The Analysis and Advanced Extensions of Canonical Correlation Analysis

Daniel V. Samarov

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2009

Approved by

Advisor: Dr. J. S. Marron

Co-Advisor: Dr. Yufeng Liu

Co-Advisor: Dr. Alexander Tropsha

Reader: Dr. Perry Haaland

Reader: Dr. D. G. Kelly

Reader: Dr. Andrew Nobel

ABSTRACT

DANIEL V. SAMAROV: The Analysis and Advanced Extensions of Canonical Correlation Analysis
(Under the direction of J. S. Marron, Yufeng Liu and Alexander Tropsha)

Drug discovery is the process of identifying compounds which have potentially meaningful biological activity. A problem that arises is that the number of compounds to search over can be quite large, sometimes numbering in the millions, making experimental testing intractable. For this reason computational methods are employed to filter out those compounds which do not exhibit strong biological activity. This filtering step, also called virtual screening reduces the search space, allowing for the remaining compounds to be experimentally tested.

In this dissertation I will provide an approach to the problem of virtual screening based on Canonical Correlation Analysis (CCA) and several extensions which use kernel and spectral learning ideas. Specifically these methods will be applied to the protein-ligand matching problem.

Additionally, theoretical results analyzing the behavior of CCA in the High Dimension Low Sample Size (HDLSS) setting will be provided.

CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

Recent advances in biology, genetics and chemistry have led to an influx in the amount of information available on a wide variety of biological processes. A major issue facing scientists is finding meaningful ways of utilizing this data to better understand the mechanisms of the diseases that affect humans. A key element in dealing with the challenges involved in understanding and analyzing this type of information and the unique problems associated with them is the development of new statistical methodology.

Of interest to scientists is using the many different ways of measuring or viewing the same (or similar) biological, genetic or chemical process in order to better understand the key elements driving them. Consider the following example:

In the field of cheminformatics, drug discovery is a key step in the process of identifying compounds which may have potentially meaningful biological activity as related to a particular disease process. The process of drug discovery typically begins with the identification of a new or existing drug target, typically these targets are proteins. Proteins are large organic compounds composed of amino acids and are the building blocks from which all cells are built and are responsible for almost all cell function. The two predominant families of target proteins in drug discovery are G-protein-coupled receptors (GPCR) and protein kinases. About half of all known drugs work through GPCR's.

GPCR's belong to the family of transmembrane receptors, proteins that span the cell membrane connecting the inside of the cell with the outside of the cell. These

transmembrane receptors bind extracellular signaling molecules, called ligands. Ligands include other proteins and small peptides (short sequences of amino acids), as well as derivatives of amino acids and fatty acids. Once bound these signaling molecules set off a chain of intracellular signaling events. These signaling events are generally mediated by protein kinases and lead to the alteration of some target proteins ultimately leading to a change in cell behavior.

The reason GPCR's and protein kinases are so important is that in both normal and abnormal cell activity, they are used as lines of communication. In the event of abnormal cell activity they are natural control points.

Consider the case where the target is a novel GPCR, ligands are then screened for their ability to inhibit or stimulate that GPCR. A problem that arises is that the number of compounds to search over can be quite large, sometimes numbering in the millions. Subsequently, experimental verification of protein-ligand interaction can be extremely time consuming and costly or in some cases simply not possible due to time and/or cost constraints. For this reason computational methods are employed to filter out those compounds that do not exhibit a strong relationship with a given receptor. This filtering step reduces the search space allowing for the remaining compounds to be experimentally tested.

A motivating example throughout this dissertation will be the prediction of protein-ligand binding, utilizing descriptive variables associated with these compounds. These descriptive variables, from here on referred to as descriptors, include information related to the electronic attributes, hydrophobicity, and steric properties of the molecules. The motivation for our model and its extensions will be based on the task of modeling these relationships. This approach to the prediction of molecular function and interaction is known as quantitative structure-activity relationship (QSAR) modeling. For a good introduction and overview of the theory, practice and history of QSAR see Selassie (2003).

In this example the proteins and ligands are represented by a set of descriptors with

the number of descriptive variables typically ranging from 150 to as many as 800 or more. The prediction problem can be generally stated as follows: for a set of n known protein ligand pairs, with d_X and d_Y descriptors, given a new protein we want to be able to predict what ligand will bind to it. Let $\mathbf{x}_i \in \mathbb{R}^{d_X}$ and $\mathbf{y}_i \in \mathbb{R}^{d_Y}$, $i = 1, \ldots, n$ denote a protein ligand pair. The sample of pairs is collected in matrices $\mathbf{X} \in \mathbb{R}^{n \times d_X}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_Y}$ with $\mathbf{x}_i$ and $\mathbf{y}_i$ as the descriptors for a row.

Our approach to this problem is based on the structural relationship between these molecules. Specifically, that there is a strong (complementary) relationship in the stereochemical layout (the relative spatial arrangement of atoms within a molecule) between the protein and its ligand(s). Thus, if we can find a way to align the space of proteins and ligands, then we may be able to exploit this structural relationship to predict which pairs match up.

1.1 General Framework

Casting the protein-ligand matching problem into a general framework, following the discussion of Bach and Jordan (2002) and Fukumizu *et al.* (2007), our example consists of two multivariate random variables X and Y belonging to $\mathbb{R}^d$. In the context of our example these random variables correspond, respectively, to proteins and ligands. Lets and ligands. Let $f_X \in \mathcal{H}_X$ and $f_Y \in \mathcal{H}_Y$ be mappings from $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$ to $\mathbb{R}$, where $\mathcal{H}_X$ and $\mathcal{H}_Y$ are spaces of functions. The type of functions we consider are, for example, bilinear maps, $f_X(X) = \langle X, \mathbf{w}_X \rangle$ and $f_Y(Y) = \langle Y, \mathbf{w}_Y \rangle$. Define $\mathcal{S} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ to be a function measuring the similarity between two random variables. An example of a similarity measure is Pearson correlation. It is important to note that the notation $\mathcal{S}$ is defined here in terms of the population random variables X and Y, as opposed to their sample counterparts. When referring to the sample, i.e. empirical similarity measure we will write $\widehat{\mathcal{S}}$ (so for example we would write $\widehat{\text{corr}}$ for sample correlation).

Returning to our example, we want to find functions f_X and f_Y such that the similar-

ity between proteins and ligands is maximized, this leads to the following optimization problem

$$\rho_{\mathcal{H}} = \max_{f_X \in \mathcal{H}_X, f_Y \in \mathcal{H}_Y} \mathcal{S}(f_X(X), f_Y(Y)) \tag{1.1}$$

where the subscript $\mathcal{H} = (\mathcal{H}_X, \mathcal{H}_Y)$ denotes the spaces of functions over which the similarity is being maximized.

The selection of a meaningful measure of similarity is context dependent. All similarity measures have relevance in certain circumstances. Examples of similarity measures include correlation, covariance, and mutual information. The one which we will focus on is correlation.

Defining $\mathcal{H}_X$ and $\mathcal{H}_Y$ to be the Hilbert spaces of bilinear maps taking the form $f_X(X) = \langle X, \mathbf{w}_X \rangle$ and $f_Y(Y) = \langle Y, \mathbf{w}_Y \rangle$ respectively, the problem as stated in (1.1) then becomes the well known Canonical Correlation Analysis (CCA) (Hotelling (1936)). The optimization problem in (1.1) then takes the form

$$\rho_{\mathcal{H}} = \max_{\mathbf{w}_X, \mathbf{w}_Y} \operatorname{corr}(\langle X, \mathbf{w}_X \rangle, \langle Y, \mathbf{w}_Y \rangle) = \max_{\mathbf{w}_X, \mathbf{w}_Y} \frac{\operatorname{cov}(\langle X, \mathbf{w}_X \rangle, \langle Y, \mathbf{w}_Y \rangle)}{\sqrt{\operatorname{var}(\langle X, \mathbf{w}_X \rangle)} \sqrt{\operatorname{var}(\langle Y, \mathbf{w}_Y \rangle)}} \tag{1.2}$$

The general framework of (1.1) will allow for a natural extension of linear CCA to kernel CCA (KCCA) by defining $\mathcal{H}_X$ and $\mathcal{H}_Y$ to be reproducing kernel Hilbert spaces (RKHS). This will be discussed in further detail in Chapter 3.

CCA has a number of appealing properties, including

1. Extensions to kernel based methods, (i.e. Kernel CCA (KCCA)), discussed in Kuss and Graepel (2003), Hardoon *et al.* (2004) and Bach and Jordan (2002).

2. An intuitive geometric interpretation to the cosine of the angle between two vectors, discussed in Anderson (2003), with extensions to KCCA, discussed in Kuss and Graepel (2003).

3. Connections to Mutual Information (MI) (Kullback (1997)). In the case when the

data is known to be normally distributed this can be shown directly. A connection between MI and KCCA is discussed in Bach and Jordan (2002).

4. An extension of CCA to more than two data sets is presented in Kettenring (1971).

5. Connection to linear discriminant analysis (LDA) (Bie (2005) and Hastie *et al.* (1995))

An illustration of the protein-ligand matching problem may help in the understanding of CCA and its application to this problem as well as its extension to other similar problems.

1.2 Two Space Toy Example

Consider the protein-ligand matching problem as outlined above. For this toy example we set $n = 10$ and $d = 2$. Suppose the descriptors for this toy example are Molecular Weight (MW) and Surface Area (SA) of the molecule. Recall that each row of $\mathbf{X}_{(10\times 2)}$ and each row of $\mathbf{Y}_{(10\times 2)}$ corresponds to an observation, a protein or a ligand respectively, and the columns correspond to the descriptors MW and SA. The pairs are identified by a unique label, corresponding to ID's from the Protein Data Bank (PDB) (www.pdb.org). Figure 1.1 shows the two toy data sets.

From Figure 1.1 it can be seen that the distribution of points in the two spaces are quite similar in the sense that the location of corresponding points in the two spaces are close. The points connected to 11gs (red) by dashed black lines are its three nearest neighbors. The cyan points are neighbors shared in both spaces and the blue and purple points are mismatched. Two of three neighbors are shared in common (in the Euclidean sense).

Consider the case where the red point in ligand space is not observed and the task is to predict its value. Using the average of the points in ligand space that correspond to the nearest neighbors of the point 11gs in the protein space (points highlighted in cyan

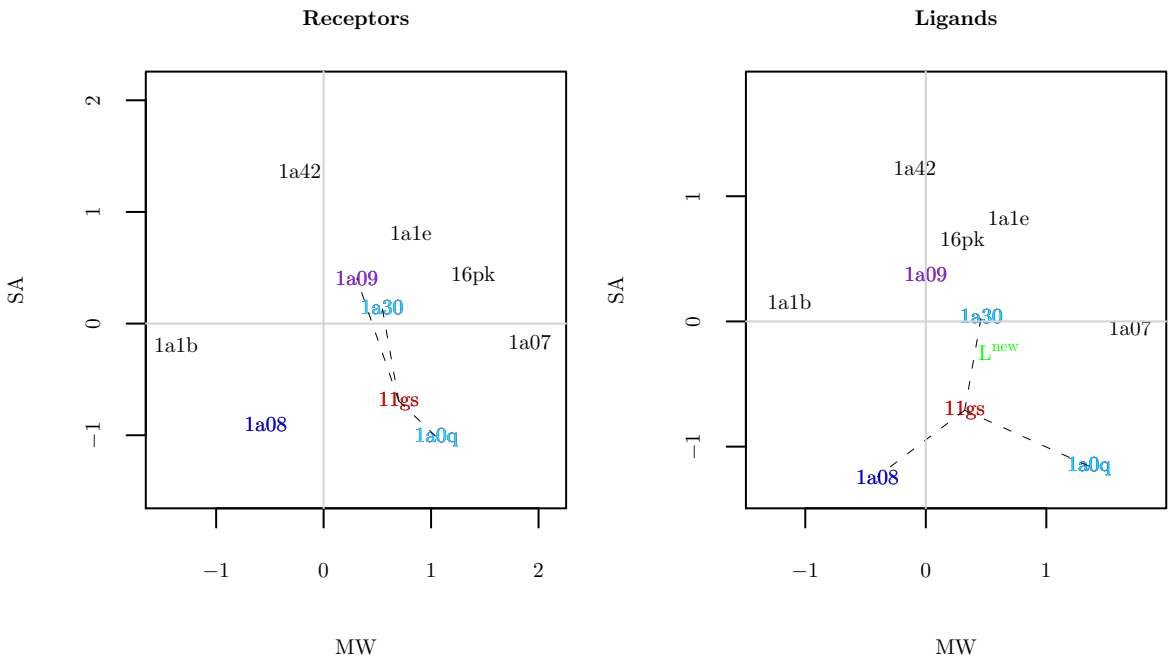

Figure 1.1: *Toy example data. The points highlighted in red correspond to the protein ligand pair 11gs, and the points connected to it by dashed black lines are its three nearest neighbors in each space. The observations highlighted cyan are neighbors in both spaces, and those highlighted in blue and purple are neighbors only in the protein, and ligand spaces respectively. The green point L^{new} in the ligand space corresponds to a simple weighted average of the cyan points and the purple point; i.e. of the nearest neighbors of 11gs in the protein space.*

and purple in ligand space) would yield a relatively poor prediction despite the strong apparent similarity between the two distributions of points. This dissertation studies more sophisticated approaches to exploiting this similarity.

In Section 1.1 the idea of similarity between two distributions was introduced. In our current example the type of similarity measure that is needed is one that tells us how well aligned the two spaces are. The functions f_X and f_Y we consider will be ones which place appropriate weights on the features (i.e. descriptors) that best align the two distributions.

To motivate our approach and justify why CCA is an appropriate method to address this problem we start by considering a simple example. Figure 1.2 shows four data

sets, each consisting of two spaces, X and Y with $d_X = d_Y = 1$ at different levels of correlation. For each data set two quite different views are considered. The top row of plots are conventional scatterplots of the data and the bottom set of plots are *connectivity plots* which provide a different view of the association between pairs of points. In the connectivity plots points are shown as the (green) segments connecting the x-coordinates (blue points) and y-coordinates (red points). This view highlights the similarity of the pairs.

As correlation increases (moving from the left panels to the right) the difference between the values of the points in the X and Y space becomes smaller. This is reflected in the top set of plots in Figure 1.2 as the observations tend to fall closer to the 45 degree line in the right hand panels. In the bottom set of plots the dashed green lines become increasingly parallel to each other. Based on these observations maximizing the correlation between the sets of points $\mathbf{x}$ and $\mathbf{y}$ is equivalent to maximizing their coordinate-wise alignment.

Yet another way to interpret correlation is as the cosine of the angle between the vectors $\mathbf{x}$ and $\mathbf{y}$ (Anderson (2003)), assuming they are mean centered. This relationship is easy to verify. Using the idea of projections provides concreteness to the interpretation of correlation as a measure of alignment. Define the projection coefficient p to be the scalar such that the vector $p\mathbf{y}$ is orthogonal to $\mathbf{x} - p\mathbf{y}$; solving the following for p

$$0 = p\mathbf{y}^T(\mathbf{x} - p\mathbf{y}) = p(\mathbf{y}^T\mathbf{x} - p\mathbf{y}^T\mathbf{y}),$$

we have that, $p = \mathbf{y}^T\mathbf{x}/\mathbf{y}^T\mathbf{y}$. Next decompose $\mathbf{x}$ as $\mathbf{x} = (\mathbf{x} - p\mathbf{y}) + p\mathbf{y}$, see Figure 1.3. The absolute value of the cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$ is the same as the length of $p\mathbf{y}$ divided by the length of $\mathbf{x}$;

$$\cos(\theta) = \sqrt{p\mathbf{y}^T(p\mathbf{y})/\mathbf{x}^T\mathbf{x}} = \mathbf{x}^T\mathbf{y}/\sqrt{(\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})} = \widehat{\text{corr}}(\mathbf{x}, \mathbf{y}).$$

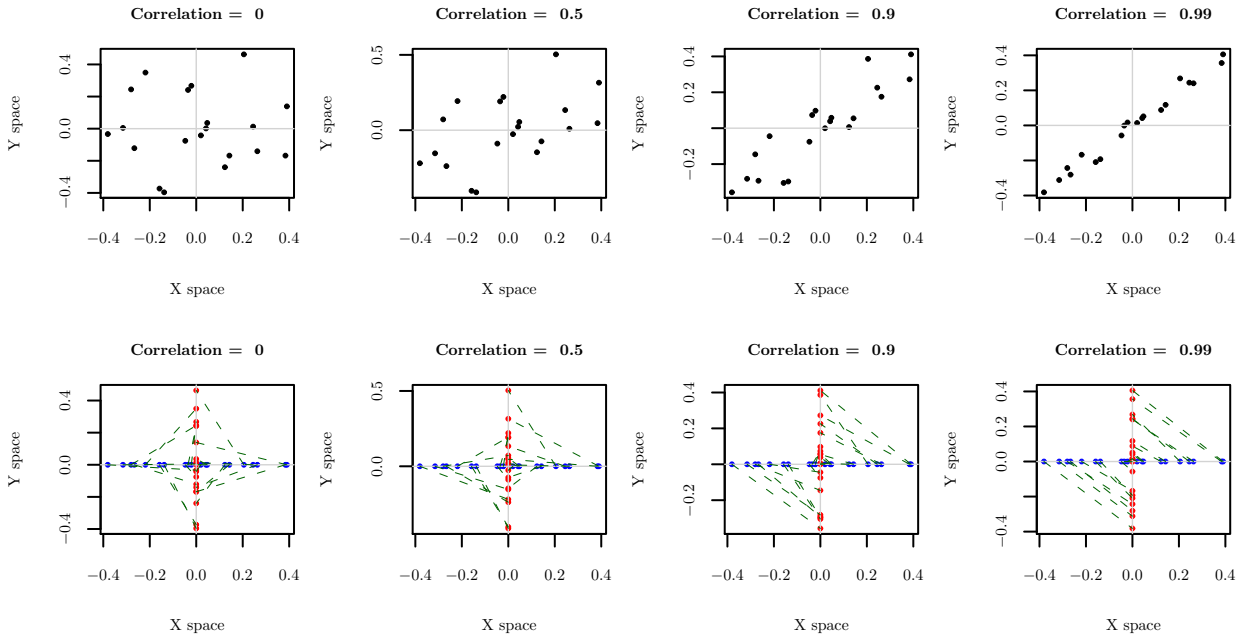

Figure 1.2: *Four bivariate toy data sets, with differing correlation. The top plots correspond to the scatterplot view of data and the bottom plots are connectivity plots of the data. The blue points, on the bottom set of plots, are the x coordinate values and the red points are the corresponding y coordinate values. In the top set of plots as correlation increases points begin falling closer to the 45 degree line. In the bottom set of plots the dashed green lines become increasingly parallel to each other.*

So in terms of (1.2) above, maximizing the correlation between $\mathbf{x}$ and $\mathbf{y}$ is equivalent to minimizing the angle between them. As the angle goes to zero the closer each pair of coordinates in both n vectors becomes (modulo a scale factor). This can be seen in Figure 1.3 as the angle goes to zero $\mathbf{x} - p\mathbf{y}$ goes to zero.

With an intuitive grasp of the relationship between correlation and alignment we return to the protein ligand example of Figure 1.1 at the beginning of this section. Solving for $\mathbf{w}_X$ and $\mathbf{w}_Y$ in (1.2), gives us the direction vectors shown in Figure 1.4 (details of these derivations will be discussed in Chapter 2). What is important to notice is how the distribution of points along the first (red) and second (green) canonical directions in both protein and ligand space are quite similar. This is due to the property of alignment that arises naturally from maximizing the correlation.

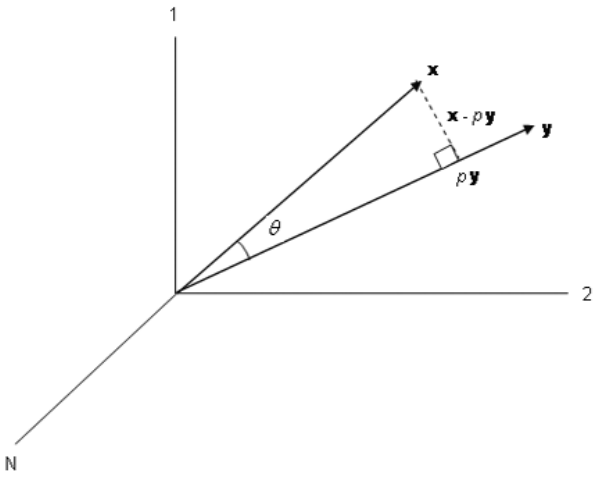

Figure 1.3: *An illustration of the relationship between correlation and angle between two vectors. Note that we assume that the vectors have been mean centered.*

Figure 1.5 shows the projections of the data onto the first two canonical vectors (note that separate directions are found in protein and ligand space). We can see that with the slight modification in alignment that has resulted from the CCA projections, the point 11gs now shares the same neighbors in both spaces. In particular note that now the predicted value in the projected ligand space is much closer to the actual value (again using the simple average).

This is a simplified example and in most cases the relationship between points in different spaces may be far more complicated. In coming sections we begin with the simplest case scenario, i.e. standard CCA and related methods. This is used as a starting point to motivate and develop methodology and theory appropriate for increasingly complex problems. Along the way we address the strong and weak points of these various methods.

1.3 Benchmark Data Sets

Two virtual drug screens will be used as a benchmark for testing the methods developed in this dissertation:

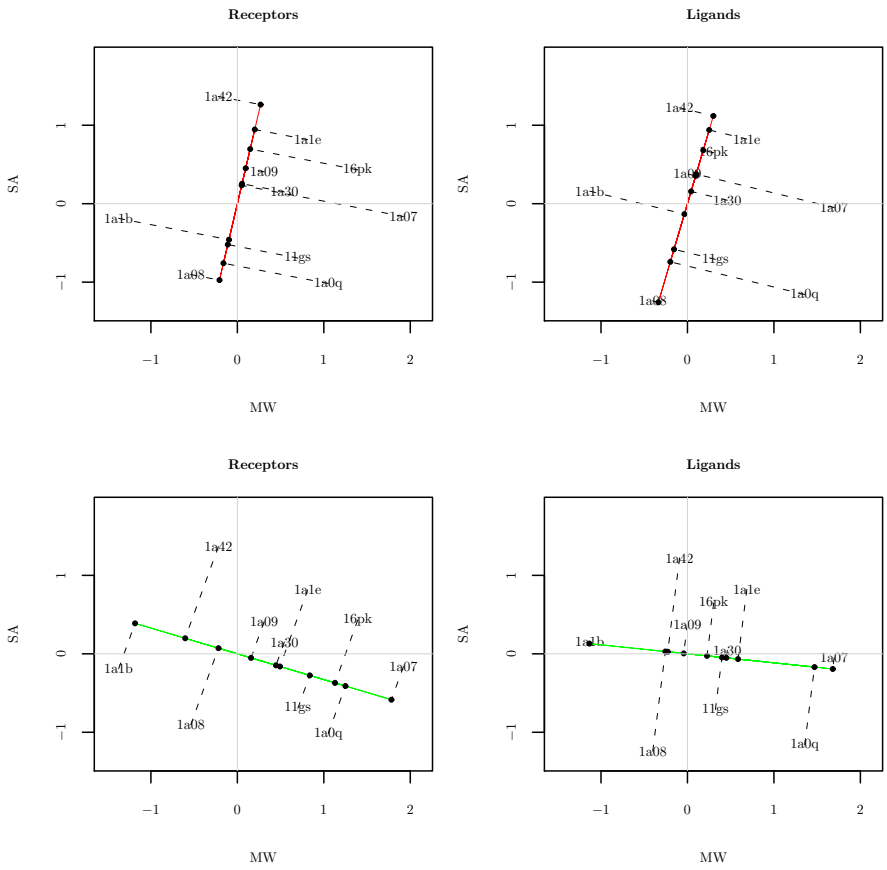

Figure 1.4: *The direction vectors and the projected value of each point. The top row of plots shows the first direction vector, in red, and the projections onto it. The bottom row of plots show the second direction vector, in green, and the projection onto it.*

1. A set of 800 chemically, and functionally diverse protein-ligand pairs obtained from the PDBbind Database (Wang *et al.* (2004)). These compounds are described by a set of 150 descriptors. We will refer to this data set as the RLP800 data.

2. The World Drug Index (WDI) (Daylight (2004)) database which contains approximately 54,000 drug candidates (ligands). Each compound in the WDI is described by the same set of 150 descriptors as the RLP800 data.

