cortex	-0.17	0.43	-0.10	-0.90	0.10	-0.04	-0.33	-0.08	0.11	0.02	-0.29	0.07
thalamus	-0.66	-0.64	-0.12	-0.19	0.00	-0.09	-0.26	0.01	0.06	0.10	-0.33	0.11
caudate	-0.40	-0.40	0.14	0.29	0.05	0.06	-0.06	0.10	0.01	0.05	-0.05	0.17
putamen	0.33	-0.79	-0.57	0.21	0.27	-0.04	-0.36	0.09	0.22	0.02	-0.31	0.14
pallidum	0.11	0.54	0.08	-0.84	0.07	0.10	-0.26	0.03	0.06	0.21	-0.28	0.05
hippocampus	0.22	0.51	0.38	0.30	0.12	0.02	-0.11	0.04	0.17	0.18	-0.19	0.13
amygdala	0.47	0.21	-0.04	-0.10	0.18	0.07	-0.18	-0.01	0.16	0.15	-0.15	0.11
cc_splenium	-0.13	-0.27	0.71	-0.32	-0.12	-0.06	0.09	-0.02	0.00	0.15	-0.03	0.09
cc_body	-0.30	-0.14	-0.27	0.08	0.12	-0.17	-0.09	-0.02	0.11	-0.06	-0.15	0.06
cc_genu	-0.14	-0.04	-0.61	-0.36	0.00	-0.02	-0.14	0.04	-0.02	0.00	-0.12	0.13

Table D.2: Large Sample Neurocognitive ACCA Application with All Response Variables: Response Set Results

	Coeffs.				**Load.**				**Cross-Load.**			
Response Var	$\hat{\mathbf{h}}_1$	$\hat{\mathbf{h}}_2$	$\hat{\mathbf{h}}_3$	$\hat{\mathbf{h}}_4$	$\hat{\mathbf{h}}_1$	$\hat{\mathbf{h}}_2$	$\hat{\mathbf{h}}_3$	$\hat{\mathbf{h}}_4$	$\hat{\mathbf{h}}_1$	$\hat{\mathbf{h}}_2$	$\hat{\mathbf{h}}_3$	$\hat{\mathbf{h}}_4$
HVLT_sum	0.51	0.24	0.33	-0.03	0.69	0.49	0.02	-0.11	0.45	0.23	0.10	-0.05
HVLT_delay	0.20	0.22	0.14	-0.30	0.63	0.41	-0.09	-0.21	0.25	0.09	0.04	-0.36
BVMT_sum	0.12	0.00	-0.31	-0.10	0.43	0.20	-0.79	-0.10	0.31	-0.02	-0.39	-0.17
BVMT_delay	0.00	0.03	-0.52	-0.08	0.42	0.20	-0.83	-0.13	0.27	0.07	-0.58	-0.09
WAIS_LNS	-0.01	0.29	0.18	-0.06	-0.04	0.47	-0.09	0.01	-0.08	0.29	0.17	0.04
WAIS_DigSym	0.39	-0.22	0.01	0.29	0.57	0.20	-0.35	0.57	0.38	-0.12	-0.14	0.21
WAIS_SymSrch	-0.06	-0.20	-0.31	0.15	0.31	0.26	-0.51	0.35	0.03	-0.12	-0.29	0.14
GPeg_dom	-0.03	0.37	-0.10	0.00	0.20	0.44	-0.23	0.60	0.02	0.45	-0.15	0.18
GPeg_nondom	-0.15	0.06	-0.14	0.24	0.02	0.48	-0.28	0.71	-0.18	0.21	-0.10	0.41
Trail_A	0.37	-0.04	0.34	0.47	0.30	0.24	0.11	0.83	0.15	0.14	0.21	0.55
Trail_B	-0.33	0.05	-0.13	0.25	-0.15	0.46	-0.20	0.52	-0.39	0.19	-0.01	0.45
COWAT	-0.29	0.43	0.00	-0.24	-0.26	0.67	-0.07	0.06	-0.28	0.38	0.08	-0.08
Animal	0.15	0.38	-0.19	0.08	0.26	0.68	-0.35	0.30	0.24	0.51	-0.32	0.26