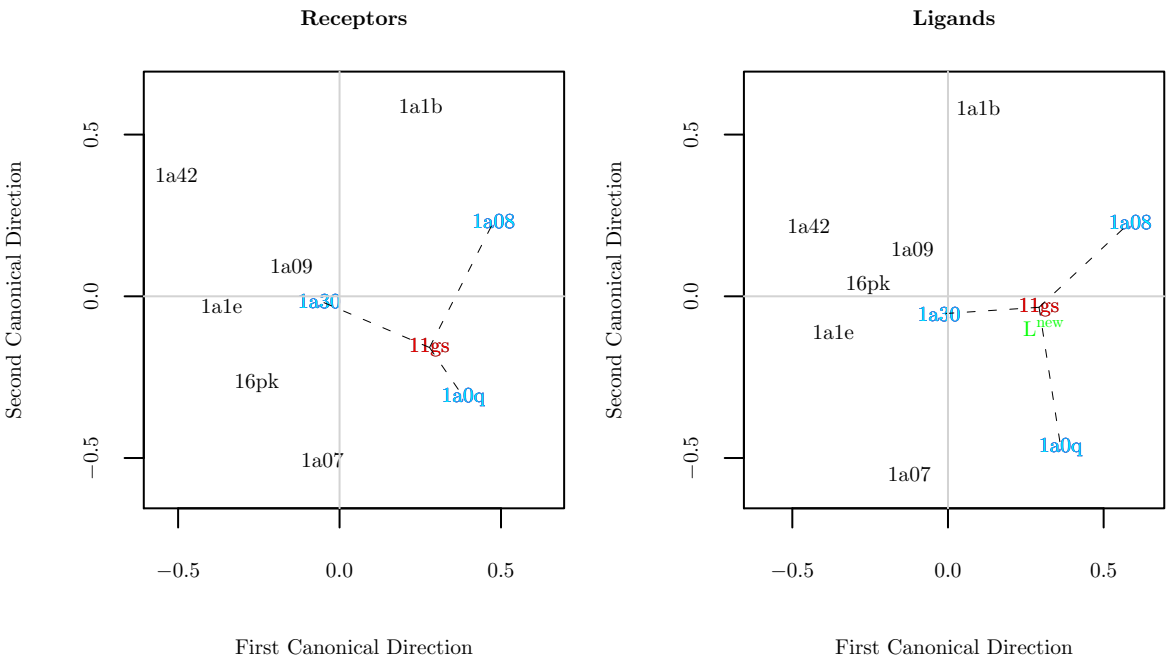

Figure 1.5: *Projection of the data in Figure 1.1 onto the first and second canonical vectors. In contrast to Figure 1.1 the point 11gs now shares the same neighbors in both spaces and the predicted value in green is much closer to the actual value.*

1.3.1 Ligand Prediction

Recall the example discussed in Section 1.2. In that example we first used CCA to define a mapping between the space of receptors and the space of ligands by projecting onto the first $p_X \leq d_X$ and $p_Y \leq d_Y$ directions (Figures 1.4 and 1.5). Let us define the projected values of the observations in X and Y space onto their first p_X and p_Y canonical vectors as

$$\mathbf{x}_{i,p}^w = \left(\mathbf{w}_X^1, \ldots, \mathbf{w}_X^{p_X}\right)^T, \mathbf{x}_i \in \mathbb{R}^{p_X}, i = 1, \ldots, n$$
$$\mathbf{y}_{i,p}^w = \left(\mathbf{w}_Y^1, \ldots, \mathbf{w}_Y^{p_Y}\right)^T, \mathbf{y}_i \in \mathbb{R}^{p_Y}, i = 1, \ldots, n$$

The sample of pairs are collected in matrices $\mathbf{X}_p^w \in \mathbb{R}^{n \times p_X}$ and $\mathbf{Y}_p^w \in \mathbb{R}^{n \times p_Y}$ with $\mathbf{x}_{i,p}^w$ and $\mathbf{y}_{i,p}^w$ as the observations for a row.

The method of prediction and assessment of model performance used, in the context of the protein-ligand matching problem, has similarities and differences with more traditional statistical definitions of these concepts. Prediction in this problem is similar to traditional definitions in the following sense: given a new input, $\mathbf{x}_{new}$, and it's projection onto the first p_X canonical vectors in X space (call this projected value $\mathbf{x}^w_{new,p}$), we want to predict the value of its unobserved pair, $\mathbf{y}^w_{new,p}$, in canonical correlation space.

There is an important distinction to draw here. Traditional methods of prediction usually assume a direction of dependence between the variables to be predicted (the *dependent variables*) and the variables predicting them (the *independent variables*), e.g. as in regression. Here we are more interested in a symmetric, not causal, type of relationship. This type of approach can be justified in the context of our, and similar problems for the following reasons: In our problem the binding between a protein and its ligand is inherently co-dependent. In similar problems, such as in information retrieval, the relationship between the input object, say a document in English, and the output object, the corresponding Japanese translation (Li and Shawe-Taylor (2006)) does not inherently imply a dependence one way or the other. Rather what we are interested in are the attributes that are held in common between them. There are also many examples in the field of bioinformatics where it is of interest to understand how multiple sources of information, for example gene expression and the corresponding metabolic pathway along which these genes fall (Vert and Kanehisa (2002)), co-depend on one another.

The accuracy of our prediction is assessed here in terms of how close, in Euclidean distance, our prediction, $\hat{\mathbf{y}}^w_{new,p}$ is to the actual value, $\mathbf{y}^w_{new,p}$. This is then compared to the set of the distances from each observation, $\mathbf{y}^w_{i,p}$, $i = 1, \ldots, n$ to the actual value. Predictive accuracy is measured by ranking these distances, from smallest to largest. Defining r_i to be the rank of our prediction of test ligand i, model performance is defined

as the average rank (over ligands) of our predictions,

$$\bar{r} = \frac{1}{n_T} \sum_{i=1}^{n_T} r_i, \tag{1.3}$$

where n_T is the number of test ligands.

The predicted value of $\mathbf{y}^w_{new,p}$ is calculated as follows (note that this is a modification of the LLE algorithm developed by Saul and Roweis (2003));

1. Compute the k neighbors of the data point $\mathbf{x}^w_{new,p}$ (the projected value of $\mathbf{x}_{new}$ into canonical correlation space). Define $N_k(\mathbf{x})$ to be the k nearest neighbors of the point $\mathbf{x}$.

2. Compute weights $\beta_{new,j}$ that best reconstruct the data point $\mathbf{x}^w_{new,p}$ from its neighbors, minimizing the cost:

$$\begin{aligned} L(\beta_{new}) &= \left(\mathbf{x}^w_{new,p} - \sum_{j:\mathbf{x}_j \in N_k(\mathbf{x}^w_{new,p})} \beta_{new,j} \mathbf{x}^w_{j,p} \right)^2, \\ \text{subject to } & \sum_{j:\mathbf{x}_j \in N_k(\mathbf{x}^w_{new,p})} \beta_{new,j} = 1. \end{aligned} \tag{1.4}$$

3. The new observation is then calculated as,

$$\hat{\mathbf{y}}^w_{new,p} = \sum_{j:\mathbf{x}_j \in N_k(\mathbf{x}^w_{new,p})} \beta_{new,j} \mathbf{y}^w_{j,p}. \tag{1.5}$$

Recall that CCA finds directions which best align two spaces. Thus, assuming that directions $\mathbf{w}^i_X$ and $\mathbf{w}^i_Y$, $i = 1, \ldots, p$, have been found such that the correlation between spaces is strong, using the weights $\beta_{new,j}$ found in X space should provide a reliable estimate of $\mathbf{y}^w_{new,p}$.

The results of our methods will be compared against those presented in Oloff *et al.* (2006). In their paper the RLP800 data was separated into 637 training points, used to

build the model, and 163 testing points, used to validate the predictive accuracy of the model. Predictive accuracy is measured by the ranking scheme described above.

To further test the predictive accuracy of our model (again following Oloff *et al.* (2006)) the WDI database is combined with the ligands from the RLP800 data set. The same ranking process is repeated but the set of ligands has been expanded to include both the WDI and RLP800 datasets.

1.3.2 Principal Component Analysis and Visualization

A parallel, but simpler tool which will prove useful in developing intuition about CCA and its extensions is principal component analysis (PCA) (Muirhead (1982)). PCA is a method used for analyzing and visualizing data. In contrast to our discussion thus far PCA looks to find linear combinations of the descriptors in an individual space, either in the space of proteins $\mathbf{X}$ or ligands $\mathbf{Y}$, which maximizes the variance (1.6). For convenience we focus on the space $\mathbf{X}$ as the same concepts hold for $\mathbf{Y}$. This variance maximization aspect of PCA can be formulated as

$$\begin{aligned} \gamma_X &= \max_{\mathbf{v}_X} \operatorname{var}(\mathbf{X}\mathbf{v}_X), \\ \text{subject to,} & \\ \mathbf{v}_X^T \mathbf{v}_X &= 1. \end{aligned} \tag{1.6}$$

The solution to (1.6) is found by defining λ_X to be the Lagrange multiplier, which gives us the corresponding Lagrangian,

$$L(\mathbf{v}_X, \lambda_X) = \mathbf{v}_X^T \Sigma_{XX} \mathbf{v}_X - \frac{\lambda_X}{2}(\mathbf{v}_X^T \mathbf{v}_X - 1). \tag{1.7}$$

Taking the derivative with respect to $\mathbf{v}_X$ and setting equal to zero gives us

$$\frac{\partial L(\mathbf{v}_X, \lambda_X)}{\partial \mathbf{v}_X} = \Sigma_{XX}\mathbf{v}_X - \lambda_X \mathbf{v}_X = 0. \tag{1.8}$$

Multiplying the left hand side of (1.8) by $\mathbf{v}_X^T$ yields

$$\mathbf{v}_X^T \Sigma_{XX} \mathbf{v}_X = \text{var}(X\mathbf{v}_X) = \lambda_X.$$

Thus $\lambda_X = \gamma_X$. Finally, rearranging terms in (1.8) gives us the eigen problem

$$\Sigma_{XX}\mathbf{v}_X = \gamma_X \mathbf{v}_X. \tag{1.9}$$

A new direction, $\mathbf{v}_X^*$ is found by repeating the process just described with the additional constraint that it be uncorrelated with $\mathbf{v}_X$. The problem in (1.6) is thus modified to be,

$$\gamma_X^* = \arg\max_{\mathbf{v}_X^*} \text{var}(X\mathbf{v}_X^*),$$

subject to,

$$\begin{aligned} &(\mathbf{v}_X^*)^T\mathbf{v}_X^* = 1 \\ &(\mathbf{v}_X^*)^T\mathbf{v}_X = 0, \\ &\text{cov}(X\mathbf{v}_X^*, X\mathbf{v}_X) = 0. \end{aligned} \tag{1.10}$$

Using Lagrange multipliers λ_X^* and μ_X gives the Lagrangian,

$$L(\mathbf{v}_X^*, \lambda_X^*, \mu_X) = (\mathbf{v}_X^*)^T X^T X \mathbf{v}_X^* - \frac{\lambda_X^*}{2}((\mathbf{v}_X^*)^T\mathbf{v}_X^* - 1) + \mu_X(\mathbf{v}_X^*)^T\mathbf{v}_X. \tag{1.11}$$

Taking the derivative of (1.11) with respect to $\mathbf{v}_X^*$ and setting equal to zero we have,

$$\frac{\partial L(\mathbf{v}_X^*, \lambda_X^*, \mu_X)}{\partial \mathbf{v}_X^*} = X^T X \mathbf{v}_X^* - \lambda_X^* \mathbf{v}_X^* + \mu_X \mathbf{v}_X = 0 \tag{1.12}$$

Multiplying the left hand side of (1.12) by $\mathbf{v}_X^T$ gives us,

$$\mathbf{v}_X^T X^T X \mathbf{v}_X^* - \lambda_X^* \mathbf{v}_X^T \mathbf{v}_X^* + \mu_X \mathbf{v}_X^T \mathbf{v}_X = \mu_X,$$

which implies that $\mu_X = 0$. Thus it can be seen that the eigenvalue γ_X^* and direction $\mathbf{v}_X^*$ are the second eigenvalue and eigenvector of Σ_{XX}. Additional linear combinations of X which maximize the variance are found in a similar fashion with the constraints in (1.10) being modified to include all previous directions.

A useful characteristic of PCA is that it allows us to visualize and gain insight into how the data are distributed. This is especially useful when the data is in a higher dimensional space. In Figures 1.6 and 1.7 we have plotted a *scatterplot matrix* showing the joint structure in the first four principal components as well as the eigenvalues for both proteins and ligands, respectively in the RLP800 training data set.

Figures 1.6 and 1.7 provide some insight into the distribution of the RLP800 data. Consider the plots of the eigenvalues in the lower left hand corner of each figure; what immediately stands out is the relatively small number of eigenvalues needed to explain a large proportion of the variation in both protein and ligand space. Here the proportion of variation measured by principal component i is given as

$$\frac{\text{var}(\mathbf{X}\mathbf{v}_X^i)}{\sum_j \text{var}(\mathbf{X}\mathbf{v}_X^j)} = \frac{\gamma_X^i}{\sum_j \gamma_X^j}. \tag{1.13}$$

This type of behavior can occur for a number of reasons, two of the more common ones are scaling and strong correlation between descriptors (also known as multicollinearity). Scaling can be an issue when the distribution of a descriptor has multiple modes, is skewed and/or is (nearly) discrete. This may have the effect of biasing the principal component vectors in the direction of these variables. In the presence of multicollinearity the PC directions will be dominated by a few larger modes of variation with the remaining ones being comparatively small. Multicollinearity will be discussed in more detail in Section

2.3.

While PCA is useful for studying a single space we are interested in studying how two spaces are related to one another. CCA is just such a tool.

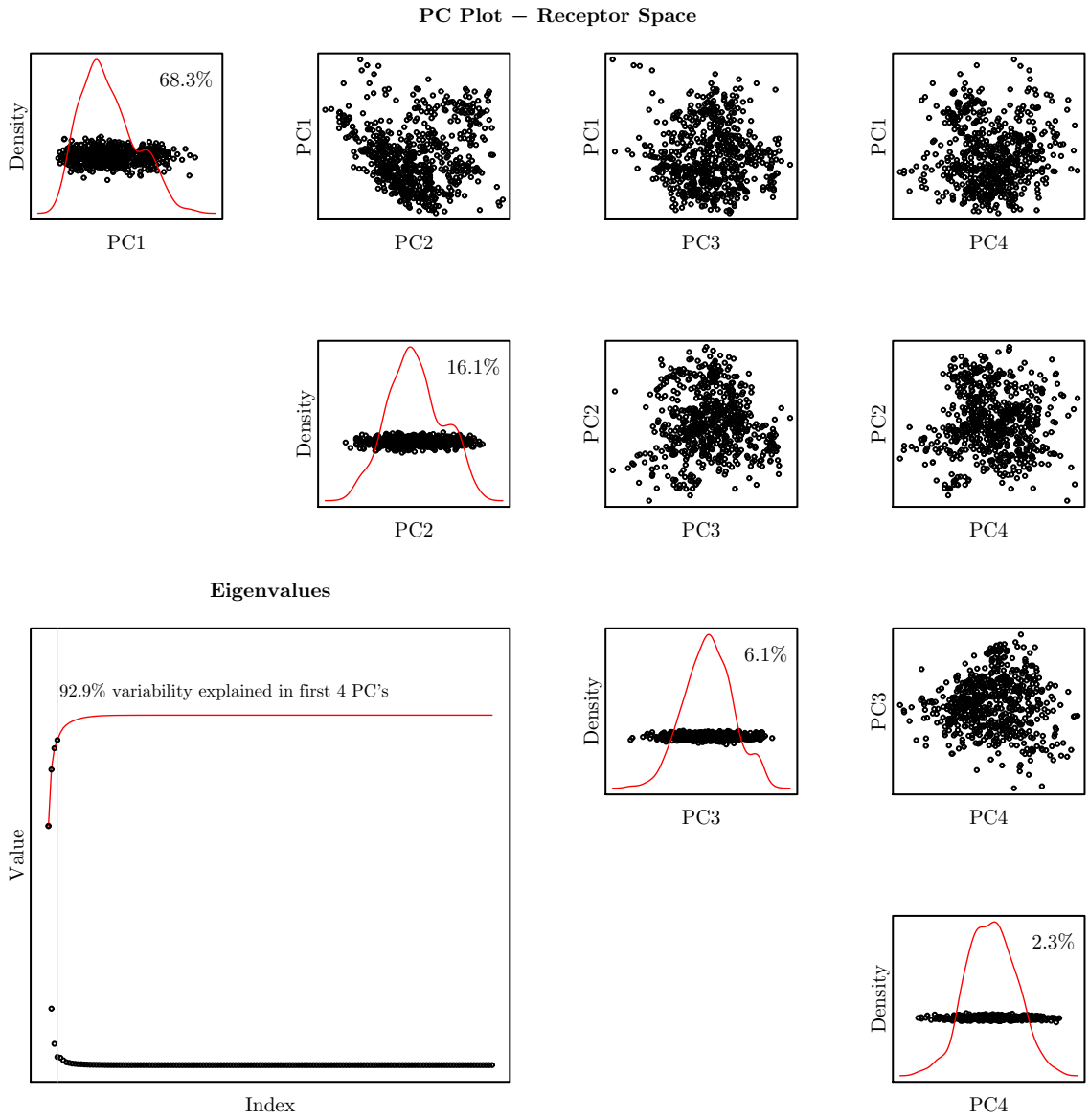

Figure 1.6: *The plots in the upper right half of the figure are the projections of the RLP800 receptor training data onto their first four principal components. The plots along the diagonal show the distribution of the projected values with the red curve being a kernel density estimate of the projections and the percentage in the upper right hand corner the proportion of variation explained by that principal component. The plot on the lower left side show the eigenvalues of all 150 principal components. The red curve is the cumulative sum of the eigenvalues.*

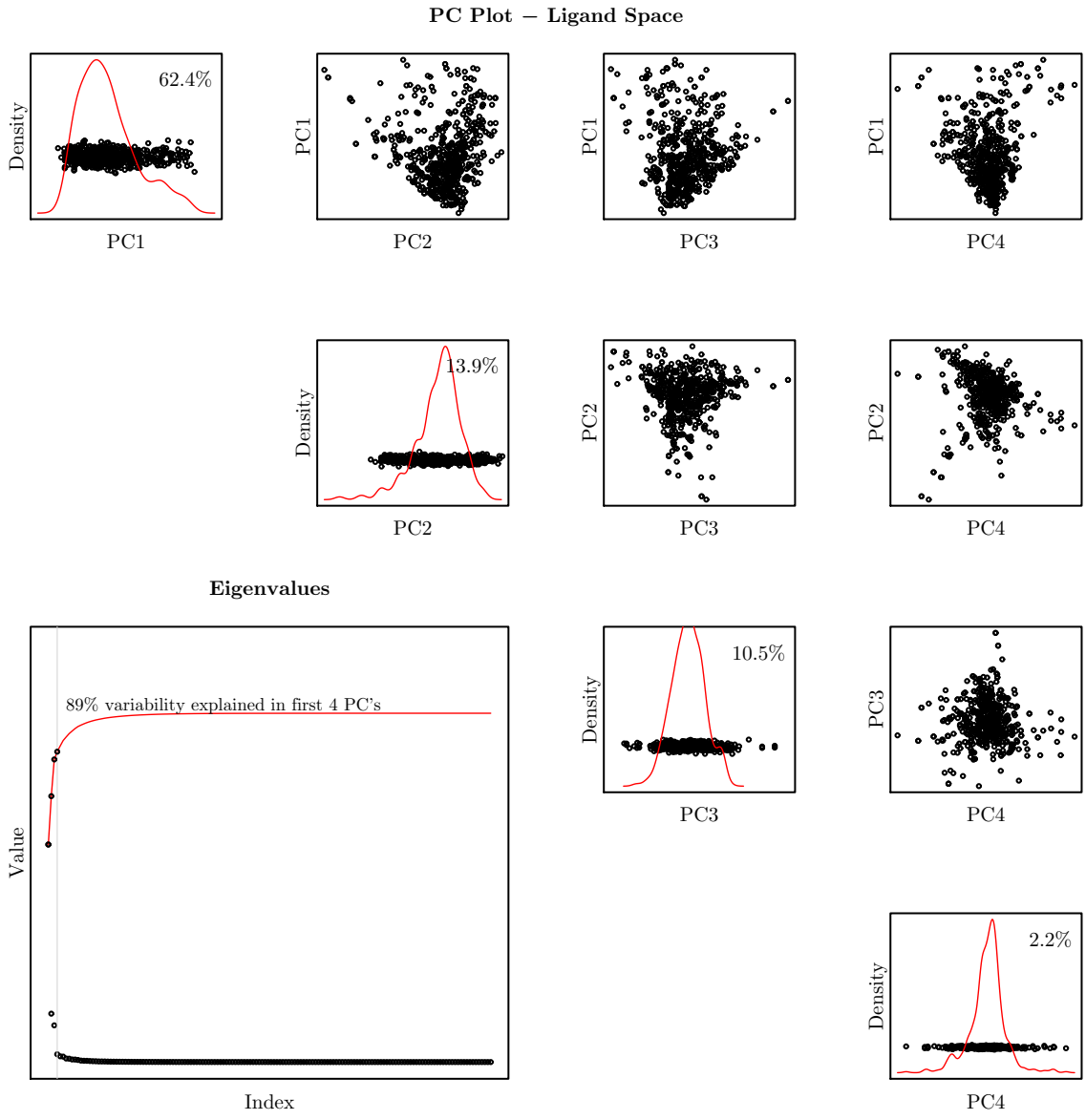

Figure 1.7: *Same layout as in Figure 1.6. but for the RLP800 ligand training data.*

CHAPTER 2

A mapping between spaces: Canonical Correlation Analysis

CCA was first proposed in 1936 (Hotelling (1936)). Since then it has seen application in a multitude of fields. In the prediction of protein-ligand binding the complexity of the data necessitates a way to model the relationship between them. CCA provides a natural framework for this type of analysis.

2.1 Linear Case

Consider the framework laid out in Section 1.1. Let $\mathbf{\Sigma}_{XX} = \text{cov}(X, X)$, $\mathbf{\Sigma}_{YY} = \text{cov}(Y, Y)$ and $\mathbf{\Sigma}_{XY} = \text{cov}(X, Y)$ denote the population covariances and $\mathbf{S_{XX}} = \widehat{\text{cov}}(\mathbf{X}, \mathbf{X})$, $\mathbf{S_{YY}} = \widehat{\text{cov}}(\mathbf{Y}, \mathbf{Y})$ and $\mathbf{S_{XY}} = \widehat{\text{cov}}(\mathbf{X}, \mathbf{Y})$ the sample covariances.

Since correlation is scale invariant we can make an arbitrary normalization of $\mathbf{w}_X$ and $\mathbf{w}_Y$. With this in mind we have the constraint

$$\text{cov}(\langle X, \mathbf{w}_X \rangle, \langle X, \mathbf{w}_X \rangle) = \text{cov}(\langle Y, \mathbf{w}_Y \rangle, \langle Y, \mathbf{w}_Y \rangle) = 1 \tag{2.1}$$

Using this constraint the optimization problem in (1.2) can be written as

$$\begin{aligned} \rho_{\mathcal{H}} &= \max_{\mathbf{w}_X, \mathbf{w}_Y} \text{corr}(\langle X, \mathbf{w}_X \rangle, \langle Y, \mathbf{w}_Y \rangle) = \mathbf{w}_X^T \Sigma_{XY} \mathbf{w}_Y, \\ &\text{subject to} \\ &\mathbf{w}_X^T \Sigma_{XX} \mathbf{w}_X = \mathbf{w}_Y^T \Sigma_{YY} \mathbf{w}_Y = 1. \end{aligned} \tag{2.2}$$

Using Lagrange multipliers ρ_X and ρ_Y the corresponding Lagrangian is

$$L(\mathbf{w}_X, \mathbf{w}_Y, \rho_X, \rho_Y) = \mathbf{w}_X^T \Sigma_{XY} \mathbf{w}_Y - \frac{\rho_X}{2}(\mathbf{w}_X^T \Sigma_{XX} \mathbf{w}_X - 1) - \frac{\rho_Y}{2}(\mathbf{w}_Y^T \Sigma_{YY} \mathbf{w}_Y - 1). \quad (2.3)$$

Taking the derivative of (2.3) with respect to $\mathbf{w}_X$ and $\mathbf{w}_Y$ and setting equal to zero we have

$$\frac{\partial L(\mathbf{w}_X, \mathbf{w}_Y, \rho_X, \rho_Y)}{\partial \mathbf{w}_X} = \Sigma_{XY} \mathbf{w}_Y - \rho_X \Sigma_{XX} \mathbf{w}_X = 0, \quad (2.4)$$

$$\frac{\partial L(\mathbf{w}_X, \mathbf{w}_Y, \rho_X, \rho_Y)}{\partial \mathbf{w}_Y} = \Sigma_{YX} \mathbf{w}_X - \rho_Y \Sigma_{YY} \mathbf{w}_Y = 0. \quad (2.5)$$

Multiplying the left hand sides of Equations (2.4) and (2.5) by, respectively, $\mathbf{w}_X^T$ and $\mathbf{w}_Y^T$ and then subtracting the resulting equations from each other gives us

$$\begin{aligned} &\mathbf{w}_X^T \Sigma_{XY} \mathbf{w}_Y - \rho_X \mathbf{w}_X^T \Sigma_{XX} \mathbf{w}_X - \mathbf{w}_Y^T \Sigma_{YX} \mathbf{w}_X + \rho_Y \mathbf{w}_Y^T \Sigma_{YY} \mathbf{w}_Y \\ &= \rho_Y \mathbf{w}_Y^T \Sigma_{YY} \mathbf{w}_Y - \rho_X \mathbf{w}_X^T \Sigma_{XX} \mathbf{w}_X = 0, \end{aligned}$$

from which it follows that

$$\rho_X = \rho_Y = \mathrm{corr}(\langle X, \mathbf{w}_X \rangle, \langle Y, \mathbf{w}_Y \rangle) = \rho_{\mathcal{H}}.$$

Assuming Σ_{YY} is invertible we have

$$\mathbf{w}_Y = \frac{\Sigma_{YY}^{-1} \Sigma_{YX} \mathbf{w}_X}{\rho_{\mathcal{H}}}. \quad (2.6)$$

Similarly we have,

$$\mathbf{w}_X = \frac{\Sigma_{XX}^{-1} \Sigma_{XY} \mathbf{w}_Y}{\rho_{\mathcal{H}}}. \quad (2.7)$$

Substituting (2.6) into (2.4) and rearranging terms gives the generalized eigenvalue problem,

$$\Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \mathbf{w}_X = \rho_{\mathcal{H}}^2 \Sigma_{XX} \mathbf{w}_X, \quad (2.8)$$

similar calculations lead to

$$\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{YX}\mathbf{w}_Y = \rho_{\mathcal{H}}^2\Sigma_{YY}\mathbf{w}_Y. \tag{2.9}$$

Equivalently using Equations (2.4) and (2.5) the generalized eigenvalue problem can be rewritten as,

$$\begin{pmatrix} \mathbf{0} & \Sigma_{XY} \\ \Sigma_{YX} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_X \\ \mathbf{w}_Y \end{pmatrix} = \rho_{\mathcal{H}} \begin{pmatrix} \Sigma_{XX} & \mathbf{0} \\ \mathbf{0} & \Sigma_{YY} \end{pmatrix} \begin{pmatrix} \mathbf{w}_X \\ \mathbf{w}_Y \end{pmatrix}. \tag{2.10}$$

We now discuss how to find second and subsequent linear combinations of X and Y. The objective is to find maximally correlated linear combinations of X, say $X\mathbf{w}_X^*$ and Y, say $Y\mathbf{w}_Y^*$ which are uncorrelated with $X\mathbf{w}_X$ and $Y\mathbf{w}_Y$ from (2.2). The optimization problem thus written as,

$$\begin{aligned} &\rho_{\mathcal{H}} = \max_{\mathbf{w}_X^*, \mathbf{w}_Y^*} \text{corr}(\langle X, \mathbf{w}_X^* \rangle, \langle Y, \mathbf{w}_Y^* \rangle) = (\mathbf{w}_X^*)^T \Sigma_{XY} \mathbf{w}_Y^*, \\ &\text{subject to} \\ &(\mathbf{w}_X^*)^T \Sigma_{XX} \mathbf{w}_X^* = (\mathbf{w}_Y^*)^T \Sigma_{YY} \mathbf{w}_Y^* = 1 \\ &\mathbf{w}_X^T \Sigma_{XX} \mathbf{w}_X^* = \mathbf{w}_Y^T \Sigma_{YY} \mathbf{w}_Y^* = 0 \\ &\mathbf{w}_X^T \Sigma_{XY} \mathbf{w}_Y^* = \mathbf{w}_Y^T \Sigma_{YX} \mathbf{w}_X^* = 0. \end{aligned} \tag{2.11}$$

Using Lagrange multipliers ρ_X^*, ρ_Y^*, μ_X and μ_Y gives the Lagrangian,

$$\begin{aligned} L(\mathbf{w}_X^*, \mathbf{w}_Y^*, \rho_X^*, \rho_Y^*, \mu_X, \mu_Y) &= (\mathbf{w}_X^*)^T \Sigma_{XY} \mathbf{w}_Y^* - \frac{\rho_X^*}{2}((\mathbf{w}_X^*)^T \Sigma_{XX} \mathbf{w}_X^* - 1) \\ &- \frac{\rho_Y^*}{2}((\mathbf{w}_Y^*)^T \Sigma_{YY} \mathbf{w}_Y^* - 1) + \mu_X \mathbf{w}_X^T \Sigma_{XX} \mathbf{w}_X^* + \mu_Y \mathbf{w}_Y^T \Sigma_{YY} \mathbf{w}_Y^*. \end{aligned} \tag{2.12}$$

Taking the derivative of (2.12) with respect $\mathbf{w}_X^*$ and $\mathbf{w}_Y^*$ and setting equal to zero we

have,

$$\frac{\partial L(\mathbf{w}_X^*, \mathbf{w}_Y^*, \rho_X^*, \rho_Y^*, \mu_X, \mu_Y)}{\partial \mathbf{w}_X^*} = \Sigma_{XY}\mathbf{w}_Y^* - \rho_X^*\Sigma_{XX}\mathbf{w}_X^* + \mu_X\Sigma_{XX}\mathbf{w}_X = \mathbf{0}, \tag{2.13}$$

$$\frac{\partial L(\mathbf{w}_X^*, \mathbf{w}_Y^*, \rho_X^*, \rho_Y^*, \mu_X, \mu_Y)}{\partial \mathbf{w}_Y^*} = \Sigma_{YX}\mathbf{w}_X^* - \rho_Y^*\Sigma_{YY}\mathbf{w}_Y^* + \mu_Y\Sigma_{YY}\mathbf{w}_Y = \mathbf{0}. \tag{2.14}$$

Multiplying the left hand side of (2.13) and (2.14) by $\mathbf{w}_X^T$ and $\mathbf{w}_Y^T$ respectively gives us

$$\begin{aligned} 0 &= \mu_X \mathbf{w}_X^T \Sigma_{XX} \mathbf{w}_X = \mu_X, \\ 0 &= \mu_Y \mathbf{w}_Y^T \Sigma_{YY} \mathbf{w}_Y = \mu_Y. \end{aligned}$$

With $\mu_X = \mu_Y = 0$ it can be seen that the canonical vectors $\mathbf{w}_X^*$ and $\mathbf{w}_Y^*$ are the second set of eigenvectors from the generalized eigenvalue problem in (2.8) and (2.9). The extension to additional linear combinations of X and Y follows along the same lines as just described with the constraints in (2.11) being modified to include orthogonality to all previous linear combinations of X and Y.

Remark 2.1.1. Eigen analyses have *ambiguous polarity* in the sense that they are only determined up to a factor of ± 1. This ambiguous polarity is resolved in a way that gives comparable results across similar data analyses by employing the following convention: The main idea is that the directions the eigenvectors follow, in each space, will always place the largest, in absolute value, projected value across both spaces on the positive (right hand) side of the axis. In other words let $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{w}_X$ and $\mathbf{w}_Y$ be as defined previously. Define the scores $a_X^i = \mathbf{x}_i^T \mathbf{w}_X$ and $a_Y^i = \mathbf{y}_i^T \mathbf{w}_Y$ to be the projected values of the i^{th} observation onto its canonical vector. Let $a_X^{(n)}$ and $a_Y^{(n)}$ be the largest score, in absolute value, across all a_X^k and a_Y^k, $k = 1, \ldots, n$, respectively. Then,

$$\begin{aligned} a_X^i &= \text{sign}(\max\{a_X^{(n)}, a_Y^{(n)}\}) \cdot a_X^i, \\ a_Y^i &= \text{sign}(\max\{a_X^{(n)}, a_Y^{(n)}\}) \cdot a_Y^i. \end{aligned} \tag{2.15}$$

In the event of a tie in the scores $a_X^{(n)}$ and $a_Y^{(n)}$ the sign is taken to be $+1$. This transformation of the data does not change the relationship of the projections between spaces. This can be seen by noting that the values of the signs by which we are multiplying the projections in both spaces will always be the same. Thus the correlation between the projections will remain unchanged.

An example of linear CCA was presented in Section 1.2. In the following section we present several toy examples illustrating where linear CCA performs well and also where it does not perform well.

2.2 Properties of CCA

CCA is invariant with respect to several common linear transformations. This point is illustrated in Figure 2.1. Plots (b), (c) and (d) in Figure 2.1 depict different transformations, orthonormal, scale and location, respectively of the data shown in Figure 2.1 (a) (not shown in the plots are the canonical correlations, 1 and 0.996 which are the same for all four groups of plots). These properties are straightforward to verify. Let X, Y, $\mathbf{w}_X$ and $\mathbf{w}_Y$ be defined as above. To ease calculation we also assume that X and Y have mean zero.

1. Orthonormal: Define $\mathbf{Q}_X \in \mathbb{R}^{d_X \times d_X}$ and $\mathbf{Q}_Y \in \mathbb{R}^{d_Y \times d_Y}$ to be orthonormal matrices. I.e.:

   (a) $\mathbf{Q}_X^T \mathbf{Q}_X = \mathbf{Q}_X \mathbf{Q}_X^T = \mathbf{I}_{d_X}$,

   (b) $\mathbf{Q}_Y^T \mathbf{Q}_Y = \mathbf{Q}_Y \mathbf{Q}_Y^T = \mathbf{I}_{d_Y}$

   Define the orthonormal transformations $X^* = X\mathbf{Q}_X$ and $Y^* = Y\mathbf{Q}_Y$. Define $\mathrm{E}[\cdot]$ to be the expected value. Using the result found in (2.8) we have

$$\mathrm{E}[(X^*)^T Y^*](\mathrm{E}[(Y^*)^T Y^*])^{-1}\mathrm{E}[(Y^*)^T X^*]\mathbf{w}_X^* = (\rho^*)^2 \mathrm{E}[(X^*)^T X^*]\mathbf{w}_X^*. \tag{2.16}$$

Substituting in for X^* and Y^* gives us,

$$\mathbf{Q}_X^T \mathrm{E}[X^T Y]\mathbf{Q}_Y(\mathbf{Q}_Y^T \mathrm{E}[Y^T Y]\mathbf{Q}_Y)^{-1}\mathbf{Q}_Y^T \mathrm{E}[Y^T X]\mathbf{Q}_X \mathbf{w}_X^* = (\rho^*)^2 \mathbf{Q}_X^T \mathrm{E}[X^T X]\mathbf{Q}_X \mathbf{w}_X^*. \tag{2.17}$$

Next we use properties (a) and (b) defined above. Multiplying the left hand side of the previous equation by $\mathbf{Q}_X$ and setting $\mathbf{w}'_X = \mathbf{Q}_X \mathbf{w}_X^*$ gives us,

$$\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{w}'_X = (\rho^*)^2 \Sigma_{XX}\mathbf{w}'_X. \tag{2.18}$$

From this it can be seen that the resulting generalized eigenvalue and eigenvector from (2.18) will be the same as those found in (2.8).

Figure 2.1(b) illustrates CCA's invariance to orthonormal transformations. In the X space the data has been rotated 30° counterclockwise and in the Y space the data has been rotated 75° clockwise (these rotations satisfy the properties (a) and (b)). The resulting projected values remain unchanged as do the canonical correlation values.

2. Scale: We use the results from CCA's invariance to orthonormal transformations to show its scale invariance. This follows immediately by substituting in scalars a and b for the orthonormal matrices $\mathbf{Q}_X$ and $\mathbf{Q}_Y$ in (2.17) which then leads to the same result as in (2.18).

   An illustration of CCA's scale invariance is presented in Figure 2.1 (c). The projected values and canonical correlations are identical to those in Figure 2.1 (a).

3. Translation: Define $\mathbf{c}_x \in \mathbb{R}_X^d$ and $\mathbf{c}_y \in \mathbb{R}_Y^d$ to be vectors of constants, and $\mathbf{1}_n \in \mathbb{R}^n$ to be a vector of ones then

$$\begin{aligned}
&\mathrm{corr}(\langle X + \mathbf{1}\mathbf{c}_x^T, \mathbf{w}_X\rangle, \langle Y + \mathbf{1}\mathbf{c}_y^T, \mathbf{w}_X\rangle) \\
&= \mathbf{w}_X^T \mathrm{E}[(X + \mathbf{1}\mathbf{c}_x^T - \mathrm{E}[X + \mathbf{1}\mathbf{c}_x^T])^T (Y + \mathbf{1}\mathbf{c}_y^T - \mathrm{E}[Y + \mathbf{1}\mathbf{c}_y^T])]\mathbf{w}_Y
\end{aligned}$$

$$= \mathbf{w}_X^T \mathrm{E}[X^T Y]\mathbf{w}_Y$$
$$= \mathrm{corr}(\langle X, \mathbf{w}_X \rangle, \langle Y, \mathbf{w}_Y \rangle).$$

Figure 2.1(d) provides an illustration of CCA's invariance to translation. Looking at the projected values and canonical correlations they are identical to those found in (a).

2.3 Regularized Canonical Correlation Analysis

There are many cases, particularly in biological problems where the data being analyzed have a large number of covariates (descriptors) as compared to the number of observations. This can lead to situations where there are potentially many highly correlated covariates, this type of behavior is referred to as multicollinearity. An approach to control the effects of multicollinearity is to add a penalty term which controls the variability of the eigenvectors of the sample covariance matrices within the X and Y spaces. There is a close relationship between variability in the eigenvectors and multicollinearity (which we discuss in greater detail below).

Recall that the eigenvalues and eigenvectors found from the eigen decomposition of the sample covariance matrix $\mathbf{S}_{XX}$ (and $\mathbf{S}_{YY}$) are also the solution to the PCA optimization problem discussed in Section 1.3.2.

An important aspect of PCA is that it gives us insight into the structure of the data. In particular it can alert us to the presence of multicollinearity. Consider a set of n observations each of which has d variables (descriptors). If there exists strong multicollinearity amongst the variables then a subset $p(< d)$ of the eigenvalues will have relatively large values and the remaining $d - p$ eigenvalues will have comparatively small values. This type of behavior in the eigenvalues can create numeric instabilities in the sample covariance matrices. Recall from equations (2.6) and (2.7) that the canonical

vectors $\mathbf{w}_X$ and $\mathbf{w}_Y$ can be written as,

$$\mathbf{w}_X = \frac{\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\mathbf{w}_Y}{\rho_{\mathcal{H}}},$$
$$\mathbf{w}_Y = \frac{\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\mathbf{w}_X}{\rho_{\mathcal{H}}}.$$

The effect of this instability on the canonical vectors can be seen by noting that their solutions depend on the inverse of the covariance matrices $\mathbf{S}_{XX}$ and $\mathbf{S}_{YY}$. An immediate consequence of this is that when the sample eigenvalues act as just mentioned it can be seen from (2.19) that small eigenvalues (near zero) will tend to inflate the elements of these matrices,

$$\begin{aligned}\mathbf{S}_{XX}^{-1} &= \mathbf{V}_X\mathbf{D}_X^{-1}\mathbf{V}_X^T,\\ \mathbf{S}_{YY}^{-1} &= \mathbf{V}_Y\mathbf{D}_Y^{-1}\mathbf{V}_Y^T.\end{aligned} \tag{2.19}$$

Here $\mathbf{V}_X = (\mathbf{v}_X^1, \ldots, \mathbf{v}_X^{d_X})$ and $\mathbf{V}_Y = (\mathbf{v}_Y^1, \ldots, \mathbf{v}_Y^{d_Y})$ are the matrices of eigenvectors (i.e. principal component direction vectors) of the sample covariance matrices $\mathbf{S}_{XX}$ and $\mathbf{S}_{YY}$. The matrices $\mathbf{D}_X^{-1}$ and $\mathbf{D}_Y^{-1}$ have elements $\frac{1}{\gamma_X^i}$, $i = 1, \ldots, d_X$ and $\frac{1}{\gamma_Y^i}$, $i = 1, \ldots, d_Y$ along their diagonals, where γ_X^i and γ_Y^i are the eigenvalues of their respective covariance matrices.

The large sample to sample variation in the canonical vectors can be understood by noting that even slight perturbations in the smallest eigenvalues of the sample covariance can lead to drastically different results in the inverse of the covariance matrices and therefore in the canonical vectors. The affect of this instability is illustrated in Figure 2.2 which shows an example of canonical vectors in X space. The data has been generated such that the first and third variables are strongly correlated. The black lines show the first canonical direction vector found from ten random samplings from this distribution. The red dashed line is the theoretical direction derived from the true variance and covariance matrices. As can be seen there is a large amount of variation from sample to

sample and a significant deviation from the theoretical direction.

The impact of this variation is felt the strongest when projecting new data onto one of these directions. For example, suppose new observations are generated from a distribution similar to that just described with the difference lying in a slight perturbation of the covariance matrix in the X space. These new observations are then projected onto the canonical vectors shown in Figure 2.2.

Ideally the projected values of the new data would vary only slightly from one set of directions to the next. Figure 2.3 shows a plot of each pair of projected values (the projection of the new data discussed above onto each of the directions shown in Figure 2.2) against one another. The observations within each of these plots should, if the directions were well behaved, fall on or near the 45° line (shown in red). However, due to the large amount of variation in the canonical vectors the resulting projections are highly variable.

One possible approach to dealing with this problem is to control how variable we allow the canonical direction vectors to be. One such penalty would be a modification of the constraint in (2.1) where an L_2 constraint on the L_2 length of the canonical vectors $\mathbf{w}_X$ and $\mathbf{w}_Y$ (Vinod (1976)) is added. This new constraint (2.20) now penalizes for how variable we allow the directions in any one space to be,

$$\text{cov}(\langle X, \mathbf{w}_X \rangle, \langle X, \mathbf{w}_X \rangle) + \kappa_X \langle \mathbf{w}_X, \mathbf{w}_X \rangle = \text{cov}(\langle Y, \mathbf{w}_Y \rangle, \langle Y, \mathbf{w}_Y \rangle) + \kappa_Y \langle \mathbf{w}_Y, \mathbf{w}_Y \rangle = 1. \tag{2.20}$$

Solving (2.2) but with new constraints (2.20) is done in a similar fashion to standard CCA. Using Lagrange multipliers ρ_X and ρ_Y we have the following modified Lagrangian as compared to (2.3),

$$\begin{aligned} L(\mathbf{w}_X, \mathbf{w}_Y, \rho_X, \rho_Y) &= \mathbf{w}_X^T \Sigma_{XY} \mathbf{w}_Y - \frac{\rho_X}{2}(\mathbf{w}_X^T \Sigma_{XX} \mathbf{w}_X + \kappa_X \mathbf{w}_X^T \mathbf{w}_X - 1) \\ &- \frac{\rho_Y}{2}(\mathbf{w}_Y^T \Sigma_{YY} \mathbf{w}_Y + \kappa_Y \mathbf{w}_Y^T \mathbf{w}_Y - 1). \end{aligned} \tag{2.21}$$

Taking the derivative with respect to $\mathbf{w}_X$ and $\mathbf{w}_Y$ and setting equal to zero gives

$$\frac{\partial L(\mathbf{w}_X, \mathbf{w}_Y, \rho_X, \rho_Y)}{\partial \mathbf{w}_X} = \Sigma_{XY}\mathbf{w}_Y - \rho_X(\Sigma_{XX}\mathbf{w}_X + \kappa_X\mathbf{w}_X) = 0, \tag{2.22}$$

$$\frac{\partial L(\mathbf{w}_X, \mathbf{w}_Y, \rho_X, \rho_Y)}{\partial \mathbf{w}_Y} = \Sigma_{YX}\mathbf{w}_X - \rho_Y(\Sigma_{YY}\mathbf{w}_Y + \kappa_Y\mathbf{w}_Y) = 0. \tag{2.23}$$

Multiplying the left hand sides of Equations (2.22) and (2.23) by, respectively, $\mathbf{w}_X^T$ and $\mathbf{w}_Y^T$ and then subtracting the resulting equations from each other gives us

$$\begin{aligned}&\mathbf{w}_X^T\Sigma_{XY}\mathbf{w}_Y - \rho_X(\mathbf{w}_X^T\Sigma_{XX}\mathbf{w}_X + \mathbf{w}_X^T\mathbf{w}_X) - \mathbf{w}_Y^T\Sigma_{YX}\mathbf{w}_X + \rho_Y(\mathbf{w}_Y^T\Sigma_{YY}\mathbf{w}_Y + \mathbf{w}_Y^T\mathbf{w}_Y)\\ &= \rho_Y(\mathbf{w}_Y^T\Sigma_{YY}\mathbf{w}_Y + \mathbf{w}_Y^T\mathbf{w}_Y) - \rho_X(\mathbf{w}_X^T\Sigma_{XX}\mathbf{w}_X + \mathbf{w}_X^T\mathbf{w}_X) = 0,\end{aligned}$$

from which it follows that

$$\rho_X = \rho_Y = \text{corr}(\langle X, \mathbf{w}_X\rangle, \langle Y, \mathbf{w}_Y\rangle) = \rho_{\mathcal{H}}.$$

Assuming $\Sigma_{YY} + \kappa_Y I_{d_Y}$ is invertible we have

$$\mathbf{w}_Y = \frac{(\Sigma_{YY} + \kappa_Y I_{d_Y})^{-1}\Sigma_{YX}\mathbf{w}_X}{\rho_{\mathcal{H}}}.$$

Substituting into (2.22) and rearranging terms gives the generalized eigenvalue problem,

$$\Sigma_{XY}(\Sigma_{YY} + \kappa_Y I_{d_Y})^{-1}\Sigma_{YX}\mathbf{w}_X = \rho_{\mathcal{H}}^2(\Sigma_{XX} + \kappa_X I_{d_X})\mathbf{w}_X, \tag{2.24}$$

similar calculations lead to

$$\Sigma_{YX}(\Sigma_{XX} + \kappa_X I_{d_X})^{-1}\Sigma_{YX}\mathbf{w}_Y = \rho_{\mathcal{H}}^2(\Sigma_{YY} + \kappa_Y I_{d_Y})\mathbf{w}_Y. \tag{2.25}$$

Equivalently using Equations (2.22) and (2.23) the generalized eigen problem can be

rewritten as,

$$\begin{pmatrix} \mathbf{0} & \Sigma_{XY} \\ \Sigma_{YX} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_X \\ \mathbf{w}_Y \end{pmatrix} = \rho_{\mathcal{H}} \begin{pmatrix} \Sigma_{XX} + \kappa_X I_{d_X} & \mathbf{0} \\ \mathbf{0} & \Sigma_{YY} + \kappa_Y I_{d_Y} \end{pmatrix} \begin{pmatrix} \mathbf{w}_X \\ \mathbf{w}_Y \end{pmatrix}. \quad (2.26)$$

Subsequent calculations to find new directions are similar to those discussed for the un-penalized case. In addition the same invariance properties that were discussed for standard CCA hold for this regularized variant of CCA (RCCA).

Consider again the example presented at the beginning of this section. Figure 2.4 shows a plot of the canonical direction vectors found from using RCCA with a value of 0.1 for the regularization parameter κ_X. In contrast to Figure 2.2, the canonical direction vectors are quite similar from one sample to the next. The dashed red line is once again the theoretical direction.

Figure 2.5 is the same plot as Figure 2.3 but with the new data being projected onto the direction vectors shown in Figure 2.4. As can be seen the projected values are quite similar from one set of directions to the next.

In the context of the protein-ligand matching problem consistent behavior of the canonical vectors is critical. Because the primary object of interest is the prediction of new protein-ligand pairs it is important that the directions that are found are not overly dependent on the training sample. As is illustrated in Figure 2.3 if measures are not taken to control the variability of the canonical vectors the projected values and any prediction based on them become unreliable.