Table D.3: Large Sample Neurocognitive ACCA Application with Learning and Memory Domain Variables: Predictor Set Results

Predictor Var	**Coeffs.**	**Load.**	**Cross-Load.**
HIV_stage	-0.01	0.04	0.09
hcv_current	0.07	-0.44	-0.42
age	0.15	-0.03	-0.01
Education	0.24	0.52	0.50
kmsk_alc	0.07	-0.08	-0.08
kmsk_cocopi	-0.48	-0.57	-0.49
fa_cc_genu	0.03	0.19	0.20
fa_cc_body	0.27	0.23	0.15
fa_cc_splenium	0.07	0.18	0.14
md_cc_genu	0.16	-0.10	-0.11
md_cc_body	-0.16	-0.11	-0.04
md_cc_splenium	-0.31	-0.20	-0.11
fa_ic1	0.12	0.16	0.16
fa_ic2	-0.33	0.08	0.05
fa_ic3	-0.12	-0.16	-0.10
fa_ic4	-0.15	-0.14	-0.06
md_ic1	-0.05	-0.04	0.01
md_ic2	-0.03	0.09	0.09
md_ic3	1.04	0.24	0.17
md_ic4	-1.11	0.11	0.10
whitematter	0.26	0.30	0.17
cortex	-0.10	0.23	0.16
thalamus	-0.08	0.29	0.21
caudate	-0.75	-0.01	-0.06
putamen	0.11	0.31	0.22
pallidum	0.41	0.31	0.21
hippocampus	0.18	0.27	0.20
amygdala	0.09	0.17	0.12
cc_splenium	-0.11	0.13	0.07
cc_body	0.30	0.29	0.15
cc_genu	-0.24	0.06	-0.01

Table D.4: Large Sample Neurocognitive ACCA Application with Learning and Memory Domain Variables: Response Set Results

Response Var	**Coeffs.**	**Load.**	**Cross-Load.**
`HVLT_sum`	0.28	0.64	0.50
`BVMT_sum`	0.42	0.87	0.75
`HVLT_delay`	0.21	0.69	0.38
`BVMT_delay`	0.35	0.88	0.62

Table D.5: Large Sample Neurocognitive ACCA Application with Executive and Information/Attention Domain Variables: Predictor Set Results

	Coeffs.		**Load.**		**Cross-Load.**	
Predictor Var	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_1$	$\hat{f}_2$
HIV_stage	-0.02	-0.15	0.00	0.04	0.03	0.07
hcv_current	-0.16	0.44	-0.33	0.45	-0.28	0.27
age	0.74	0.28	-0.05	0.66	0.03	0.59
Education	-0.36	0.14	0.13	-0.11	0.15	0.06
kmsk_alc	-0.17	-0.05	-0.29	0.09	-0.25	0.03
kmsk_cocopi	-0.53	0.29	-0.31	0.52	-0.26	0.28
fa_cc_genu	0.32	0.00	0.21	-0.04	0.19	0.04
fa_cc_body	-0.76	0.19	0.13	-0.02	0.14	0.11
fa_cc_splenium	0.44	-0.13	0.41	-0.30	0.32	-0.19
md_cc_genu	-0.02	-0.12	-0.22	0.10	-0.28	-0.04
md_cc_body	-1.09	-0.35	-0.41	-0.08	-0.37	-0.11
md_cc_splenium	-0.07	0.11	-0.38	0.33	-0.36	0.18
fa_ic1	-0.16	0.20	0.16	-0.06	0.17	0.03
fa_ic2	0.18	0.43	0.06	0.12	0.12	0.17
fa_ic3	-0.33	-0.35	0.03	-0.20	0.06	-0.12
fa_ic4	0.29	0.19	0.12	-0.16	0.13	-0.14
md_ic1	-0.38	0.31	-0.14	-0.04	-0.19	-0.09
md_ic2	0.23	-0.18	-0.06	-0.10	-0.09	-0.08
md_ic3	-0.21	-0.25	-0.11	-0.10	-0.16	-0.10
md_ic4	1.12	0.17	-0.12	-0.12	-0.19	-0.14
whitematter	0.39	0.63	0.18	0.06	0.20	0.11
cortex	-0.37	-0.35	0.11	-0.27	0.15	-0.14
thalamus	0.64	0.21	0.10	-0.11	0.18	0.04
caudate	0.31	0.20	0.11	0.01	0.13	0.06
putamen	0.75	-0.52	0.21	-0.26	0.24	-0.11
pallidum	-0.99	-0.54	-0.01	-0.08	0.08	0.07
hippocampus	-0.17	0.46	0.14	-0.05	0.21	0.09
amygdala	-0.20	-0.34	0.05	-0.17	0.15	0.01
cc_splenium	-0.26	-0.05	-0.01	0.08	0.10	0.21
cc_body	-0.10	0.07	0.11	-0.18	0.16	-0.05
cc_genu	0.01	-0.21	0.03	-0.05	0.09	0.02

Table D.6: Large Sample Neurocognitive ACCA Application with Executive and Information/Attention Domain Variables: Response Set Results

	Coeffs.		**Load.**		**Cross-Load.**	
Response Var	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_1$	$\hat{h}_2$
`Trail_B`	0.04	0.58	0.37	0.84	0.19	0.65
`COWAT`	-0.35	0.35	-0.09	0.76	-0.33	0.30
`WAIS_LNS`	0.01	0.30	0.14	0.60	0.09	0.33
`WAIS_DigSym`	0.57	-0.21	0.87	0.24	0.62	-0.10
`WAIS_SymSrch`	0.38	-0.08	0.70	0.34	0.43	0.01
`Trail_A`	0.31	0.31	0.61	0.45	0.45	0.42

Table D.7: Large Sample Neurocognitive ACCA Application with Verbal Fluency and Motor Domain Variables: Predictor Set Results

Predictor Var	**Coeffs.**	**Load.**	**Cross-Load.**
HIV_stage	0.33	0.24	0.16
hcv_current	0.30	0.10	0.09
age	0.15	0.11	0.31
Education	0.53	0.28	0.25
kmsk_alc	0.02	-0.12	0.01
kmsk_cocopi	0.06	-0.17	-0.02
fa_cc_genu	-0.05	0.12	0.15
fa_cc_body	0.36	0.22	0.22
fa_cc_splenium	-0.06	-0.13	-0.12
md_cc_genu	-0.29	-0.03	-0.10
md_cc_body	0.28	0.09	0.06
md_cc_splenium	0.20	0.06	0.02
fa_ic1	0.39	0.14	0.24
fa_ic2	-0.35	0.23	0.36
fa_ic3	0.54	0.07	-0.00
fa_ic4	-0.48	-0.05	-0.00
md_ic1	0.53	0.02	-0.10
md_ic2	-0.66	-0.09	-0.10
md_ic3	1.28	0.19	0.09
md_ic4	-1.27	0.08	-0.01
whitematter	-0.99	-0.07	0.13
cortex	0.84	0.12	0.14
thalamus	-0.25	-0.03	0.16
caudate	-0.27	0.03	0.09
putamen	-0.50	0.03	0.11
pallidum	0.90	0.14	0.28
hippocampus	-0.14	-0.03	0.16
amygdala	0.24	0.11	0.22
cc_splenium	-0.35	-0.01	0.14
cc_body	0.16	-0.00	0.02
cc_genu	0.46	0.10	0.12

Table D.8: Large Sample Neurocognitive ACCA Application with Verbal Fluency and Motor Domain Variables: Response Set Results