2.4 A Toy Example

Linear CCA, in both its standard and regularized form, encounters greater challenges when the relationship between distributions of points is more complex, for example if some type of non-linearity is introduced. In the same framework as the example presented in Section 1.2 we consider a new toy data set shown in Figure 2.6, with a much more

complex relationship between proteins and ligands. Recall the task is the following: given a new observation in the space of proteins can we accurately predict the corresponding point in the space of ligands.

The data in the space of proteins falls into three distinct groups and the data in the space of ligands falls into two distinct groups. This scenario is relevant in the context of our example for the following reason: a single protein can bind many different ligands, based on the conformation, i.e. steric layout, of the binding site. Thus in the context of our example the three different clusters could be thought of as representing three different proteins. The slight perturbation in each group is attributed to the change in conformation of the binding sites of each protein to allow the binding of different ligands. The two groups in the space of ligands could be thought of as representing ligands corresponding to proteins, larger macro molecules or shorter sequences of peptides, small molecules.

The data has been generated such that those points which fall into the same group in both protein and ligand space are highly correlated. The result is that the global structure of the data is non-linear in the following sense: the underlying correlation structure of protein ligand pairs is localized, as a result this relationship cannot be captured by a simple (global) linear combination of the descriptors.

Observations in Figure 2.7 are highlighted according to whether they fall into the same cluster in both spaces. This plot helps illustrate just how different the neighborhood structures are in protein space versus ligand space. Consider, for example, the point labeled 1a7t (cyan). Its neighbors in protein space are all different from the corresponding point in ligand space.

In addition to the failure of the simple nearest neighbor method, as shown in Figure 2.6, linear CCA is also challenged by this new toy example. Since both CCA and RCCA essentially provide identical results in this example only the results from CCA are presented. In contrast to Figure 1.4 in Section 1.2 the distribution of points along the first

canonical directions, shown in red in the top row of plots in Figure 2.8, do not show a strong alignment of points between spaces. The same is true for the second canonical direction, shown in green in the bottom set of plots. The canonical correlation values, 0.46 and 0.34 confirm our visual assessment.

Looking at the projections onto the first two canonical vectors shown in Figure 2.9 we can see little if any change has been made to the structure of the data in protein space, relative to the raw data shown in Figure 2.6. In ligand space the directions found appear to have made the prediction of L^{new} worse.

In Chapter 3 a variant of CCA will be discussed which can capture this non-linear relationship between spaces.

2.5 Connection Between Linear Discriminant Analysis and CCA

A question of interest in many problems is the classification of a set of observations into one of several distinct categories. This is one example of *supervised learning*, see Duda *et al.* (2000) for an overview of the large literature on this topic. In contrast to supervised learning is clustering, a specific area of *unsupervised learning*. In clustering the categories are unknown and the task is to determine what "natural" groupings can be found in the data. Linear Discriminant Analysis (LDA) (Fisher (1936)) is a standard tool used in classification. In Section 2.5.1 we outline LDA and in Section 2.5.2 we show LDA in terms of CCA.

2.5.1 Linear Discriminant Analysis

Consider the k class (i.e. k category) discrimination problem. Suppose we have a set of n observation-label pairs, $(\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^d \times \{0,1\}^k$, $i = 1, \ldots, n$. Let C_j, $j = 1, \ldots, k$ be the collection of points $\mathbf{x}_i$ which belong to class j. To fit this problem into a similar context as the protein-ligand example of Section 1.2 we consider the following variation:

let the observations $\mathbf{x}_i$ be a collection of drug descriptors (i.e. ligands) and $\mathbf{y}_i$ be the labels representing whether a drug is active or inactive. Define $\mathbf{X} \in \mathbb{R}^{n \times d}$ to be a matrix whose rows are the observations $\mathbf{x}_i$. Define $\mathbf{Y} \in \mathbb{R}^{n \times k}$ to be the label matrix whose ij^{th} entry is defined as $y_{ij} = I_{\{\mathbf{x}_j \in C_i\}}$, where I is the indicator function. One way to think of LDA is that it looks to find a vector of weights, $\mathbf{w}_X$, associated with the columns of $\mathbf{X}$, such that the linear combination, $\mathbf{X}\mathbf{w}_X$ maximizes the ratio of its between-class variance to its within-class variance, defined in (2.30) and (2.29). To ease notation we assume that $\mathbf{X}$ has been mean centered.

Define $n_j = \sum_i y_{ij} = |C_j|$, where $|C_j|$ denotes the cardinality of C_j, to be the number of observations in class j and let $\mathbf{m}_j$ in (2.27) be defined as the mean of the observations that belong to class j,

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{i:\mathbf{x}_i \in C_j} \mathbf{x}_i. \tag{2.27}$$

Define the total sum of squares to be

$$\mathbf{S}_T = \sum_{i=1}^{k} \sum_{j:\mathbf{x}_j \in C_i} \mathbf{x}_j \mathbf{x}_j^T = (n-1)\mathbf{S}_{XX}, \tag{2.28}$$

where $\mathbf{S}_{XX}$ is the sample covariance matrix, discussed in Section 2.1. The total sum of squares, S_T can be decomposed into the sum of the *within-class sum of squares*,

$$\mathbf{S}_W = \sum_{i=1}^{k} \sum_{j:\mathbf{x}_j \in C_i} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T, \tag{2.29}$$

and *between-class sum of squares*

$$\mathbf{S}_B = \sum_{i=1}^{k} n_i \mathbf{m}_i \mathbf{m}_i^T. \tag{2.30}$$

Specifically,

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B. \tag{2.31}$$

With these definitions we can now state the LDA optimization problem,

$$\begin{aligned} \mathbf{w}_X^* &= \arg\max_{\mathbf{w}_X} \mathbf{w}_X^T \mathbf{S}_B \mathbf{w}_X, \\ \text{subject to,} & \\ \mathbf{w}_X^T \mathbf{S}_W \mathbf{w}_X &= 1. \end{aligned} \tag{2.32}$$

Using the Lagrange multiplier λ gives the corresponding Lagrangian

$$L(\mathbf{w}_X, \lambda) = \mathbf{w}_X^T \mathbf{S}_B \mathbf{w}_X - \lambda(\mathbf{w}_X^T \mathbf{S}_W \mathbf{w}_X - 1). \tag{2.33}$$

Taking the derivative of (2.33) with respect to $\mathbf{w}_X$ and setting equal to zero gives us,

$$\frac{\partial L(\mathbf{w}_X, \lambda)}{\partial \mathbf{w}_X} = \mathbf{S}_B \mathbf{w}_X - \lambda \mathbf{S}_W \mathbf{w}_X = \mathbf{0},$$

which yields the following generalized eigenvalue problem,

$$\mathbf{S}_B \mathbf{w}_X = \lambda \mathbf{S}_W \mathbf{w}_X. \tag{2.34}$$

Points are then projected onto the resulting eigenvectors $\mathbf{w}_X$ giving $\mathbf{x}_i^* = \mathbf{x}_i^T \mathbf{w}_X$. An observation $\mathbf{x}_i^*$ is assigned to a class based on which class center $\mathbf{m}_j^* = \mathbf{m}_j^T \mathbf{w}_X$, $j = 1 \ldots, k$ is nearest,

$$\arg\min_j ||\mathbf{x}_i^* - \mathbf{m}_j^*||^2 \tag{2.35}$$

We show that for the two class problem a simple closed form solution exists for the direction $\mathbf{w}_X$.

Theorem 2.5.1. *Given the optimization problem in (2.32) when the number of classes k is equal to 2 then*

$$\mathbf{w}_X^* = \frac{1}{\sqrt{\lambda^*}} \mathbf{S}_W^{-1} (\mathbf{m}_1 - \mathbf{m}_2), \tag{2.36}$$

where $\lambda^* = \frac{n\lambda}{n_1 n_2}$.

Proof. First we observe that when the number of classes is equal to two the between-class sum of squares can be expressed as

$$\mathbf{S}_B = \frac{n_1 n_2}{n}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T.$$

For notational purposes we rewrite the generalized eigenvalue problem in (2.34) as

$$\mathbf{S}_B^* \mathbf{w}_X = \lambda^* \mathbf{S}_W \mathbf{w}_X, \tag{2.37}$$

where $\mathbf{S}_B^* = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$. From the generalized eigenvalue problem in (2.37) and using the constraints in the optimization problem (2.32) we have

$$\begin{aligned}
\lambda^* &= \mathbf{w}_X^T \mathbf{S}_B^* \mathbf{w}_X \\
&= \mathbf{w}_X^T (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w}_X \\
&= \frac{1}{\sqrt{\lambda^*}}(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \frac{1}{\sqrt{\lambda^*}} \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2).
\end{aligned}$$

Rearranging terms gives us

$$\lambda^* = (\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{2.38}$$

Next, starting with the left hand side of (2.37) and substituting in for $\mathbf{w}_X$ and $\mathbf{S}_B^*$, we have

$$\begin{aligned}
\mathbf{S}_B^* \mathbf{w}_X &= \frac{1}{\sqrt{\lambda^*}}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \\
&= \sqrt{\lambda^*}(\mathbf{m}_1 - \mathbf{m}_2)
\end{aligned}$$

Now looking at the right hand side of (2.37) we have

$$\begin{aligned}\lambda^* \mathbf{S}_W \mathbf{w}_X &= \lambda^* \frac{1}{\sqrt{\lambda^*}} \mathbf{S}_W \mathbf{S}_W^{-1} (\mathbf{m}_1 - \mathbf{m}_2) \\ &= \sqrt{\lambda^*} (\mathbf{m}_1 - \mathbf{m}_2).\end{aligned}$$

Thus we have shown that the left and right sides of (2.37) are equal. Also note that conditions in (2.32) are satisfied

$$\begin{aligned}\mathbf{w}_X^T \mathbf{S}_W \mathbf{w}_X &= \frac{1}{\lambda^*} (\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{S}_W^{-1} \mathbf{S}_W \mathbf{S}_W^{-1} (\mathbf{m}_1 - \mathbf{m}_2) \\ &= \frac{(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{S}_W^{-1} (\mathbf{m}_1 - \mathbf{m}_2)}{(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{S}_W^{-1} (\mathbf{m}_1 - \mathbf{m}_2)} \\ &= 1.\end{aligned}$$

From this we can see that (2.36) is an eigenvector of the generalized eigenvalue problem in (2.34). In order to show that this is in fact the leading eigenvector note that because the rank of the between-class scatter matrix is 1 there are at most 1 non-zero eigenvalues in the generalized eigenvalue problem (2.37). However, from (2.38) it is clear that λ^* and therefore λ will be strictly positive so long as $\mathbf{m_1} \neq \mathbf{m_2}$ and $\mathbf{S}_W$ is non-singular. Therefore we have that (2.36) is the leading eigenvector of (2.34). □

2.5.2 LDA Solved by CCA

In this section we derive the connection between LDA and CCA. It will be shown that the generalized eigen problem in (2.4), is essentially the same, modulo a scalar, as the generalized eigen problem in (2.34). Letting $\mathbf{Y}$ be defined as in Section 2.5.1 (the

matrix of class labels), we have,

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1}_{n_1} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_{n_k} \end{pmatrix}.$$

From this it is easy to see that,

$$\mathbf{S}_{YX} = \mathbf{Y}^T\mathbf{X} = \begin{pmatrix} n_1\mathbf{m}_1^T \\ n_2\mathbf{m}_2^T \\ \vdots \\ n_k\mathbf{m}_k^T \end{pmatrix}.$$

It follows that

$$\mathbf{S}_{YY}^{-1} = (\mathbf{Y}^T\mathbf{Y})^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{pmatrix}.$$

Using these results we have

$$\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX} = \sum_{i=1}^{k} n_i\mathbf{m}_i\mathbf{m}_i^T = \mathbf{S}_B \tag{2.39}$$

Starting with

$$\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\mathbf{w}_X = \rho_{\mathcal{H}}^2\mathbf{S}_{XX}\mathbf{w}_X,$$

and using (2.28) and (2.39) this can be rewritten as,

$$\mathbf{S}_B\mathbf{w}_X = \rho_{\mathcal{H}}^2\mathbf{S}_T\mathbf{w}_X.$$

Finally using (2.31) and rearranging terms gives us,

$$\mathbf{S}_B\mathbf{w}_X = \frac{\rho_{\mathcal{H}}^2}{1-\rho_{\mathcal{H}}^2}\mathbf{S}_W\mathbf{w}_X. \tag{2.40}$$

This is identical to (2.34) but with $\lambda = \frac{\rho_{\mathcal{H}}^2}{1-\rho_{\mathcal{H}}^2}$.

This relationship will prove useful later in developing intuition and theory about CCA and its ability to find and understand the co-dependence of subpopulation's between spaces.

2.6 CCA Performance on Real Data

We now apply the methods described in this chapter on the RLP800 and WDI data sets described previously. Figure 2.10 is a scatterplot matrix showing the projection of the training (black) and testing (red) data onto the first three canonical vectors. Figure 2.11 is a plot of all the canonical correlations and density plots of the canonical directions themselves.

Regularized CCA was used, with parameters $\kappa_X = \kappa_Y = 0.1$, the number of dimensions projected onto was $p_X = p_Y = 100$ and the number of neighbors used in the prediction was 60. These parameters were selected via a simple cross validation scheme using a randomly selected subset of 537 and 100 points from the training data as "training" and "testing" sets. Values for the tuning parameters were found by searching over values of $\kappa_X = \kappa_Y = \{0.1, 1, 10, 20\}$, $p_X = p_Y = \{25, 50, 75, 100, 125\}$ and $k = \{5, 10, 20, 40, 60, 80\}$, the final set of parameters were selected based on which produced the lowest average rank (see Section 1.3.1 for details), which in this case was approximately 8.5.

Figure 2.11 shows the distribution of each of the first three canonical variates (left) as well as the canonical correlations for each of the 150 variates (right). As can be seen the leading canonical correlations are fairly large indicating that a strong relationship exists

between spaces.

Figure 2.12 is a scatterplot matrix of the first three pairs of canonical variates in protein and ligand space respectively with one test point highlighted (red) and its predicted value (green, ligand space only). As can be seen the prediction is fairly accurate.

Figure 2.13 is a histogram of the ranks associated with our prediction using regularized CCA. The average rank in this case was approximately 10, indicated by the vertical red line. This is a significant improvement over the current methodology implemented in Oloff *et al.* (2006), where the average rank was 18.1 (vertical green line).

Figure 2.14 shows the results from prediction on the WDI data set. The mean predicted rank using CCA is approximately 67 (green line), the previous method yielded a mean result of 310 (red line).

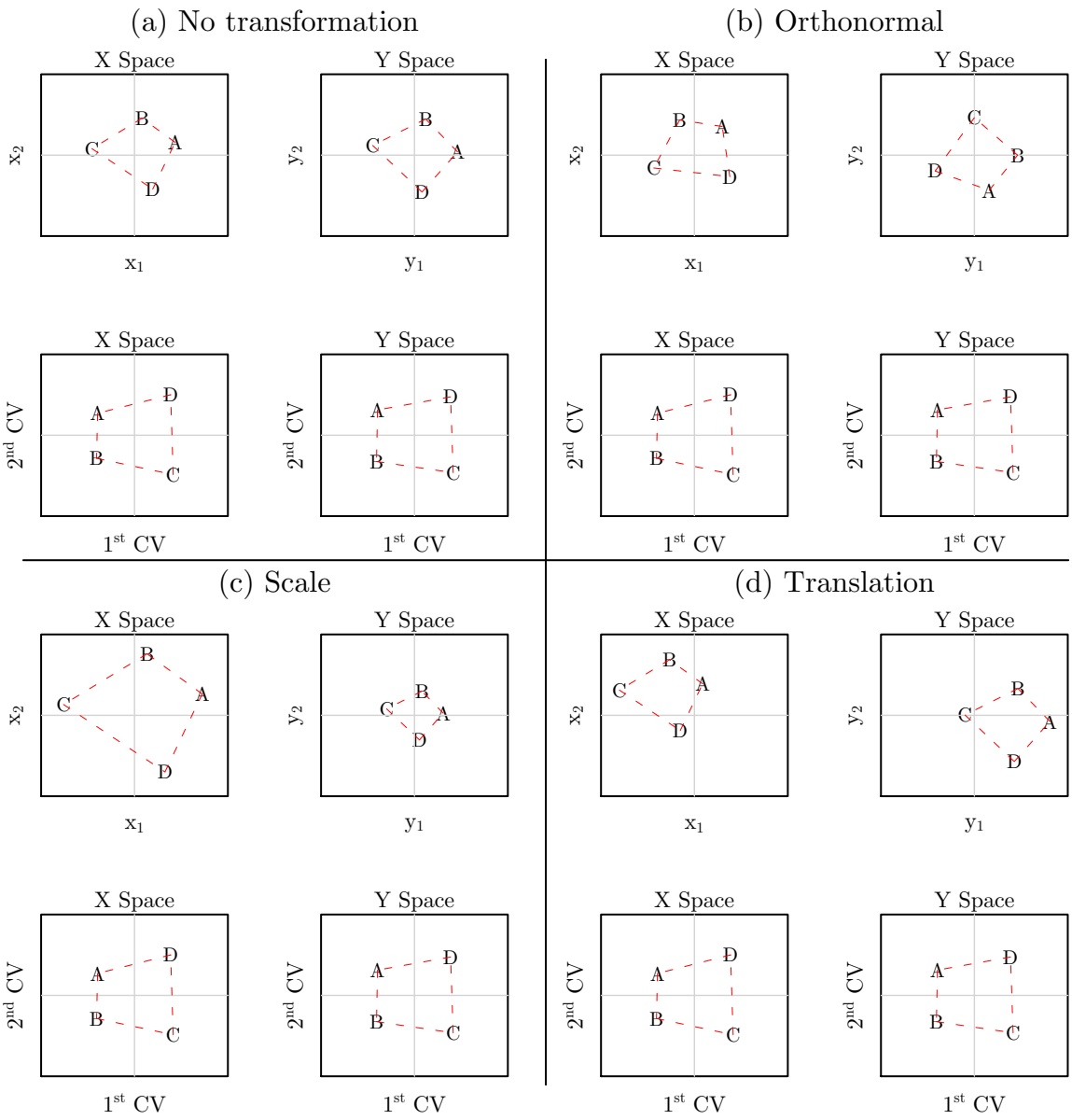

Figure 2.1: *Four groups of four plots, each group consists of a plot of the X and Y raw data spaces (top left and right) and the projections of these spaces onto their respective first and second canonical directions (bottom left and right). Group (a) shows the data with no transformation. All subsequent groups have been transformed. In group (b) The data in the X space have been rotated* $30°$ *counterclockwise and in the Y space the data have been rotated* $75°$ *clockwise. In group (c) the points in the X space have been scaled by* $\frac{5}{3}$ *and in the space Y by* $\frac{2}{3}$*. In group (d) the means of the points have been shifted such that the centers are now at* $\left(-\frac{3}{4}, \frac{1}{2}\right)$ *and* $\left(\frac{3}{4}, -\frac{1}{4}\right)$*. The point of all these illustrations is that in all four groups of plots the bottom left and right plots, the projections into the canonical correlation space, are all the same. This provides visual confirmation of CCA's invariance properties.*

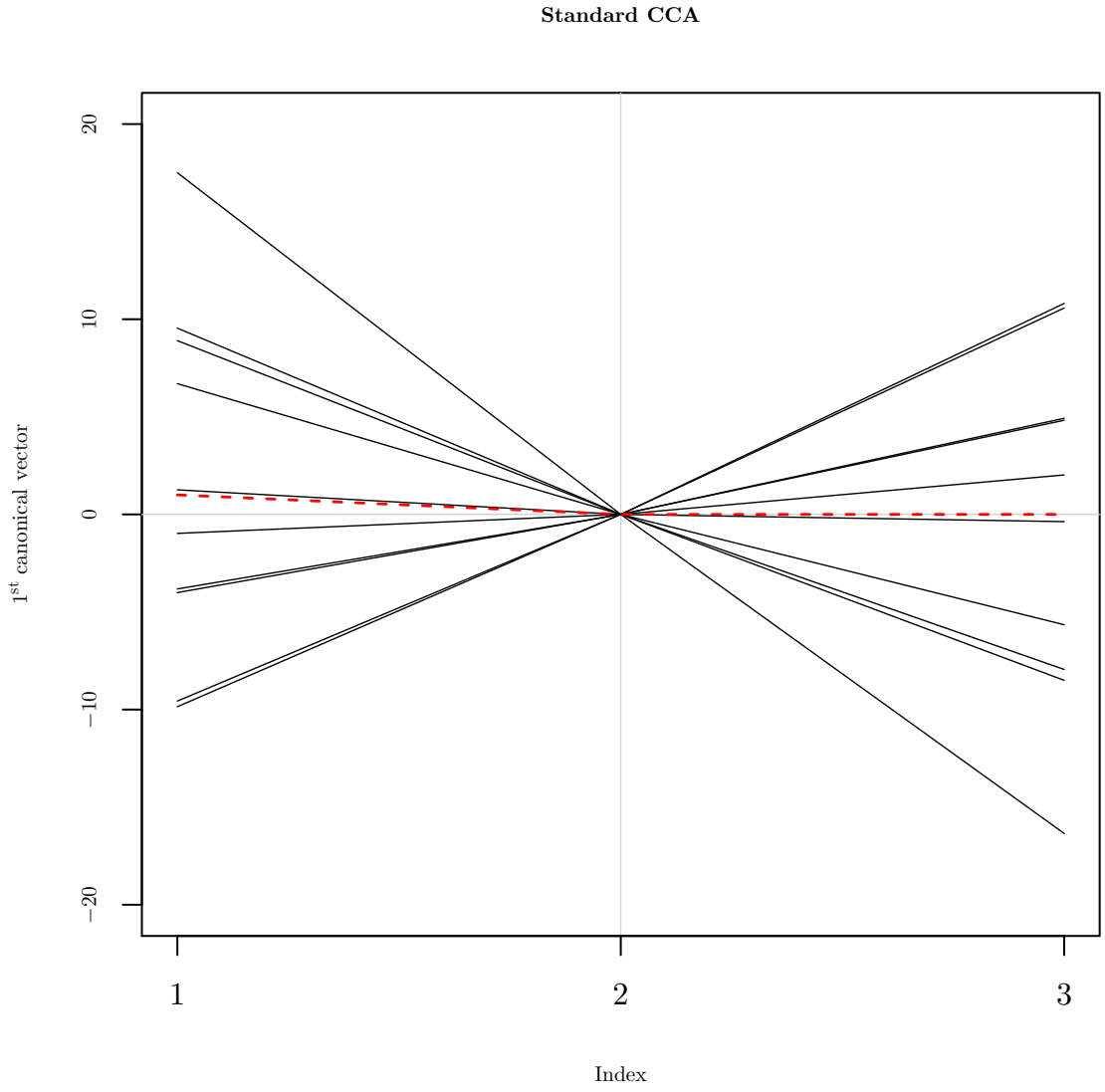

Figure 2.2: *A simulated example of the canonical vectors in X space in the presence of strong multicollinearity between the first and third descriptors. The major issue here is the large amount of variation in the canonical directions from one sample to the next despite the fact that the data are drawn from the same distribution.*