Response Var	**Coeffs.**	**Load.**	**Cross-Load.**
COWAT	0.63	0.72	0.34
Animal	0.48	0.72	0.26
Gpeg_dom	0.60	0.44	0.33
Gpeg_nondom	-0.09	0.43	-0.05
Trail_A	-0.46	0.05	-0.25

APPENDIX E

HIGH-DIMENSIONAL NEUROIMAGAING DATA ANALYSIS RESULTS

Table E.1: High-Dimensional Neurocognitive ACCA Application after GLASSO with All Response Variables: Predictor Set Results

	Coeffs.		**Load.**		**Cross-Load.**	
Predictor Variable	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$
age	0.58	0.23	0.68	0.19	0.46	0.21
Education	-0.17	0.51	-0.24	0.69	-0.21	0.43
kmsk_cocopi	0.56	-0.36	0.68	-0.59	0.46	-0.34
fa_ic2_quad	0.51	0.58	0.36	0.66	0.28	0.28

Table E.2: High-Dimensional Neurocognitive ACCA Application after GLASSO with All Response Variables: Response Set Results

	Coeffs.		**Load.**		**Cross-Load.**	
Response Variable	$\hat{\mathbf{h}}_1$	$\hat{\mathbf{h}}_2$	$\hat{\mathbf{h}}_1$	$\hat{\mathbf{h}}_2$	$\hat{\mathbf{h}}_1$	$\hat{\mathbf{h}}_2$
HVLT_sum	-0.09	0.09	-0.20	0.50	-0.14	0.19
HVLT_delay	-0.16	0.07	-0.29	0.49	-0.26	0.15
BVMT_sum	-0.22	0.23	-0.41	0.73	-0.35	0.46
BVMT_delay	-0.18	0.25	-0.43	0.72	-0.29	0.51
WAIS_LNS	0.05	0.12	0.15	0.44	0.07	0.24
WAIS_DigSym	-0.09	0.14	0.14	0.69	-0.14	0.28
WAIS_SymSrch	-0.02	0.18	0.09	0.74	-0.03	0.37
GPeg_dom	0.16	0.12	0.51	0.42	0.26	0.25
GPeg_nondom	0.24	0.10	0.68	0.50	0.38	0.20
Trail_A	0.22	-0.02	0.61	0.27	0.34	-0.04
Trail_B	0.32	0.12	0.63	0.51	0.50	0.24
COWAT	0.23	0.08	0.51	0.37	0.35	0.15
Animal	0.18	0.18	0.39	0.54	0.27	0.37

Table E.3: High-Dimensional Neurocognitive ACCA Application after GLASSO with Learning and Memory Response Variables: Predictor Set Results

Predictor Variable	**Coeffs.**	**Load.**	**Cross-Load.**
hcv_current	0.03	-0.58	-0.44
Education	0.49	0.64	0.50
kmsk_cocopi	-0.45	-0.69	-0.50
fa_cc_genu_quad	0.22	0.40	0.24
fa_ic2_quad	0.02	0.21	0.08
HIV_stage*cc_body	0.14	0.20	0.19
fa_cc_genu*putamen	0.16	0.48	0.30
md_cc_splenium*caudate	-0.48	-0.11	-0.09
md_ic3*putamen	0.30	0.48	0.32

Table E.4: High-Dimensional Neurocognitive ACCA Application after GLASSO with Learning and Memory Response Variables: Response Set Results

Response Variable	**Coeffs.**	**Load.**	**Cross-Load.**
HVLT_sum	0.18	0.56	0.31
HVLT_delay	0.20	0.64	0.34
BVMT_sum	0.42	0.91	0.72
BVMT_delay	0.42	0.92	0.71

Table E.5: High-Dimensional Neurocognitive ACCA Application after GLASSO with Executive and Information/Attention Response Variables: Predictor Set Results

	Coeffs.		**Load.**		**Cross-Load.**	
Predictor Variable	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$	$\hat{\mathbf{f}}_1$	$\hat{\mathbf{f}}_2$
age	0.72	0.59	0.81	0.21	0.58	0.12
kmsk_cocopi	0.51	-0.78	0.72	-0.64	0.49	-0.34
md_cc_body_quad	-0.31	-0.66	-0.13	-0.55	-0.07	-0.29

Table E.6: High-Dimensional Neurocognitive ACCA Application after GLASSO with Executive and Information/Attention Response Variables: Response Set Results

	Coeffs.		Load.		Cross-Load.	
Response Variable	$\hat{\mathbf{h}}_1$	$\hat{\mathbf{h}}_2$	$\hat{\mathbf{h}}_1$	$\hat{\mathbf{h}}_2$	$\hat{\mathbf{h}}_1$	$\hat{\mathbf{h}}_2$
Trail_B	0.57	0.04	0.69	0.51	0.51	0.27
COWAT	0.46	-0.09	0.70	0.25	0.35	0.09
WAIS_LNS	0.06	0.23	0.28	0.49	0.17	0.28
WAIS_DigSym	-0.38	0.54	0.00	0.85	-0.04	0.46
WAIS_SymSrch	-0.30	0.49	-0.01	0.87	-0.01	0.43
Trail_A	0.49	0.00	0.51	0.36	0.42	0.20

Table E.7: High-Dimensional Neurocognitive ACCA Application after GLASSO with Verbal and Motor Response Variables: Predictor Set Results

Predictor Variable	**Coeffs.**	**Load.**	**Cross-Load.**
age	0.60	0.54	0.37
md_cc_body_quad	-0.28	-0.22	-0.19
fa_ic2_quad	0.70	0.47	0.35
HIV_stage*whitematter	-0.26	0.54	0.27
HIV_stage*amygdala	0.69	0.56	0.26
fa_cc_genu*amygdala	0.67	0.46	0.31
fa_ic2*amygdala	-0.53	0.51	0.34

Table E.8: High-Dimensional Neurocognitive ACCA Application after GLASSO with Verbal and Motor Response Variables: Response Set Results

Response Variable	**Coeffs.**	**Load.**	**Cross-Load.**
COWAT	0.24	0.47	0.38
Animal	0.35	0.71	0.54
GPeg_dom	0.30	0.75	0.47
GPeg_nondom	0.26	0.81	0.40
Trail_A	0.27	0.71	0.43

BIBLIOGRAPHY

[1] T.W. Anderson (1951).“Estimating linear restrictions on regression coefficients for multivariate normal distributions.” *Annals of Mathematical Statistics,* 22, 327-351.

[2] T.W. Anderson (1999).“Asymptotic distribution of the reduced rank regression estimator under general conditions.” *Annals of Mathematical Statistics,* 27(4), 1141- 1154.

[3] T.W. Anderson (2002). “Specification and misspecification in reduced rank regression.” *Sankya*, (64), Series A, 193-205.

[4] P.J. Basser and C. Pierpaoli (1996). “Microstructural and phsyiological features of tissues elucidated by quantitative-diffusion-tensor MRI.” *Journal of Magnetic Resonance, Series B.* 111, 209-219.

[5] P. Bickel and E. Levina (2006). “Regularized estimation of large covariance matrices.” *Annals of Statistics*, 36(1) 199.

[6] P. Bickel and E. Levina (2008). “Covariance regularization.” *The Annals of Statistics*, 36(6), 2577-2604.

[7] F. Bunea, Y. She, and M. Wegkamp (2011). “Joint variable and rank selection for parsimonious estimation of high-dimensional matrices.” *Submitted,* http://arxiv.org/abs/1110.3556.

[8] F. Bunea and Y. She and M. Wegkamp (2011). “Optimal selection of reduced rank estimators of high-dimensional matrices.” to appear in the *Annals of Statistics*.