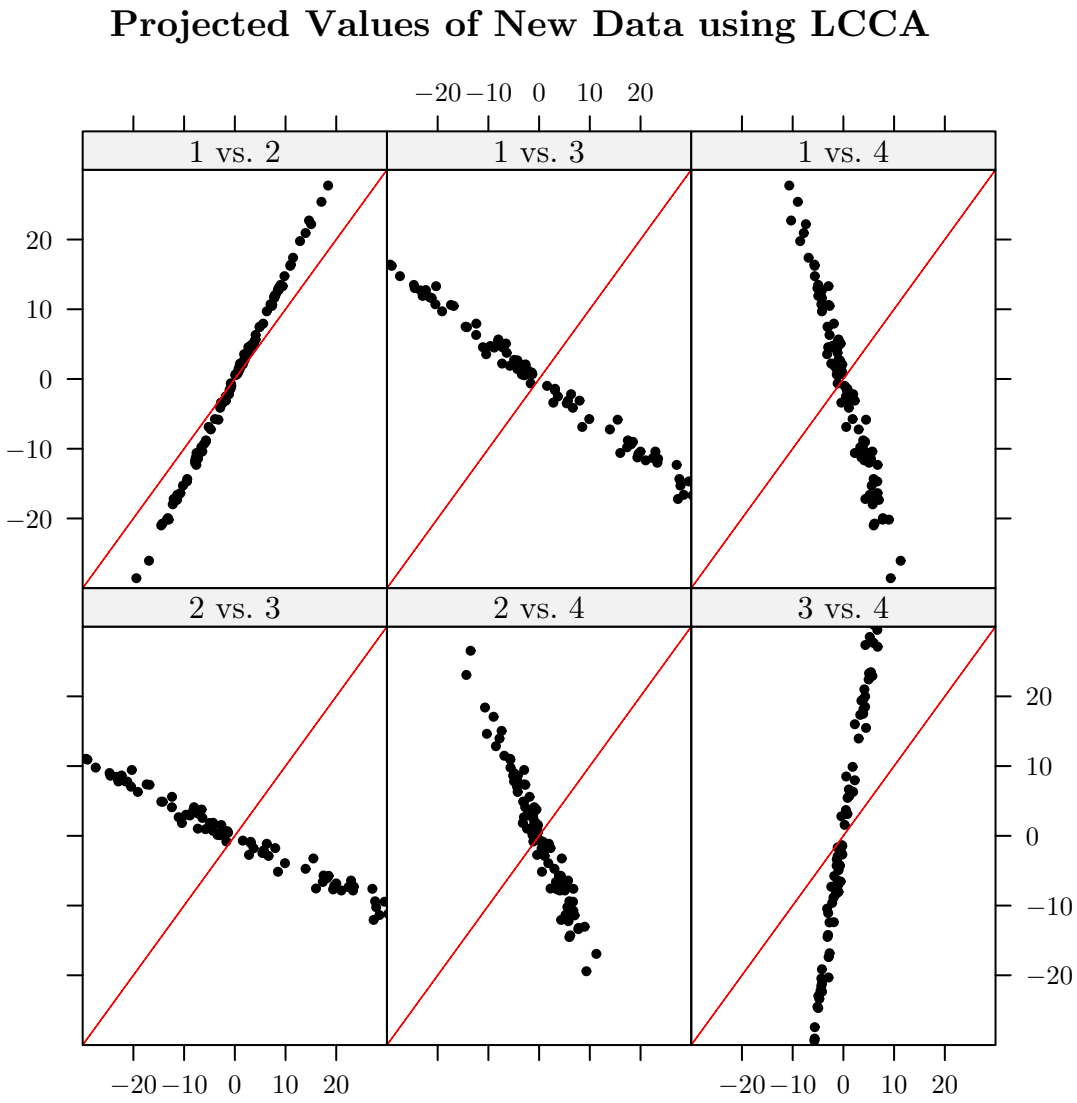

Figure 2.3: *Plot of the projected values of a new set of observations onto the canonical direction vectors shown in Figure 2.2. Each panel shows the plot of one projection versus another (only four projections are shown).*

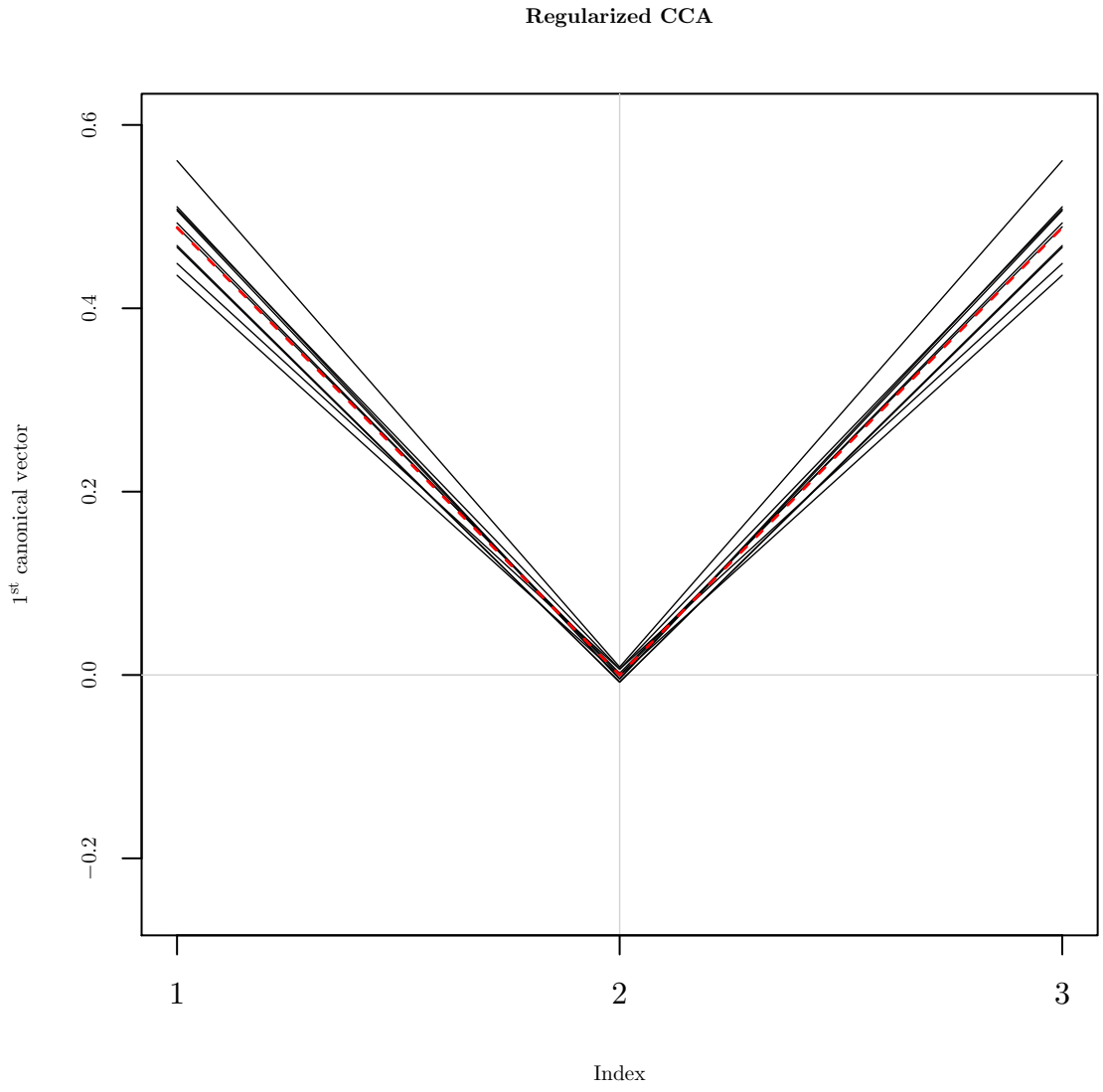

Figure 2.4: *This is a plot of the canonical direction vectors found from RCCA. The dashed red line is the theoretical direction. In contrast to the direction found by linear CCA the directions found by regularized CCA display little variation from one sample to the next and lie near the theoretical direction.*

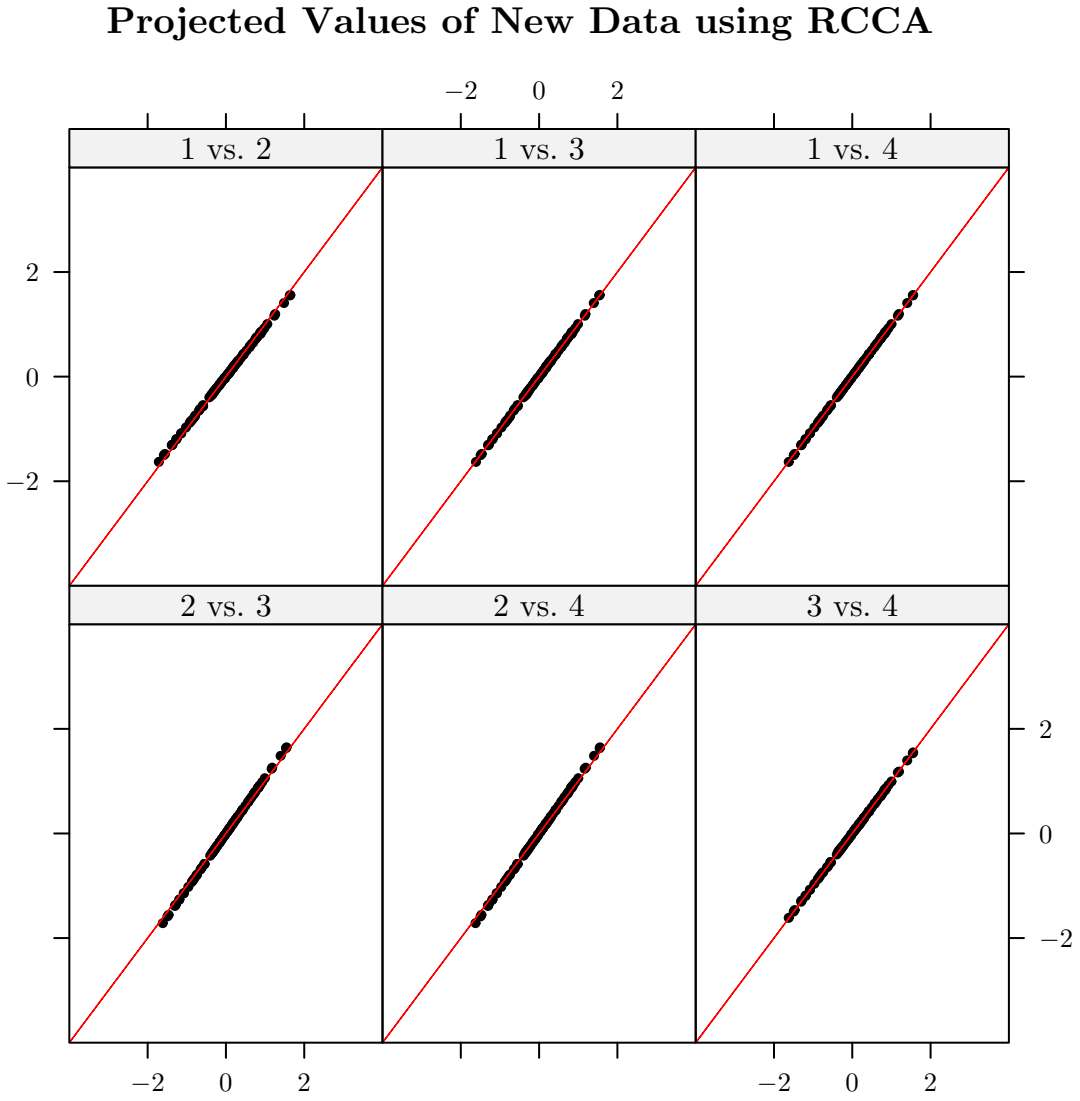

Figure 2.5: *A plot of each pair of projected values of the new data onto each of the direction vectors shown in Figure 2.4 against one another. As can be seen the projections are all quite similar to one another, in contrast to standard CCA where there was a great deal of variation from one set of directions to the next.*

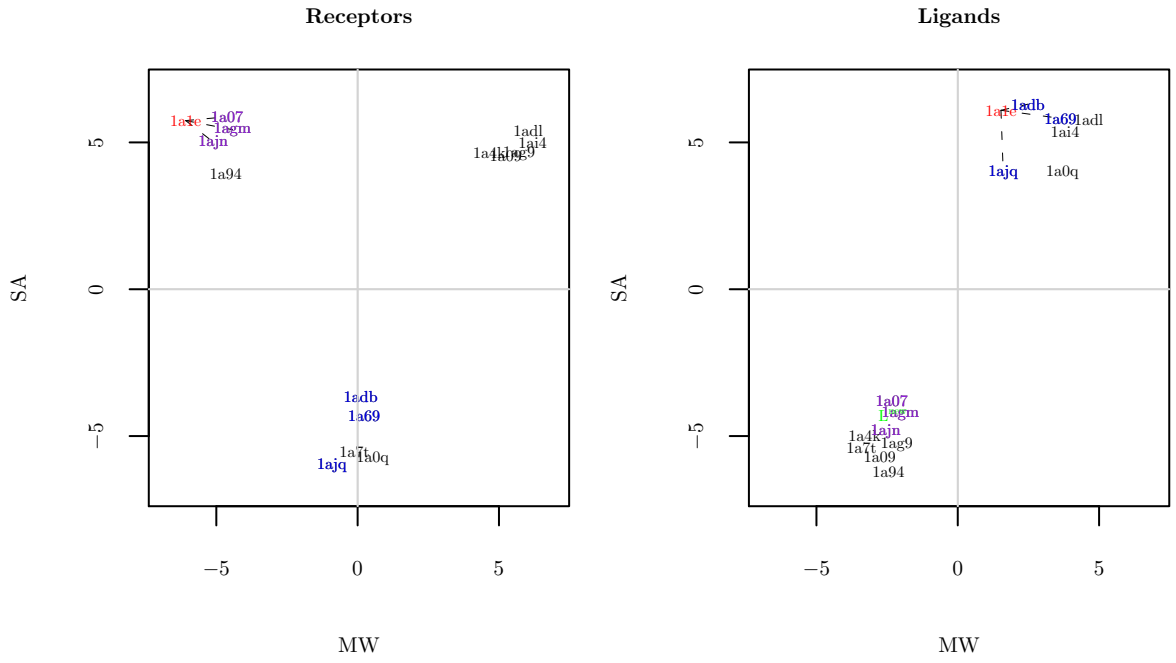

Figure 2.6: *New toy example data. The points highlighted in red correspond to the protein ligand pair 1a1e, and the points connected to it by dashed black lines are its three nearest neighbors in each space. The observations highlighted in blue and purple are neighbors only in the protein and ligand spaces respectively. The green point L^{new} in the ligand space corresponds to a simple weighted average of the cyan point and the purple points, i.e. of the nearest neighbors of 1a1e in the protein space.*

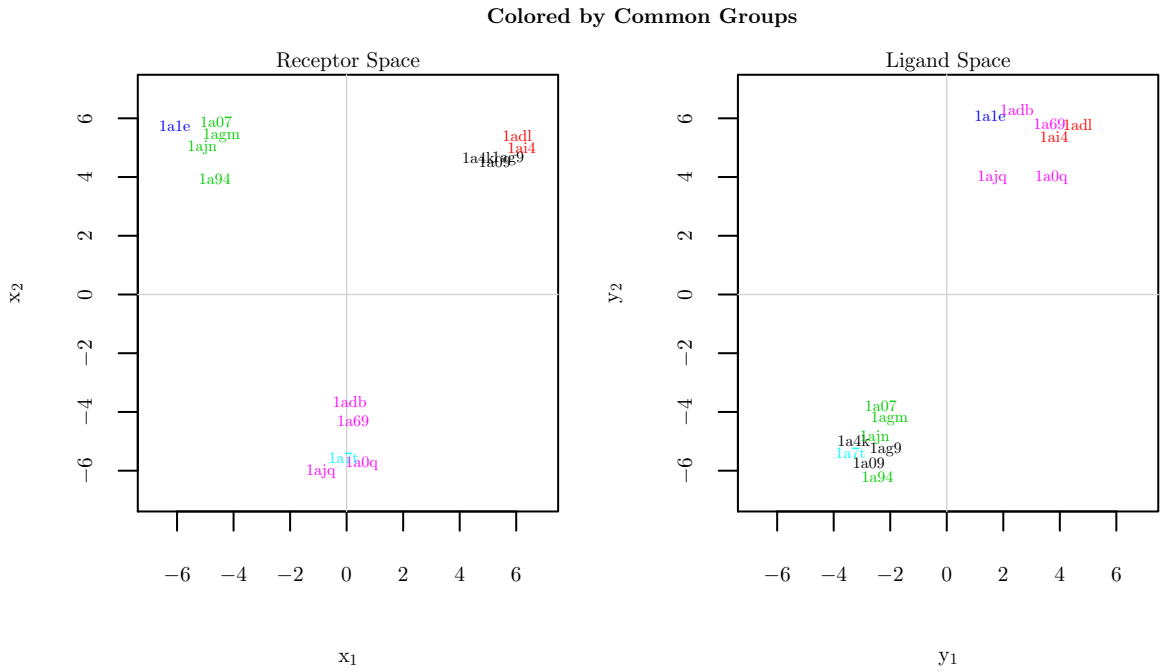

Figure 2.7: *These plots depict the same data as in Figure 2.6 with points highlighted according to whether they appear in the same cluster in both spaces. For example, consider the green points, these observations appear in the same cluster in both protein and ligand space. The data has been generated such that points that appear in the same cluster in both spaces are highly correlated.*

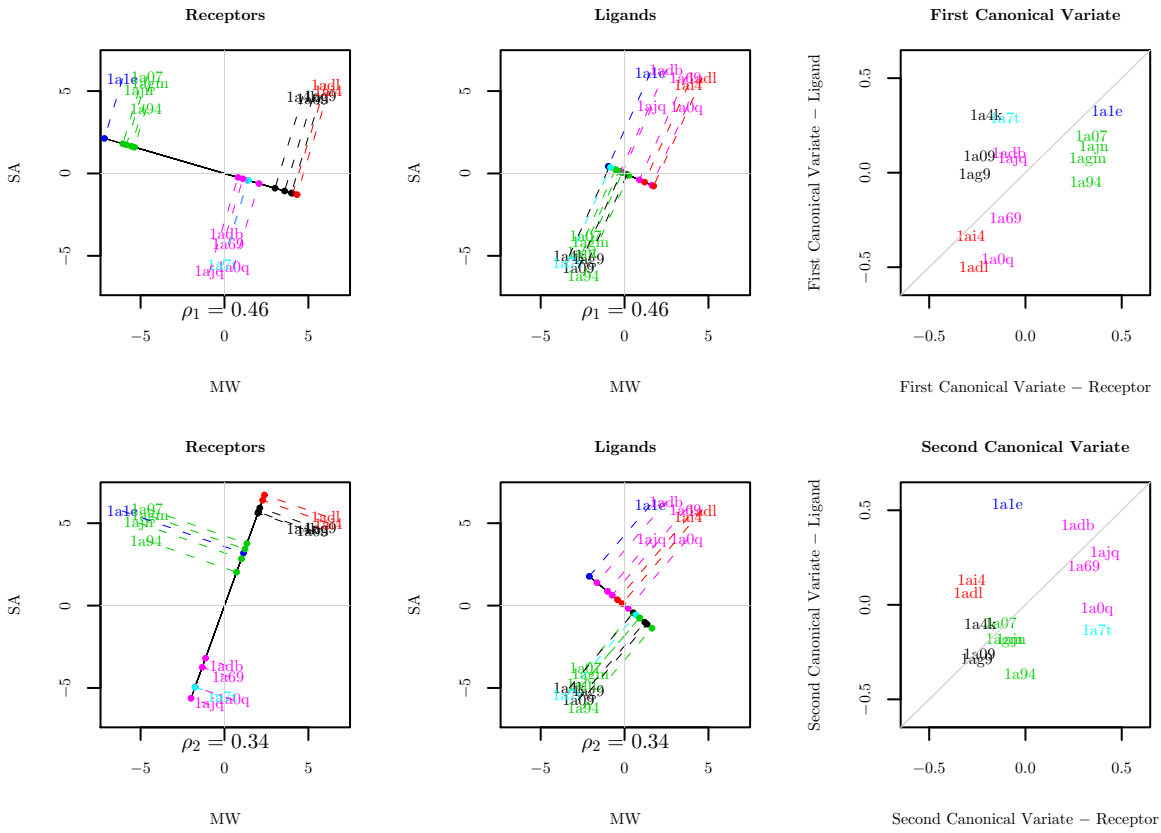

Figure 2.8: *The linear CCA direction vectors and the projected value of each point colored as in Figure 2.7. On the first row of plots the first two panels show the first direction vector and the projections onto it in protein and ligand space respectively. The last panel on the top row of plots is a plot of the first canonical variate in protein space against the first canonical vector in ligand space. If the directions we found were able to capture the underlying relationship between the two spaces we would expect these points to fall along the 45° line. The second row of plots shows the same set of plots as the top row of plots but for the second canonical direction. A visual assessment of the projected values of the observations in each space shows how different the distribution of points is along the canonical vectors. This discrepancy is further highlighted by noting how different the location of the colored points are along the canonical vectors. The implication of this is that the correlation, i.e. alignment is not very good as reflected by the canonical correlations of 0.46 and 0.34.*

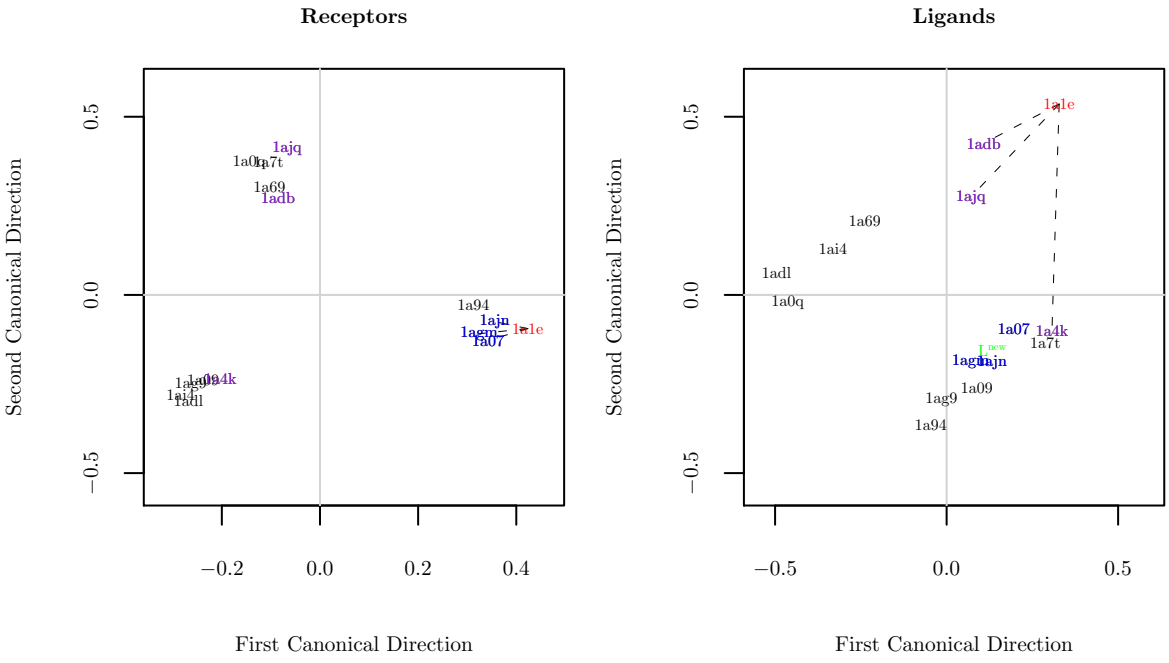

Figure 2.9: *CCA Projected space. In contrast to Figure 1.4, linear CCA appears to have made the prediction worse.*

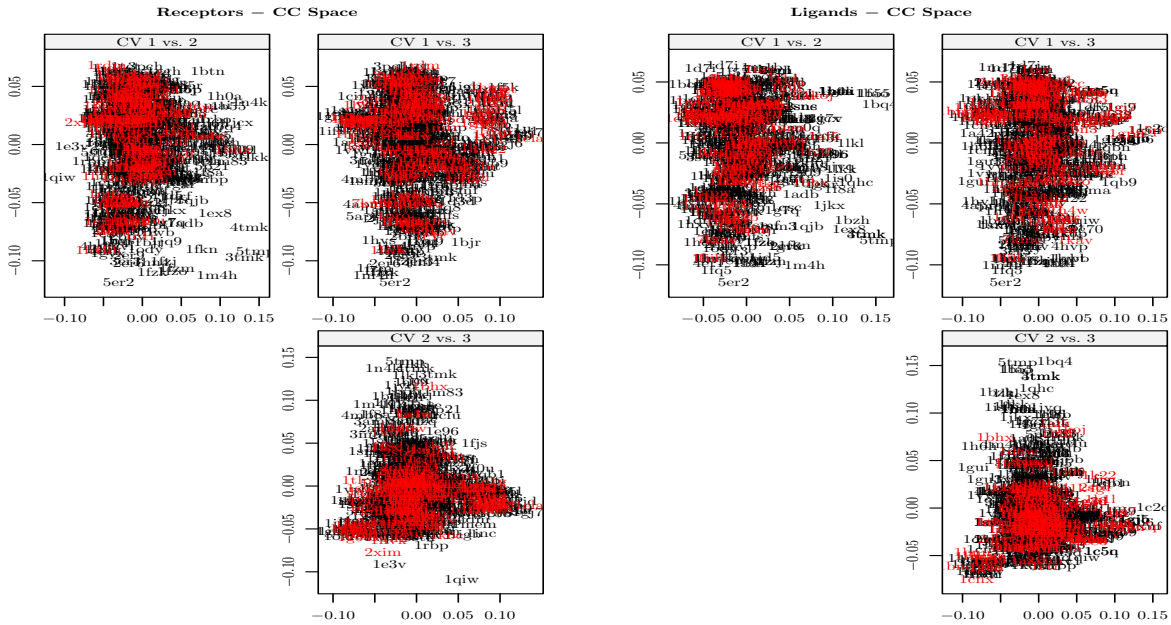

Figure 2.10: *A plot of the canonical correlations from the RLP800 data set with the training data shown in black and the test data shown in red. From a visual assessment of the data it appears as though the two spaces are fairly well aligned.*