[9] F. Bunea, Y. She, H. Ombao, A. Gongavatana, K. Devlin, and R. Cohen (2010). “Penalized least squares methods and their application to neuroimaging.” to appear in *NeuroImage*.

[10] Chen *et al.* (2010) “Shrinkage Algorithms for MMSE Covariance Estimation.” *IEEE Trans on Sign. Proc.* 58(10);

[11] M.L. Eaton and M.D. Perlman (1973). “The Non-Singularity of Generalized Sample Covariance Matrices.” *The Annals of Statistics,* 1(4) 710-717.

[12] B. Efron (1982). “Maximum likelihood and decision theory.” *Annals of Statistics.* 10, 340Ű356.

[13] B. Efron and C. Morris (1977). Stein's paradox in statistics. *Scientific American*, 236(5):119-127.

[14] G. Givens and J. Hoeting (2005). *Computational Statistics.* John Wiley & Sons, Inc., NJ.

[15] A. Gongvatana, R. Cohen, S. Correia, K. Devlin, J. Miles, U. Clark, M. Westbrook, G. Hana, H. Ombao, B. Navia, D. Laidlaw, and K. Tashima (2010). Impact of Hepatitis C and HIV Coinfection on Cerebral White Matter Integrity. *Neurology,* under review.

[16] I. González and S. Déjean and P. Martin and A. Baccini (2008). "CCA: An R Package to Extend canonical Correlation Analysis." *Journal of Statistical Software,* 23(12), 1-14. http://www.jstatsoft.org/

[17] J. Hair, R. Anderson, R. Tatham, W. Black. (1998). *Multivariate Data Analysis, 5th edition.* Prentice Hall, Inc.

[18] J. Hair, R. Anderson, R. Tatham, and W. Black. (1998). "Canonical Correlation: A Supplement to Multivariate Data Analysis." *Multivariate Data Analysis, 5th edition.* Prentice Hall, Inc. Retrieved from http://www.mvstats.com/

[19] J. Hair, R. Anderson, R. Tatham, and W. Black. (1998)."Advanced Diagnostics for Multiple Regression: A Supplement to Multivariate Data Analysis." *Multivariate Data Analysis, 5th edition.* Prentice Hall, Inc. Retrieved from http://www.mvstats.com/

[20] C.H. Hinkin, S.A. Castellon, A.J. Levine, T.R. Barclay, and E.J. Singer (2008). Neurocognition in Individuals Co-Infected with HIV and Hepatitis C. *J Addict Dis.* 27(2): 11-17.

[21] A. F. Hoerl and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

[22] H. Hotelling (1936). "Relations between two sets of variates." *Biometrika.* Vol. 28, 321-377.

[23] A.J. Izenman (1975). "Reduced-Rank Regression for the Multivariate Linear Model." *Journal of Multivariate Analysis,* 5, 248-2762.

[24] A.J. Izenman (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning.* New York: Springer.

[25] R. Johnson and D. Wichern (2002). *Applied Multivariate Statistical Analysis.* Prentice-Hall, Inc., NJ.

[26] Ledoit and Wolf (2004). "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis.* 88,2, 365-411.

[27] S.E. Leurgans and R.A. Moyeed and B.W. Silverman (1993). "Canonical Correlation Analysis when the Data are Curves." *Journal of the Royal Statistical Society B,* 55(3), 725-740.

[28] W. Mittenberg and S. Motta (1993). Effects of chronic cocaine abuse on memory and learning. *Archives of Clinical Neuropsychology,* 8(6), 477-483.

[29] F.A. Nielsen and L.K. Hansen and S.C. Strother (1998). "Canonical ridge analysis with ridge parameter optimization," *NeuroImage* 7, S758.

[30] S.H. Patel, D.L. Kolson, G. Glosser, I. Matozzo, Y. Ge, J.S. Babb, L.J. Mannon, and R.I. Grozzman (2002) Correlation between Percentage of Brain Parenchymal Volume and Neurocognitive Performance in HIV-Infected Patients. *AJNR Am J. Neuroradiol.* 23:543-549.

[31] C.R. Rao (1978). "Matrix Approximations and Reduction of Dimensionality in Multivariate Statistical Analysis." *Multivariate Analysis V,* Proceedings of the fifth international symposium of multivariate analysis; P.R. Krishnaiah Editor, North-Holland Publishing.

[32] G. Reinsel and R. Velu (1998). *Multivariate Reduced-Rank Regression.* New York: Springer.

[33] A. Rencher (2000). *Linear Models in Statistics.* John Wiley & Sons, Inc., New York.

[34] P.M. Robinson (1973). "Generalized canonical analysis for time series." *Journal of Multivariate Analysis,* 3, 141-160.

[35] P.M. Robinson (1974). "Identification, estimation and large sample theory for regression containing unobservable variables." *International Economic Review,* 15, 680-692.

[36] E.L. Ryan, S. Morgello, K. Isaacs, M. Phil, M. Naseer, P. Gerits, and the Manhattan HIV Brain Bank (2004). Neuropsychiatric impact of hepatitis C on advanced HIV. *Neurology*, 62(6), 957-962.

[37] Y. She (2011) "An Iterative Algorithm for Fitting Nonconvex Penalized Generalized Linear Models with Grouped Predictors." *Computational Statistics & Data Analysis,*" Available online 23 November 2011, ISSN 0167-9473, 10.1016/j.csda.2011.11.013. http://www.sciencedirect.com/science/article/pii/S0167947311004105

[38] C. Stein (1956). "Inadmissibility of the usual estimator for the mean of a multivariate distribution." In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* Volume 1, Berkeley, pp. 197Ű206. University of California Press.

[39] D. Stewart and W. Love (1968). *A General Canonical Correlation Index. Psychological Bulletin,* 70 160Ű63.

[40] D. Stuss and R.T. Knight (Editors) (2002). *The Frontal Loabes.* New York: Oxford University Press.

[41] H.D. Vinod (1976). "Canonical Ridge and Econometrics of Joint Production." *J. Econometrics*, 4, 147-166.

[42] C. Watson, M. Kirkcaldie, and G. Paxinos (2010) *The Brain: An Introduction to Functional Neuroanatomy.* Academic Press, London, UK.

[43] Y. Wu, P. Storey, B.A. Cohen, L.G. Epstein, R.R. Edelman, and A.B. Ragin (2006) "Diffusion Alterations in Corpus Callosum of Patients with HIV." *AJNR Am J Neuroradiol.* 27, 656-660.

[44] M. Yuan and Y. Lin (2006) "Model selection and estimation in regression with grouped variables," *JRSSB*, 68, 49-67.

[45] "HIV-Associated Dementia and Other Neurocognitive Disorders."(2011) *AIDS Education and Training Centers Nation Resource Center (AETC NRC).* Retrieved from http://www.aids-ed.org/aidsetc?page=cg-802_dementia.

BIOGRAPHICAL SKETCH

The author, born in South Korea, adopted by Beth and Jerome Geis, was raised in the frozen tundra, the Twin Cities of Minnesota, with her siblings: Jason, Joan, and Janice Geis. After dropping out of high school, she returned to higher education at Augsburg College in Minneapolis where she completed a Bachelor of Arts in Mathematics and a Bachelor of Science in Actuarial Science. She continued her education at Northern Illinois University, receiving a Master of Science in Applied Probability and Statistics. Following this, she moved to Tallahassee, Florida and was enrolled at Florida State University where she earned a second Master of Science in Biostatistics. After the completion of this manuscript, a Doctor of Philosophy in Biostatistics was also conferred on her. She currently resides in San Diego, California with her cat, Murs, working in genomic and genetic analysis research in molecular diagnostics and is forever indebted to many great statisticians and mathematicians, including two from her mathematical beginnings: Mr. Donald Holmes, her middle school teacher, and Dr. Kenneth Kaminsky, professor at Augsburg College.