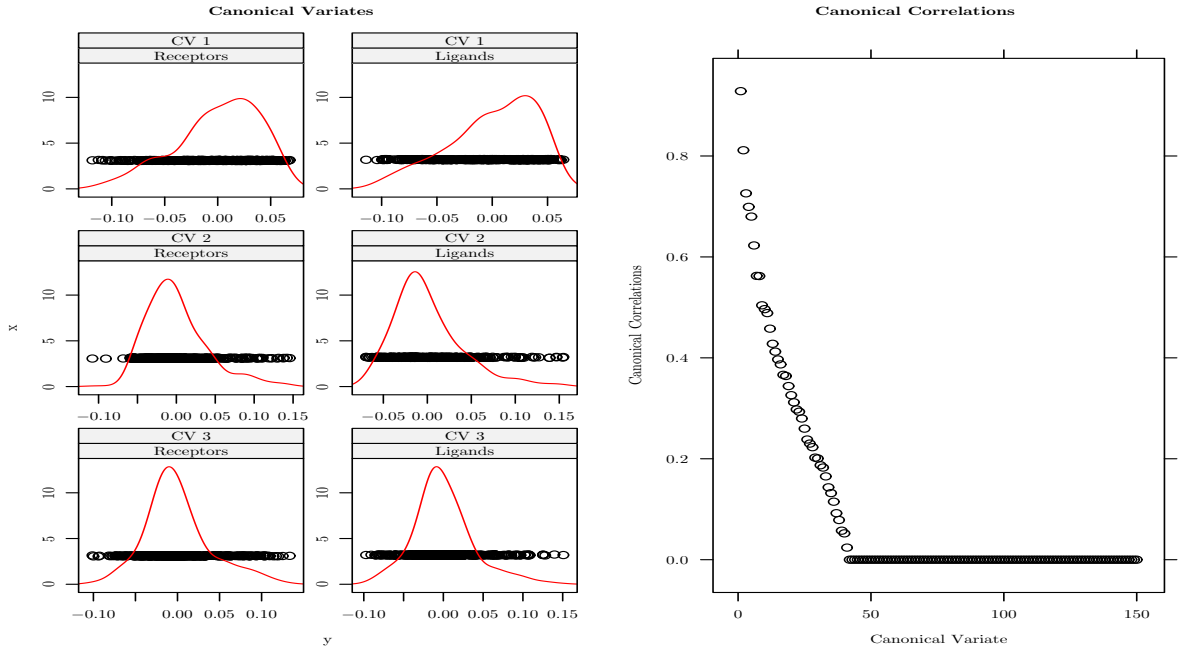

Figure 2.11: *On the left are plots of the first three canonical variates in protein and ligand space respectively. The red curves are the associated density estimates of the canonical variates. This is meant to provide some insight into the distribution of the data within a space as well as how well aligned points are between spaces. On the right is a plot of the canonical correlations associated with each of the 150 canonical vectors.*

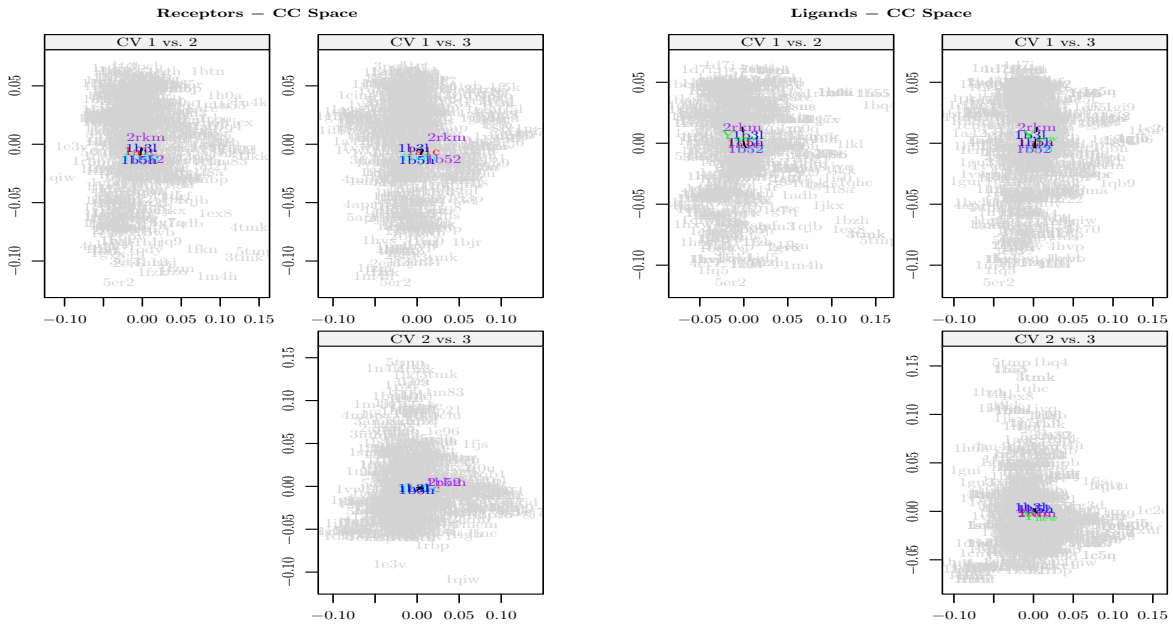

Figure 2.12: *Similar to Figure 2.10 but with one of the test points highlighted as well as its three nearest neighbors. The color scheme is similar to that of the toy examples discussed earlier this section.*

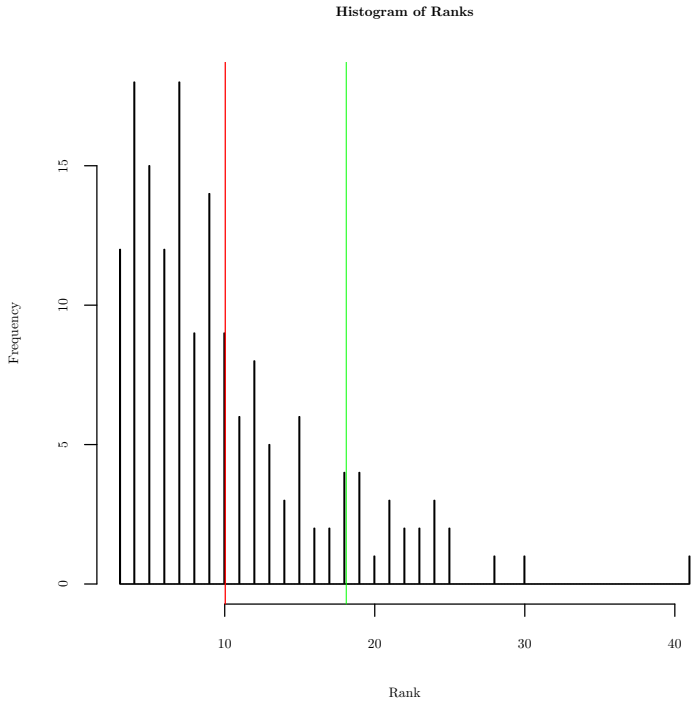

Figure 2.13: *A histogram showing the ranks resulting from prediction on the test data from the RLP800 dataset. The vertical red line indicates the average rank (approximately 10) using CCA and the vertical green line the method implemented in Oloff* et al. *(2006) (approximately 18).*

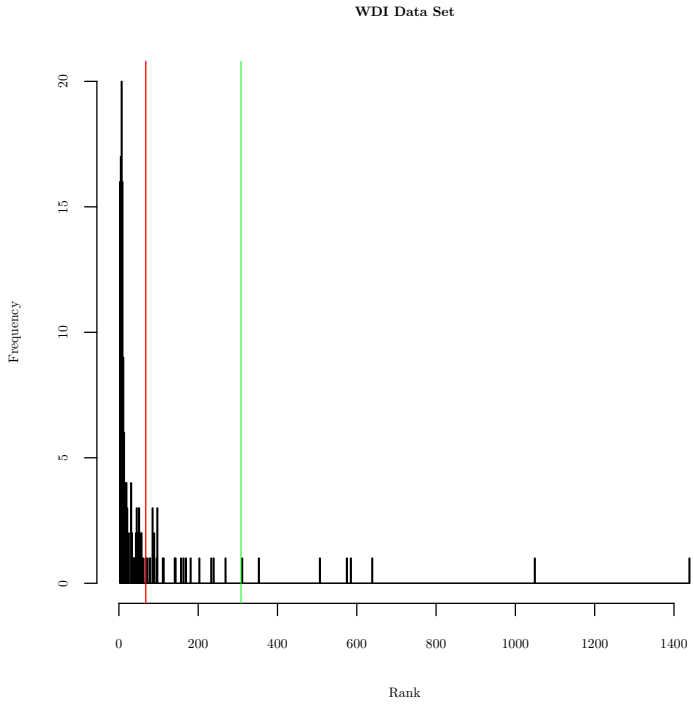

Figure 2.14: *Similar to the histogram above but using the WDI data. The mean rank using CCA is approximately 67 while the previous method yielded a mean result of approximately 310*

CHAPTER 3

Kernel Methods

In Section 1.1 we introduced the concept of a similarity measure $\mathcal{S}$ and a pair of spaces of functions $\mathcal{H} = \{\mathcal{H}_X, \mathcal{H}_Y\}$ over which it was defined. In the context of standard CCA, $\mathcal{S}$ is the correlation function and $\mathcal{H}$ contains the Hilbert spaces of functions containing the bilinear maps $f_X(X) = \langle X, \mathbf{w}_X \rangle$ and $f_Y(Y) = \langle Y, \mathbf{w}_Y \rangle$. As we saw in Section 2.2 standard CCA may encounter problems when the relationships between and within distributions of points cannot be described by a simple linear combination of the descriptors. For this reason it is useful to consider an alternative space of functions which is more appropriate for learning these complex relationships. In the following sections examples and details surrounding such a space of functions will be discussed.

In developing intuition and methodology related to kernel methods we follow the discussion of Schölkopf and Smola (2002) (pp 25-60). From here on we will refer to the spaces $Y \in \mathcal{X}_X$ and $X \in \mathcal{X}_Y$ as the object space representations of the data. The spaces $\mathcal{X}_X$ and $\mathcal{X}_Y$ are nonempty sets from which the observations $\mathbf{x}_i$ and $\mathbf{y}_i$ are sampled. This general definition of the object space is meant to emphasize the fact that the data can be any of a number of different types. For example, we may be interested in using the amino acid sequence of a protein in place of its descriptors in our analysis. However, unless stated otherwise we only consider the object spaces in $\mathcal{X}_X = \mathbb{R}^{d_X}$ and $\mathcal{X}_Y = \mathbb{R}^{d_Y}$.

The spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$, containing the functions f_X and f_Y discussed in Section 1.1 will be referred to as the feature spaces. The maps Φ_X and Φ_Y define a mapping from

object space (the original protein and ligand space) into feature space,

$$\begin{aligned} \Phi_X &: \mathcal{X}_X \to \mathcal{H}_X, \\ \Phi_Y &: \mathcal{X}_Y \to \mathcal{H}_Y. \end{aligned} \tag{3.1}$$

3.1 Example: Feature Maps

To illustrate the type of feature maps we may encounter consider the following toy example: Recall the general framework of the examples discussed in Section 2 but rather than having both protein and ligand space characterized by MW and SA, suppose that the protein space has two descriptors, call them d_X^1 and d_X^2 and the ligand space has two descriptors, call them d_Y^1 and d_Y^2, shown in Figure 3.1. The observation highlighted in red, 1a94, corresponds to a new protein whose corresponding ligand we are trying to predict. The point highlighted in cyan is one of the 3-nearest neighbors of 1a94 in both spaces. Those points highlighted in blue (and purple) are nearest neighbors in only the protein (and ligand) spaces, respectively. The point L^{new} in the ligand space, highlighted in green is a simple average of the nearest neighbors of the point 1a94 in protein space. Using L^{new} as a prediction of the new ligand would not provide a particularly accurate prediction.

As before we use CCA to try and find a linear combination of the descriptors which best align the two spaces. Figure 3.2 is a plot of the projections onto the first and second canonical variates in protein and ligand space. The color scheme is the same as in Figure 3.1. As can be seen standard CCA does not seem to be able to find a good alignment between the two spaces, which is confirmed by the low values of the canonical correlations, 0.47 and 0.15 respectively for the first and second directions.

Suppose it is believed that some type of functional relationship exists between the descriptors across spaces that is best characterized by looking at the second order poly-

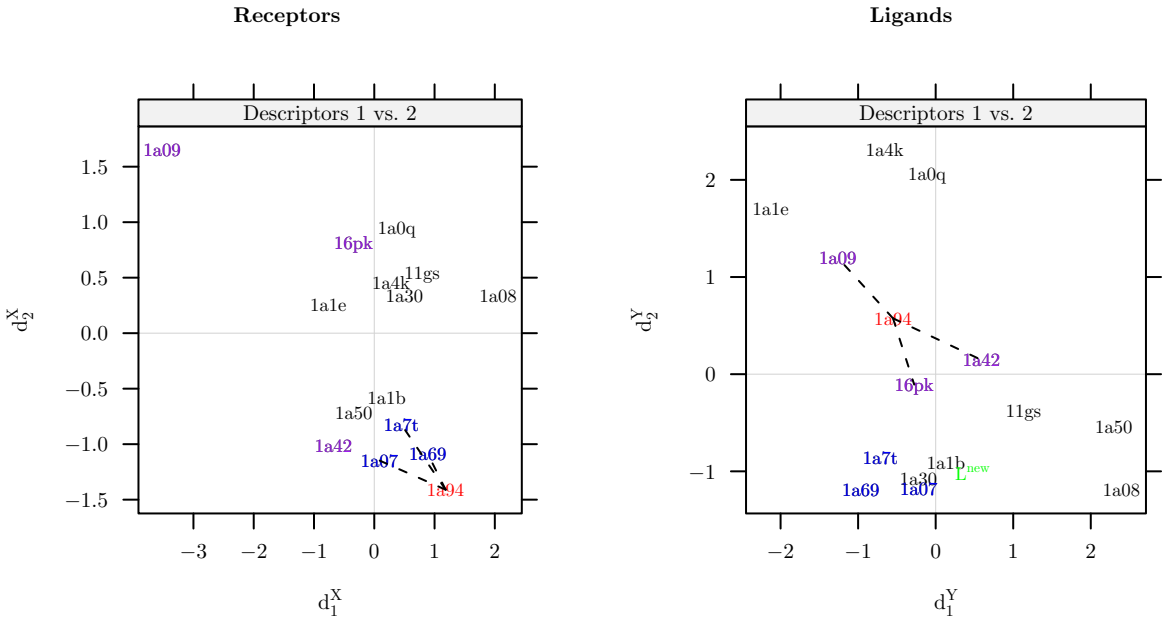

Figure 3.1: *A plot of the data generated such that the underlying relationship between points is non-linear. The observation highlighted in red, 1a94, is the new observations which we are trying to predict. The points joined to it by dashed black lines are its nearest neighbors. The points highlighted in cyan correspond to a point which is a nearest neighbor of 1a94 in both spaces. Points highlighted in blue and purple correspond to points which are only neighbors in either protein or ligand space respectively. The point labeled L^{new} in ligand space corresponds to a simple average of the points 1a08, 1a09 and 1a1b, i.e. the nearest neighbors of the point 1a94 in protein space.*

nomials of the descriptors within each space, that is,

$$\begin{aligned} \Phi_X &: (d_X^1, d_X^2) \rightarrow ((d_X^1)^2, (d_X^2)^2, d_X^1 d_X^2), \\ \Phi_Y &: (d_Y^1, d_Y^2) \rightarrow ((d_Y^1)^2, (d_Y^2)^2, d_Y^1 d_Y^2). \end{aligned} \tag{3.2}$$

Figures 3.3 and 3.4 are plots of proteins and ligands respectively embedded in this three dimensional space. As can be seen there are now two neighbors shared in common between spaces (colored in cyan). Furthermore the prediction of the new observation, L^{new} (in green) by a simple average of its three nearest neighbors in feature space is, by comparison, much closer to the actual value than the corresponding prediction in object space.

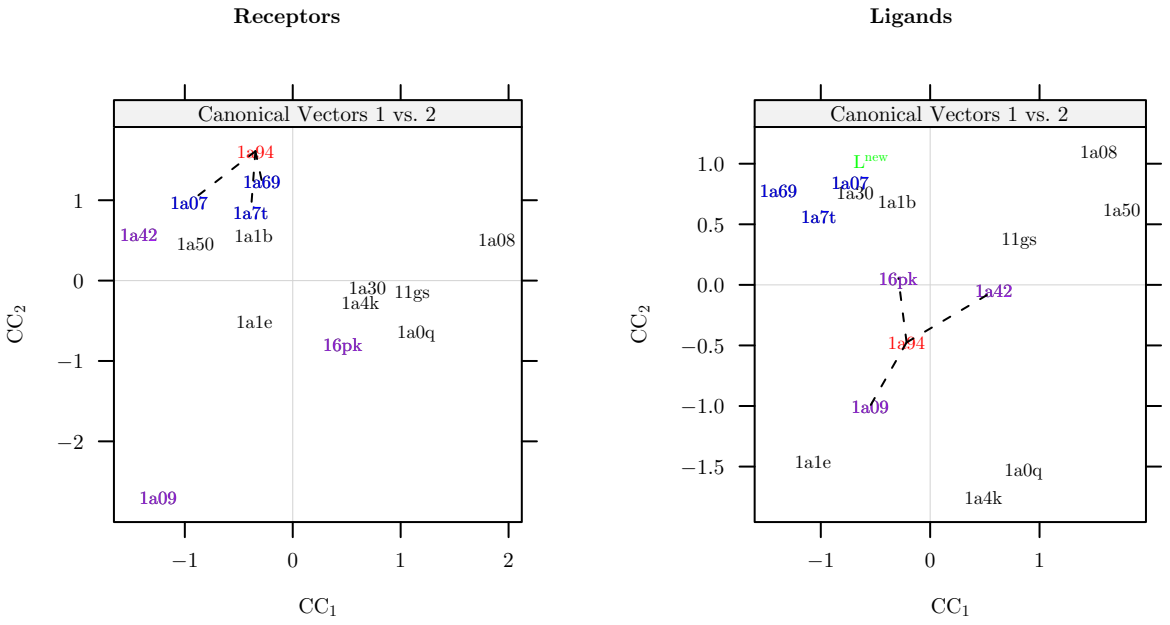

Figure 3.2: *A plot of the data projected onto the first two canonical vectors in both protein and ligand spaces. The directions found by standard CCA do not provide a good alignment between the two spaces.*

As before CCA is used on this transformed data, now in feature space, to align the space of proteins and ligands. Figures 3.5 and 3.6 show the plots of the projected data. Note that now both the new protein and its ligand (highlighted in red) share the same neighbors. The quality of the alignment is further confirmed by looking at the canonical correlation values which are equal to 1 for each of the directions.

It is worth noting that, due to overfitting, the kernel canonical correlation values can sometimes be artificially large due to strong correlation between features in kernel space. The intuitive ideas are similar to those discussed for linear CCA in Section 2.3. Regularization methods for helping to control these effects in the kernel case will be discussed in Section 3.4.

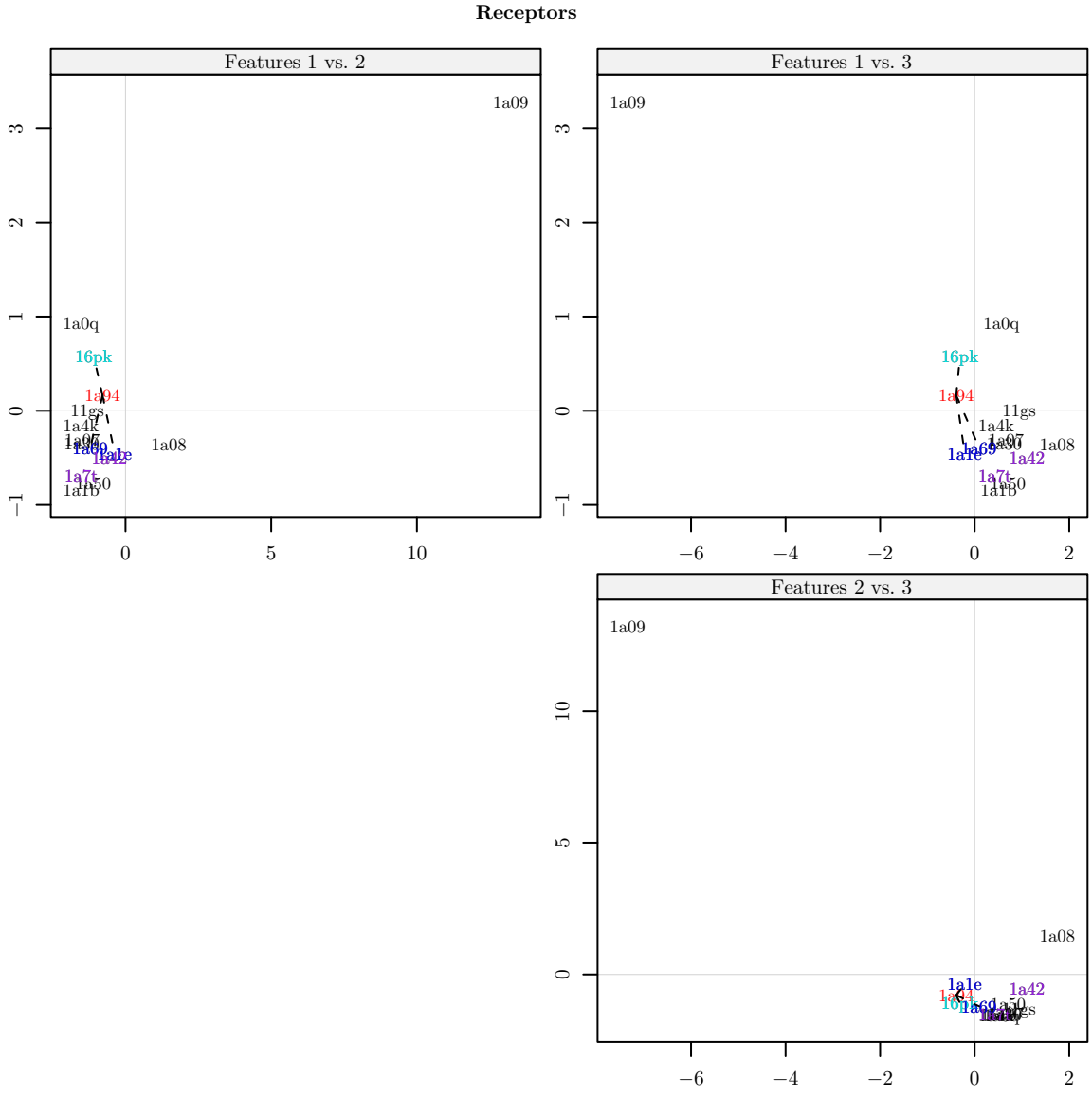

Figure 3.3: *A plot of the protein data in kernel space. The color scheme is the same as in Figure 3.1. Looking at Figure 3.4 the overall correspondence between points in protein space and ligand space is much better than in the original (object) space.*

3.2 Kernels

In contrast to the example just discussed, there are many cases where the types of feature spaces best suited for describing the relationship between spaces cannot be

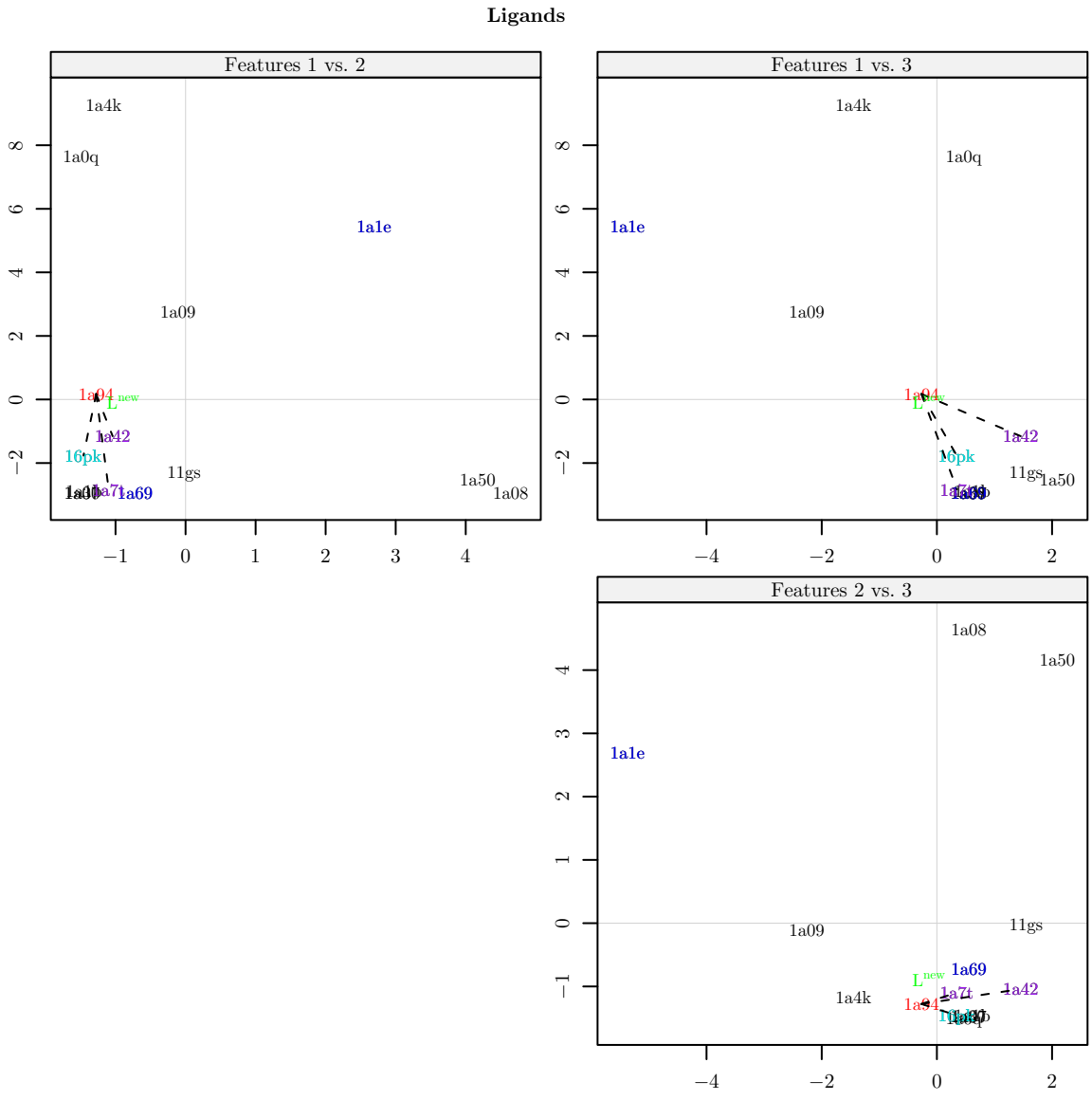

Figure 3.4: *A plot of the ligand data in kernel space. The color scheme is the same as in Figure 3.1. As discussed in Figure 3.3 the correspondence between points in ligand and protein space is much better than in the original object space. This improved mapping will allow CCA to do a better job aligning the two spaces.*

explicitly defined. Specifically, difficulties arise when the space of functions to which Φ_X and Φ_Y belong, define mappings into large or possibly infinite dimensions. However, what is more important than explicitly defining these feature spaces is showing that such

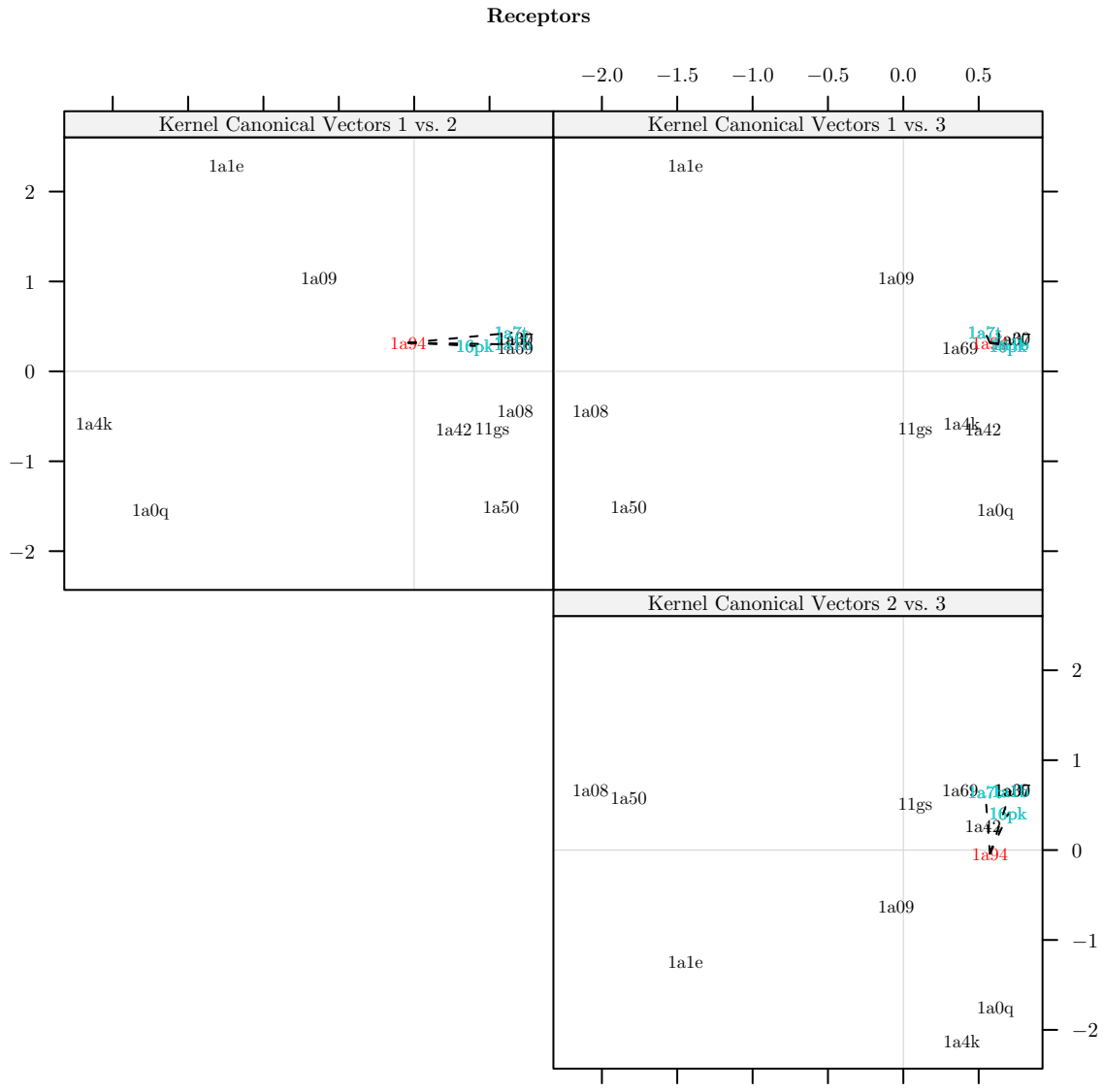

Figure 3.5: *This is a plot of the projection of the data in protein feature space onto the first, second and third canonical vectors. As can be seen not only does the new observation 1a94 (red) have the same 3 nearest neighbors in both protein and ligand space but the prediction of the new ligand, L^{new} highlighted in green below in Figure 3.6 is close to the actual value of 1a94.*

spaces exist and that a inner product can be defined in them. A space equipped with a inner product allows us to understand how points are related to one another in that space. Thus we want to show that given a similarity measure in object space, called a

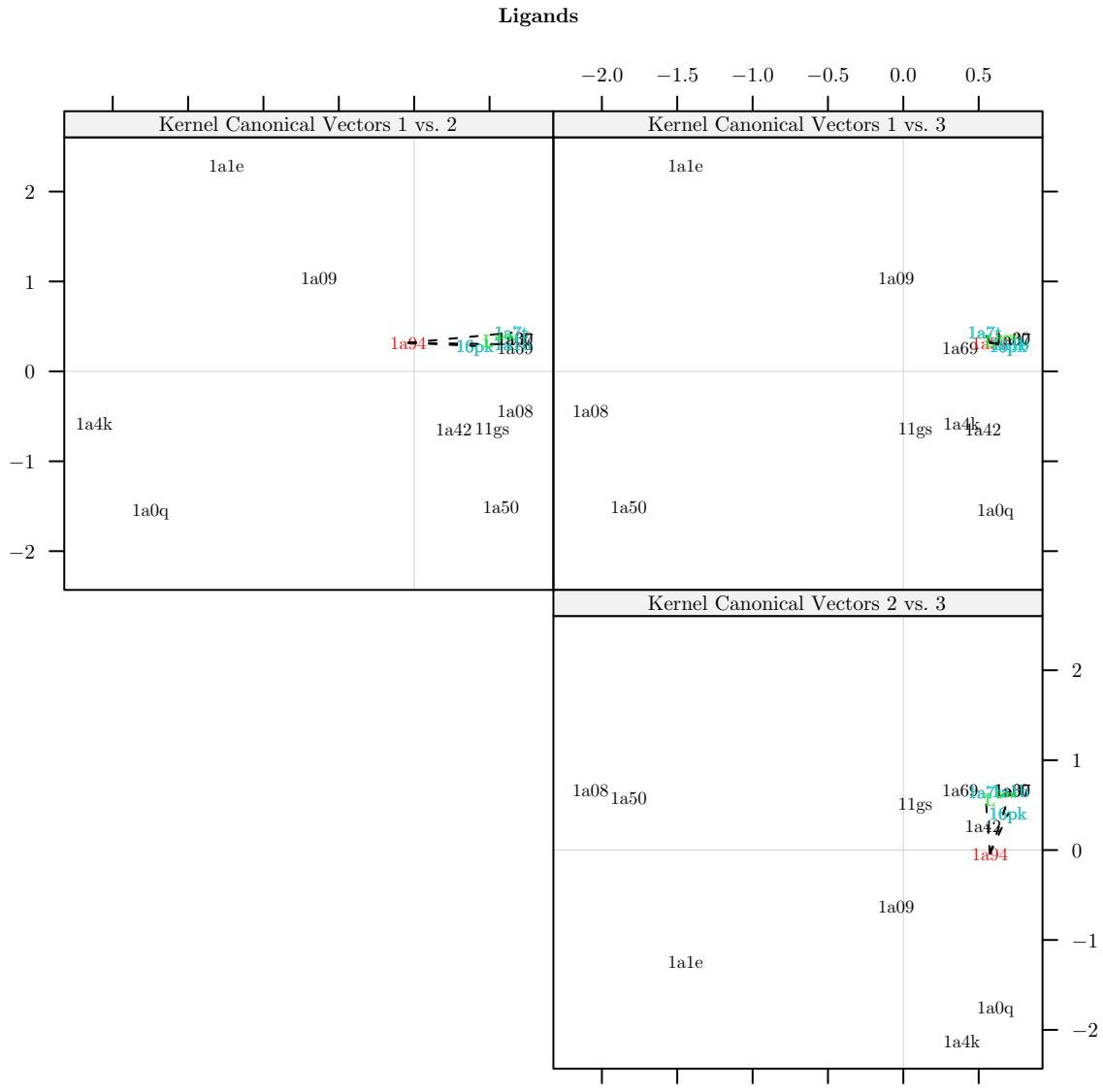

Figure 3.6: *See Figure 3.5 for details.*

kernel, under certain conditions this kernel also defines an inner product in feature space. We give a few definitions associated with kernels, following the development of Schölkopf and Smola (2002).

Definition 3.2.1. *(Gram Matrix) Given a function* $K : \mathcal{X}^2 \to \mathbb{R}$ *and observations* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$*, the* $n \times n$ *matrix* $\mathbf{K}$ *with elements*

$$K_{ij} := K(\mathbf{x}_i, \mathbf{x}_j) \tag{3.3}$$

is called the Gram matrix (or kernel matrix) of K *with respect to* $\mathbf{x}_1, \ldots, \mathbf{x}_n$*.*

Definition 3.2.2. *(Positive Definite Matrix) A real* $n \times n$ *matrix* $\mathbf{K}$ *satisfying*

$$\sum_{i,j} c_i c_j K_{ij} \geq 0 \tag{3.4}$$

for all $c_i \in \mathbb{R}$ *is called positive definite.*

Definition 3.2.3. *((Positive Definite) Kernel) Let* $\mathcal{X}$ *be a nonempty set. A function* K *on* $\mathcal{X} \times \mathcal{X}$ *which for all* $n \in \mathbb{N}$ *and all* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ *gives rise to a positive definite Gram matrix is called a positive definite (pd) kernel.*

With these definitions in place recall the feature maps defined in (3.1) (we restrict our discussion to the space X as the same holds for the space Y). Assuming K_X is a real valued positive definite kernel, replacing $\mathcal{H}_X$ by $\mathbb{R}^{\mathcal{X}_X} := \{f : \mathcal{X}_X \to \mathbb{R}\}$ we have

$$\begin{aligned} \Phi_X :& \mathcal{X}_X \to \mathbb{R}^{\mathcal{X}_X}, \\ & \mathbf{x} \mapsto K_X(., \mathbf{x}). \end{aligned} \tag{3.5}$$

Intuitively the function $\Phi_X(\mathbf{x})$ can be thought of as a function measuring the similarity between $\mathbf{x}$ and all points $\mathbf{x}' \in \mathcal{X}_X$. Here similarity is measured by the function $K_X(\mathbf{x}', \mathbf{x})$ with

$$\Phi_X(\mathbf{x})(.) = K_X(., \mathbf{x}). \tag{3.6}$$

From these definitions it can be shown (Schölkopf and Smola (2002))

1. The image of Φ_X can be represented as a vector space,

2. a dot product can be defined in this vector space, and

3. this dot product satisfies $K(\mathbf{x}, \mathbf{x}') = \langle \Phi_X(\mathbf{x}), \Phi_X(\mathbf{x}') \rangle$.

In particular we have

$$\langle K_X(., \mathbf{x}), f_X \rangle = f_X(\mathbf{x}), \tag{3.7}$$

and

$$\langle K_X(., \mathbf{x}), K_X(., \mathbf{x}') \rangle = K_X(\mathbf{x}, \mathbf{x}'). \tag{3.8}$$

The kernel function K_X as defined above is referred to as a *reproducing kernel*. The space of functions to which the function K_X, endowed with properties (3.7) and (3.8), belongs is called a *reproducing kernel Hilbert space (RKHS)* which is defined as follows

Definition 3.2.4. *Let $\mathcal{X}$ be a nonempty set and $\mathcal{H}$ a Hilbert space of functions $f : \mathcal{H} \to \mathbb{R}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with the dot product $\langle ., . \rangle$ if there exists a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties,*

1. *K has the reproducing property*

$$\langle f, K(\mathbf{x}, .) \rangle = f(\mathbf{x}), \text{ for all } f \in \mathcal{H}.$$

   In particular,

$$\langle K(., \mathbf{x}), K(., \mathbf{x}') \rangle = K(\mathbf{x}, \mathbf{x}').$$

2. *K spans $\mathcal{H}$, $\mathcal{H} = \overline{span\{K(\mathbf{x}, .) | \mathbf{x} \in \mathcal{X}\}}$, where $\overline{X}$ denotes the completion of the set X.*

Furthermore it can be shown that a RKHS uniquely determines the kernel K. With these definitions in place we can now define kernel CCA (KCCA).

3.3 Kernel CCA

The objective of standard CCA is now restated as follows (following the discussion of Hardoon *et al.* (2004),

$$\begin{aligned}\rho_{\mathcal{H}} &= \max_{\mathbf{w}_X,\mathbf{w}_Y} \operatorname{corr}(\langle \Phi_X, \mathbf{w}_X \rangle, \langle \Phi_Y, \mathbf{w}_Y \rangle) \\ &= \max_{\mathbf{w}_X,\mathbf{w}_Y} \frac{\operatorname{cov}(\langle \Phi_X, \mathbf{w}_X \rangle, \langle \Phi_Y, \mathbf{w}_Y \rangle)}{\sqrt{\operatorname{var}(\langle \Phi_X, \mathbf{w}_X \rangle)}\sqrt{\operatorname{var}(\langle \Phi_Y, \mathbf{w}_Y \rangle)}}.\end{aligned} \tag{3.9}$$

Now note that because $\mathbf{w}_X$ (and $\mathbf{w}_Y$) lie in the span of Φ_X (and Φ_Y) these can be re-expressed by the linear transformations

$$\mathbf{w}_X = \Phi_X \alpha_X,$$

$$\mathbf{w}_Y = \Phi_Y \alpha_Y.$$

Plugging this into (3.9) gives us

$$\begin{aligned}\rho_{\mathcal{H}} &= \max_{\alpha_X,\alpha_Y} \frac{\operatorname{cov}(\langle \Phi_X, \Phi_X^T \alpha_X \rangle, \langle \Phi_Y, \Phi_Y^T \alpha_Y \rangle)}{\sqrt{\operatorname{var}(\langle \Phi_X, \Phi_X^T \alpha_X \rangle)}\sqrt{\operatorname{var}(\langle \Phi_Y, \Phi_Y^T \alpha_Y \rangle)}} \\ &= \max_{\alpha_X,\alpha_Y} \frac{\operatorname{cov}(\mathbf{K}_X \alpha_X, \mathbf{K}_Y \alpha_Y)}{\sqrt{\operatorname{var}(\mathbf{K}_X \alpha_X)}\sqrt{\operatorname{var}(\mathbf{K}_Y \alpha_Y)}}.\end{aligned} \tag{3.10}$$

Following the same intuition as discussed in Section 2.1 we impose the constraints

$$\alpha_X^T \mathbf{K}_X^2 \alpha_X = 1,$$

$$\alpha_Y^T \mathbf{K}_Y^2 \alpha_Y = 1.$$

The optimization problem thus becomes

$$\begin{aligned}&\rho_{\mathcal{H}} = \max_{\alpha_X,\alpha_Y} \operatorname{corr}(\langle \mathbf{K}_X, \alpha_X \rangle, \langle \mathbf{K}_Y, \alpha_Y \rangle) = \alpha_X^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y, \\ &\text{subject to,} \\ &\alpha_X^T \mathbf{K}_X^2 \alpha_X = \alpha_Y^T \mathbf{K}_Y^2 \alpha_Y = 1.\end{aligned} \tag{3.11}$$

The corresponding Lagrangian of (3.11) is

$$L(\rho_X, \rho_Y, \alpha_X, \alpha_Y) = \alpha_X^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y - \frac{\rho_X}{2}(\alpha_X^T \mathbf{K}_X^2 \alpha_X - 1) - \frac{\rho_Y}{2}(\alpha_Y^T \mathbf{K}_Y^2 \alpha_Y - 1).$$

Taking the derivatives with respect to α_X and α_Y gives us

$$\frac{\partial L}{\partial \alpha_X} = \mathbf{K}_X \mathbf{K}_Y \alpha_Y - \rho_X \mathbf{K}_X^2 \alpha_X = 0, \tag{3.12}$$

$$\frac{\partial L}{\partial \alpha_Y} = \mathbf{K}_Y \mathbf{K}_X \alpha_X - \rho_Y \mathbf{K}_Y^2 \alpha_Y = 0. \tag{3.13}$$

Multiplying (3.12) by α_X^T and (3.13) by α_Y^T and subtracting the two gives us,

$$\begin{aligned} 0 &= \alpha_X^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y - \rho_X \alpha_X^T \mathbf{K}_X^2 \alpha_X - \alpha_Y^T \mathbf{K}_Y \mathbf{K}_X \alpha_X + \rho_Y \alpha_Y^T \mathbf{K}_Y^2 \alpha_Y \\ &= \rho_Y \alpha_Y^T \mathbf{K}_Y^2 \alpha_Y - \rho_X \alpha_X^T \mathbf{K}_X^2 \alpha_X. \end{aligned}$$

Using the constraints in (3.11) we then have that $\rho_X = \rho_Y$. Setting $\rho_X = \rho_Y = \rho_{\mathcal{H}}$ and assuming that the matrices $\mathbf{K}_X$ and $\mathbf{K}_Y$ are invertible, we have

$$\begin{aligned} \alpha_X &= \frac{\mathbf{K}_X^{-1} \mathbf{K}_X^{-1} \mathbf{K}_X \mathbf{K}_Y \alpha_Y}{\rho_{\mathcal{H}}} \\ &= \frac{\mathbf{K}_X^{-1} \mathbf{K}_Y \alpha_Y}{\rho_{\mathcal{H}}}. \end{aligned} \tag{3.14}$$

Similarly,

$$\alpha_Y = \frac{\mathbf{K}_Y^{-1} \mathbf{K}_X \alpha_X}{\rho_{\mathcal{H}}}. \tag{3.15}$$

Next substituting (3.15) into (3.12) gives us

$$\mathbf{K}_X \mathbf{K}_Y \mathbf{K}_Y^{-1} \mathbf{K}_X \alpha_X - \rho_{\mathcal{H}}^2 \mathbf{K}_X \mathbf{K}_X \alpha_X = 0,$$

leading the generalized eigen problem

$$\mathbf{I}\alpha_X = \rho^2_{\mathcal{H}}\alpha_X. \tag{3.16}$$

From the solution in (3.16) it can be seen that the eigenvalues for each of the corresponding eigenvectors will be equal to 1. Furthermore the corresponding eigenvectors will be equal to the unit vector $\mathbf{e}_i$ for α^i_X and will be equal to $\frac{1}{\rho_{\mathcal{H}}}\mathbf{K}_Y^{-1}\mathbf{K}_X\mathbf{e}_i$ for α^i_Y, $i = 1, \ldots, n$. This will be true so long as the kernel matrices $\mathbf{K}_X$ and $\mathbf{K}_Y$ are invertible.

As was the case with linear CCA we need to control how flexible we allow the directions to be. In the following section we discuss a regularized variant of KCCA which allows us to find non-trivial directions and relationships between spaces.

3.4 Regularized KCCA

Two standard regularization techniques used with KCCA are

$$\begin{aligned} \alpha_X^T\mathbf{K}_X^2\alpha_X + \kappa_X\alpha_X^T\alpha_X = 1, \\ \alpha_Y^T\mathbf{K}_Y^2\alpha_Y + \kappa_Y\alpha_Y^T\alpha_Y = 1, \end{aligned} \tag{3.17}$$

discussed in Kuss and Graepel (2003), and

$$\begin{aligned} \alpha_X^T\mathbf{K}_X^2\alpha_X + \alpha_X^T\mathbf{K}_X\alpha_X = 1, \\ \alpha_Y^T\mathbf{K}_Y^2\alpha_Y + \alpha_Y^T\mathbf{K}_Y\alpha_Y = 1. \end{aligned} \tag{3.18}$$

discussed in Hardoon *et al.* (2004). We focus on (3.17) since its behavior, generally speaking, is similar to (3.18), but looks, and as a result has a more intuitively appealing

connection to standard CCA. The optimization problem in (3.11) is rewritten as,

$$\rho_{\mathcal{H}} = \max_{\alpha_X, \alpha_Y} \text{corr}(\langle \mathbf{K}_X, \alpha_X \rangle, \langle \mathbf{K}_Y, \alpha_Y \rangle) = \alpha_X^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y,$$
$$\text{subject to} \tag{3.19}$$
$$\alpha_X^T \mathbf{K}_X^2 \alpha_X + \kappa_X \alpha_X^T \alpha_X = \alpha_Y^T \mathbf{K}_Y^2 \alpha_Y + \kappa_Y \alpha_Y^T \alpha_Y = 1.$$

The corresponding Lagrangian is

$$L(\rho_X, \rho_Y, \alpha_X, \alpha_Y) = \alpha_X^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y - \frac{\rho_X}{2}(\alpha_X^T \mathbf{K}_X^2 \alpha_X + \kappa_X \alpha_X^T \alpha_X - 1) - \frac{\rho_Y}{2}(\alpha_Y^T \mathbf{K}_Y^2 \alpha_Y \kappa_Y \alpha_Y^T \alpha_Y - 1).$$

Taking the derivative with respect to α_X and α_Y and setting equal to zero we have

$$\frac{\partial L}{\partial \alpha_X} = \mathbf{K}_X \mathbf{K}_Y \alpha_Y - \rho_X(\mathbf{K}_X^2 \alpha_X + \kappa_X \alpha_X) = 0, \tag{3.20}$$

$$\frac{\partial L}{\partial \alpha_Y} = \mathbf{K}_Y \mathbf{K}_X \alpha_X - \rho_Y(\mathbf{K}_Y^2 \alpha_Y + \kappa_Y \alpha_Y) = 0. \tag{3.21}$$

Multiplying (3.20) by α_X^T and (3.21) by α_Y^T and subtracting the two gives us,

$$0 = \alpha_X^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y - \rho_X \alpha_X^T (\mathbf{K}_X^2 + \kappa_X \mathbf{I}) \alpha_X - \alpha_Y^T \mathbf{K}_Y \mathbf{K}_X \alpha_X + \rho_Y \alpha_Y^T (\mathbf{K}_Y^2 + \kappa_Y \mathbf{I}) \alpha_Y +$$
$$= \rho_Y - \rho_X,$$

where the last equality holds by the constraints in (3.19). We then have that $\rho_X = \rho_Y$. Setting $\rho_X = \rho_Y = \rho_{\mathcal{H}}$ and assuming that the matrices $\mathbf{K}_X^2 + \kappa_X \mathbf{I}$ and $\mathbf{K}_Y^2 + \kappa_Y \mathbf{I}$ are invertible, we find

$$\alpha_X = \frac{(\mathbf{K}_X^2 + \kappa_X \mathbf{I})^{-1} \mathbf{K}_X \mathbf{K}_Y \alpha_Y}{\rho_{\mathcal{H}}}. \tag{3.22}$$

Similarly we have,

$$\alpha_Y = \frac{(\mathbf{K}_Y^2 + \kappa_Y \mathbf{I})^{-1} \mathbf{K}_Y \mathbf{K}_X \alpha_X}{\rho_{\mathcal{H}}}. \tag{3.23}$$

Next substituting (3.23) into (3.20) gives us the generalized eigen problem

$$\mathbf{K}_X\mathbf{K}_Y(\mathbf{K}_Y^2+\kappa_Y\mathbf{I})^{-1}\mathbf{K}_Y\mathbf{K}_X\alpha_X=\rho_{\mathcal{H}}^2(\mathbf{K}_X^2+\kappa_X\mathbf{I})\alpha_X.$$

This can also be expressed as

$$\begin{aligned}&\begin{pmatrix}\mathbf{0} & \mathbf{K}_X\mathbf{K}_Y\\ \mathbf{K}_Y\mathbf{K}_X & \mathbf{0}\end{pmatrix}\begin{pmatrix}\alpha_X\\ \alpha_Y\end{pmatrix}\\ &=\rho_{\mathcal{H}}\begin{pmatrix}\mathbf{K}_X^2+\kappa_X\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{K}_Y^2+\kappa_Y\mathbf{I}\end{pmatrix}\begin{pmatrix}\alpha_X\\ \alpha_Y\end{pmatrix}.\end{aligned}\tag{3.24}$$

In a similar fashion to linear CCA subsequent canonical correlations and vectors are found by solving for the remaining eigenvalue eigenvector pairs of the generalized eigenvalue problem in (3.24).

3.5 A Simultaneous Formulation of KCCA

An alternative formulation of the KCCA problem which will be of use later in Chapter 4 combines the successive subproblems described previously in Section 3.4 into one problem. The formulation of the simultaneous optimization problem is

$$\rho_{\mathcal{H}}=\arg\max_{(\alpha_X^1,\alpha_Y^1),\ldots,(\alpha_Y^n,\alpha_Y^n)}\sum_{i=1}^{n}(\alpha_X^i)^T\mathbf{K}_X\mathbf{K}_Y\alpha_Y^i$$

subject to,

$$(\alpha_X^i)^T(\mathbf{K}_X+\kappa\mathbf{I}_n)\alpha_X^j=\begin{cases}1 & \text{if } i\neq j,\\ 0 & \text{otherwise,}\end{cases}$$

$$(\alpha_Y^i)^T(\mathbf{K}_Y+\kappa\mathbf{I}_n)\alpha_Y^j=\begin{cases}1 & \text{if } i\neq j,\\ 0 & \text{otherwise.}\end{cases}$$

$$(\alpha_X^i)^T\mathbf{K}_X\mathbf{K}_Y\alpha_Y^j=0,\forall i\neq j,\tag{3.25}$$

$$i, j = 1, \ldots, n. \tag{3.26}$$

The corresponding Lagrangian can be written as

$$\begin{aligned} &L((\alpha_X^1, \alpha_Y^1), \ldots, (\alpha_X^n, \alpha_Y^n), \{\rho_X^{ij}\}_{i,j=1}^n, \{\rho_X^{ij}\}_{i,j=1}^n) \\ &= \sum_{i=1}^{n} (\alpha_X^i)^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y^i - \frac{1}{2} \sum_{i,j=1}^{n} \rho_X^{ij} (\alpha_X^i)^T (\mathbf{K}_X^2 + \kappa \mathbf{I}_n) \alpha_X^j - \frac{1}{2} \sum_{i,j=1}^{n} \rho_Y^{ij} (\alpha_Y^i)^T (\mathbf{K}_Y^2 + \kappa \mathbf{I}_n) \alpha_Y^j, \end{aligned} \tag{3.27}$$

where $\{\rho_X^{ij}\}_{i,j=1}^n$ and $\{\rho_X^{ij}\}_{i,j=1}^n$ are Lagrange multipliers.

Theorem 3.5.1. *The optimization problem in (3.25) can be restated as*

$$\rho_{\mathcal{H}} = arg \max_{\mathbf{A}_X, \mathbf{A}_Y} Tr(\mathbf{A}_X^T \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y)$$

subject to,

$$\begin{aligned} &\mathbf{A}_X^T (\mathbf{K}_X^2 + \kappa \mathbf{I}_n) \mathbf{A}_X = \mathbf{I}_n \\ &\mathbf{A}_T^T (\mathbf{K}_Y^2 + \kappa \mathbf{I}_n) \mathbf{A}_Y = \mathbf{I}_n \\ &(\alpha_X^i)^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y^j = 0, \forall i \neq j, \end{aligned} \tag{3.28}$$

where Tr denotes the matrix trace and $\mathbf{A}_X = (\alpha_X^1, \ldots, \alpha_X^n)$ *and* $\mathbf{A}_Y = (\alpha_Y^1, \ldots, \alpha_Y^n)$ *are the* $n \times n$ *matrices of canonical vectors.*

Proof. Let $\mathbf{R}_X = \{\rho_X^{ij}\}_{i,j=1}^n$ and $\mathbf{R}_Y = \{\rho_Y^{ij}\}_{i,j=1}^n$ be the $n \times n$ matrices of Lagrange multipliers, note that these matrices are symmetric. The Lagrangian in (3.27) can be written as

$$\begin{aligned} &L(\mathbf{A}_X, \mathbf{A}_Y, \mathbf{R}_X, \mathbf{R}_Y) \\ &= \mathrm{Tr}(\mathbf{A}_X^T \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y) - \frac{1}{2} \mathrm{Tr}(\mathbf{A}_X^T (\mathbf{K}_X^2 + \kappa \mathbf{I}_n) \mathbf{A}_X \mathbf{R}_X) - \frac{1}{2} \mathrm{Tr}(\mathbf{A}_Y^T (\mathbf{K}_Y^2 + \kappa \mathbf{I}_n) \mathbf{A}_Y \mathbf{R}_Y). \end{aligned} \tag{3.29}$$

In solving the Lagrangian in (3.29) we use the following identities related to the derivatives of the trace function

$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}^T \mathbf{A}) = \mathbf{A}, \tag{3.30}$$

and

$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}^T \mathbf{B} \mathbf{X} \mathbf{C}) = \mathbf{B}\mathbf{X}\mathbf{C} + \mathbf{B}^T \mathbf{X} \mathbf{C}^T. \tag{3.31}$$

Taking the derivative of (3.29) with respect to $\mathbf{A}_X$ and $\mathbf{A}_Y$ and setting equal zero gives us

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{A}_X} &= \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y - (\mathbf{K}_X^2 + \kappa \mathbf{I}_n) \mathbf{A}_X \mathbf{R}_X = \mathbf{0}, \\ \frac{\partial L}{\partial \mathbf{A}_Y} &= \mathbf{K}_Y \mathbf{K}_X \mathbf{A}_X - (\mathbf{K}_Y^2 + \kappa \mathbf{I}_n) \mathbf{A}_Y \mathbf{R}_Y = \mathbf{0}. \end{aligned} \tag{3.32}$$

Multiplying these by $\mathbf{A}_X^T$ and $\mathbf{A}_Y^T$ respectively, using the constraints in (3.28) and rearranging terms gives us

$$\begin{aligned} \mathbf{R}_X &= \mathbf{A}_X^T \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y, \\ \mathbf{R}_Y &= \mathbf{A}_Y^T \mathbf{K}_Y \mathbf{K}_X \mathbf{A}_X. \end{aligned}$$

But note that

$$\mathbf{A}_X^T \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y = \mathbf{A}_Y^T \mathbf{K}_Y \mathbf{K}_X \mathbf{A}_X = \mathrm{diag}\left(\{ (\alpha_X^i)^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y^i \}_{i=1}^n \right),$$

therefore

$$\mathbf{R}_X = \mathbf{R}_Y = \mathbf{R} = \begin{pmatrix} (\alpha_X^1)^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y^1 & 0 & \cdots & 0 \\ 0 & (\alpha_X^2)^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (\alpha_X^n)^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y^n \end{pmatrix}$$

$$= \begin{pmatrix} \rho^1_{\mathcal{H}} & 0 & \cdots & 0 \\ 0 & \rho^2_{\mathcal{H}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho^n_{\mathcal{H}} \end{pmatrix}.$$

Next solving for $\mathbf{A}_X$ and $\mathbf{A}_Y$ above we have

$$\begin{aligned} \mathbf{A}_X &= (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y \mathbf{R}^{-1}, \\ \mathbf{A}_Y &= (\mathbf{K}_Y^2 + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_Y \mathbf{K}_X \mathbf{A}_X \mathbf{R}^{-1}, \end{aligned} \tag{3.33}$$

which are the same as the solutions we found for α_X^i and α_Y^i in Section 3.4. Plugging in the solution for $\mathbf{A}_Y$ into the first equation in (3.32) and rearranging terms gives us

$$\mathbf{K}_X \mathbf{K}_Y (\mathbf{K}_Y^2 + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_Y \mathbf{K}_X \mathbf{A}_X = (\mathbf{K}_X^2 + \kappa \mathbf{I}_n) \mathbf{A}_X \mathbf{R}^2.$$

Let $\mathbf{B}_X = (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{\frac{1}{2}} \mathbf{A}_X$ then

$$\mathbf{K}_X \mathbf{K}_Y (\mathbf{K}_Y^2 + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_Y \mathbf{K}_X (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{\frac{1}{2}} (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{A}_X = (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{\frac{1}{2}} (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{\frac{1}{2}} \mathbf{A}_X \mathbf{R}^2.$$

Rearranging terms gives us

$$(\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{K}_X \mathbf{K}_Y (\mathbf{K}_Y^2 + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_Y \mathbf{K}_X (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{B}_X = \mathbf{B}_X \mathbf{R}^2. \tag{3.34}$$

Let $\mathbf{M}_{XY} = (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{K}_X \mathbf{K}_Y (\mathbf{K}_Y^2 + \kappa \mathbf{I}_n)^{-\frac{1}{2}}$. Suppose $\mathbf{B}_X$ are the eigenvectors of the matrix $\mathbf{M}_{XY} \mathbf{M}_{XY}^T$ and $\mathbf{\Lambda}_X$ the corresponding eigenvalues, then

$$\mathbf{M}_{XY} \mathbf{M}_{XY}^T = \mathbf{B}_X \mathbf{\Lambda}_X \mathbf{B}_X^T,$$

Plugging this into (3.34) we have

$$\begin{aligned}\mathbf{M}_{XY}\mathbf{M}_{XY}^T\mathbf{B}_X &= \mathbf{B}_X\mathbf{\Lambda}_X\mathbf{B}_X^T\mathbf{B}_X \\ &= \mathbf{B}_X\mathbf{\Lambda}_X \\ &= \mathbf{B}_X\mathbf{R}^2.\end{aligned}$$

Left multiplying both sides by $\mathbf{B}_X^T$ shows us that $\mathbf{\Lambda}_X = \mathbf{R}^2$. From this we can see that the matrices $\mathbf{R}$ and $\mathbf{B}_X$ must be the singular values and left singular vectors of $\mathbf{M}_{XY}$. Similar calculations show us that $\mathbf{B}_Y$ are the right singular vectors of $\mathbf{M}_{XY}$. This is in agreement with our calculations from Section 3.4. □

3.6 Kernel Centering

In order to maintain our understanding of KCCA as maximizing correlation in feature space we need to ensure that the data is *centered* in feature space. The following calculation shows how this can be done. Let $\bar{\Phi} = \frac{1}{n}\mathbf{J}\Phi$ where $\mathbf{J}$ is an $n \times n$ matrix of ones, then

$$\begin{aligned}(\Phi - \bar{\Phi})(\Phi - \bar{\Phi})^T &= \Phi\Phi^T - \Phi\bar{\Phi}^T - \bar{\Phi}\Phi^T + \bar{\Phi}\bar{\Phi}^T \\ &= \mathbf{K} - \frac{1}{n}\mathbf{J}\mathbf{K} - \frac{1}{n}\mathbf{K}\mathbf{J} + \frac{1}{n^2}\mathbf{J}\mathbf{K}\mathbf{J} \\ &= \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{K}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) \qquad (3.35)\end{aligned}$$

Unless stated otherwise we assume throughout that the kernel matrices are centered.

3.7 Toy Example: Non-standard data

We saw in Section 3.1 that KCCA was able to overcome some of the obstacles encountered by standard CCA. Where KCCA begins to encounter problems is when the distribution of points within a space is non-standard and/or heterogeneous. To illus-

trate this consider the example shown in Figure 3.7, as with the protein-ligand matching problem there is a one-to-one correspondence between points in the two spaces.

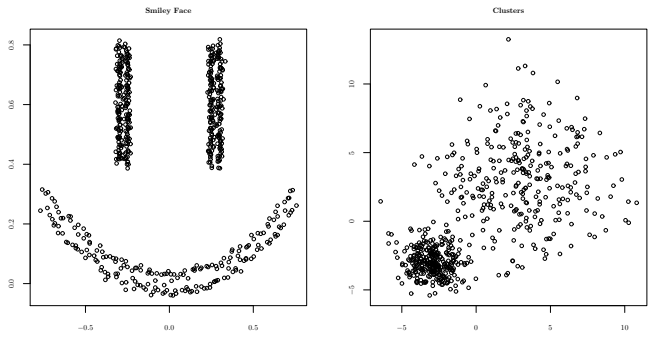

Figure 3.7: *A toy example illustrating the cases when the distribution of points within a space is non-standard and heterogeneous.*

The underlying structure between these spaces is illustrated in Figure 3.8. The top row of plots tells us about how the distribution of points on the right (cluster space) relates to the distribution of points on the left (smiley face space). The bottom set of plots tells us about how the distribution of points on the left is related to distribution of points on the right.

If we were to look at the two spaces as marginal distributions, there is a distinct impression of the three clusters in the left, and two in the right. However, the joint distribution has six distinct groups. Looking at the plots on the left in Figure 3.8 each of the three clusters is in fact composed of two subclusters. Likewise each of the two clusters in the plots on the right are composed of three subclusters. Ideally the projections onto the KCCA directions would identify each of these six groups, shown in Figure 3.9.

Using an RBF kernel with $\sigma = 1/2$ we look at the first 5 canonical directions. Ideally what we would see is a separation of each of the groups as well as a strong alignment between each of the spaces. What we find looking at Figure 3.10, a scatter plot matrix of the first five canonical directions, is that while the leading correlations are large (0.98, 0.97, 0.95, 0.80, 0.75), we are not able to find the structure in the data we were looking for, i.e. separating out the six groups (with each of the colors corresponding to one of

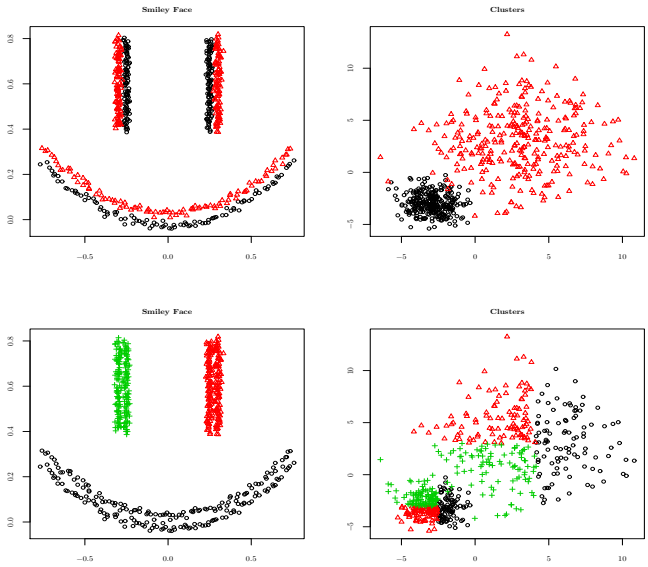

Figure 3.8: *These plots highlight how the distribution of points in one space is related to the distribution of points in the other. Looking at the plots on the left in Figure 3.8 each of the three clusters is in fact composed of two subclusters. Likewise each of the two clusters in the plots on the right are composed of three subclusters.*

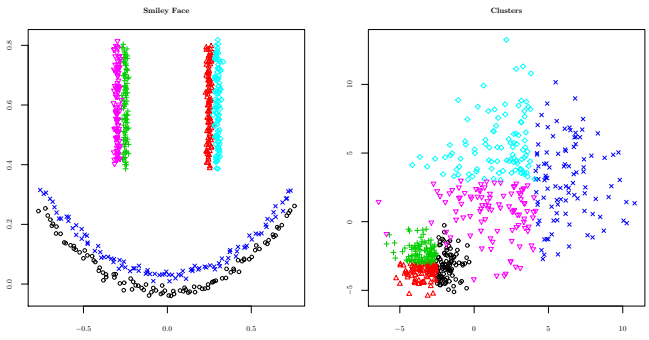

Figure 3.9: *In this plot each of the six underlying subgroups shown in Figure 3.8 is highlighted.*

the six groups). Note that only the projections in the smiley face space are shown since the cluster space projections look essentially the same.

In the context of the protein-ligand matching problem this type of situation presents a potential problem. Suppose a new point, say in the space with the smiley face, is projected into KCCA space. As can be seen in Figure 3.10 there is a great deal of overlap between each of the six subgroups in the projected space. In particular note that

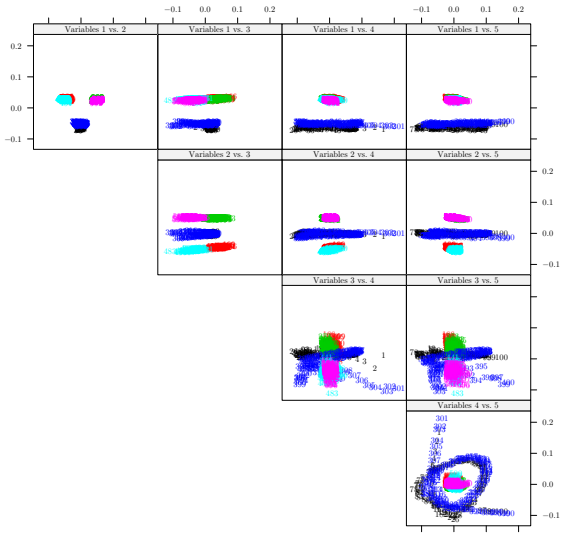

Figure 3.10: *Scatterplot matrix of the first five KCCA direction vectors for the data shown in Figure 3.7. Each of the colors in this plot corresponds to one of the six underlying subpopulation in the data (see Figure 3.8 for details).*

each of the overlapped groups is composed of, respectively, the left eye, right eye and mouth. The reason this type of behavior presents a problem is that each of the eyes and the mouth are actually composed of two different subpopulations where each of the populations correspond to very different groups in the space with the two clusters. So while we may be able to accurately predict the location of a new point in KCCA space the interpretation of its surrounding neighbors may not be so meaningful.

3.8 KCCA Performance on Real Data

As in Section 2.6 we apply the methods described in this chapter on the RLP800 and WDI data sets described in Section 1.3. The kernel used in our analysis is the radial basis function (RBF)

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{1}{2\sigma^2}||\mathbf{x}_i - \mathbf{x}_j||^2\right\}. \tag{3.36}$$

Regularized KCCA was used with tuning parameters selected via a cross validation

scheme similar to that described in Section 2.6, the difference being the addition of the bandwidth parameter σ, whose with candidate values are $\{0.5, 1, 2, 5, 10\}$. The resulting set of parameter values were $\sigma = 2$, $\kappa_X = \kappa_Y = 0.01$, the number of dimensions projected onto was $p_X = p_Y = 400$ and the number of neighbors used in the prediction was 60.

Figure 3.11 shows the distribution of each of the first three kernel canonical variates (left) as well as the canonical correlations for each of the 400 variates (right). As can be seen the leading canonical correlations are fairly large indicating that a strong relationship exists between spaces.

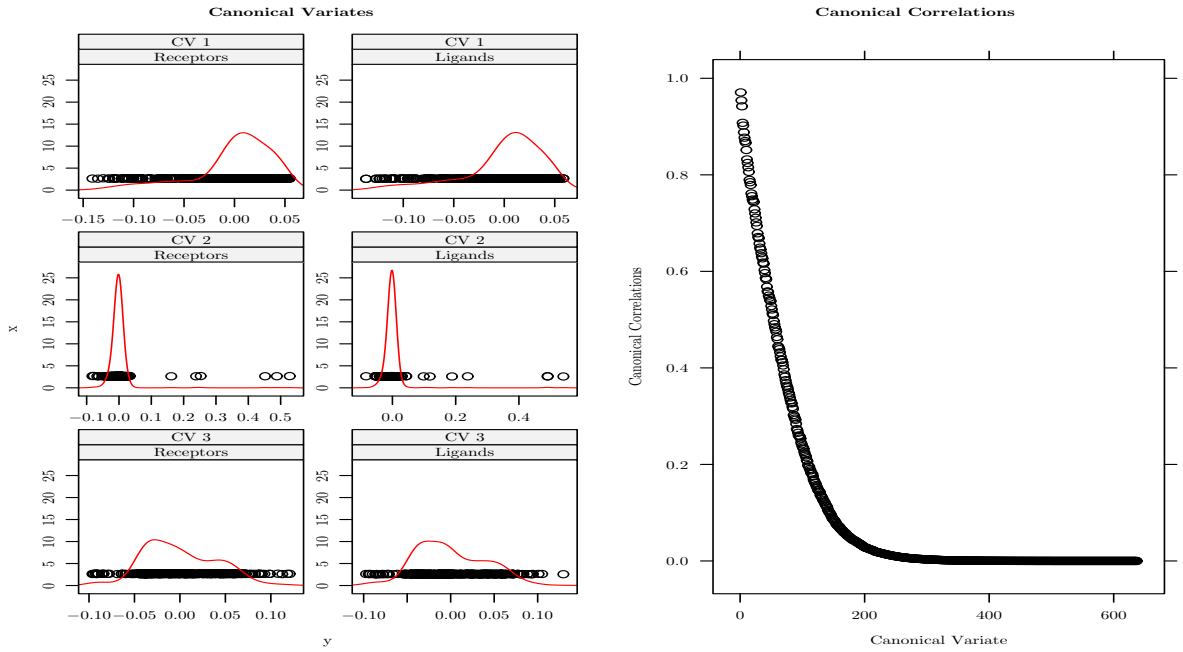

Figure 3.11: *On the left are plots of the first three canonical variates in protein and ligand space respectively. The red curves are the associated density estimates of the canonical variates. This is meant to provide some insight into the distribution of the data within a space as well as how well aligned points are between spaces. On the right is a plot of the canonical correlations associated with each of the 637 canonical vectors.*

Figure 3.12 is a scatterplot matrix showing the projections of the training (black) and testing (red) points from the RLP800 data onto the first three canonical vectors. From a visual assessment of the data it appears as though both the training and testing points in each of the two spaces is fairly well aligned.

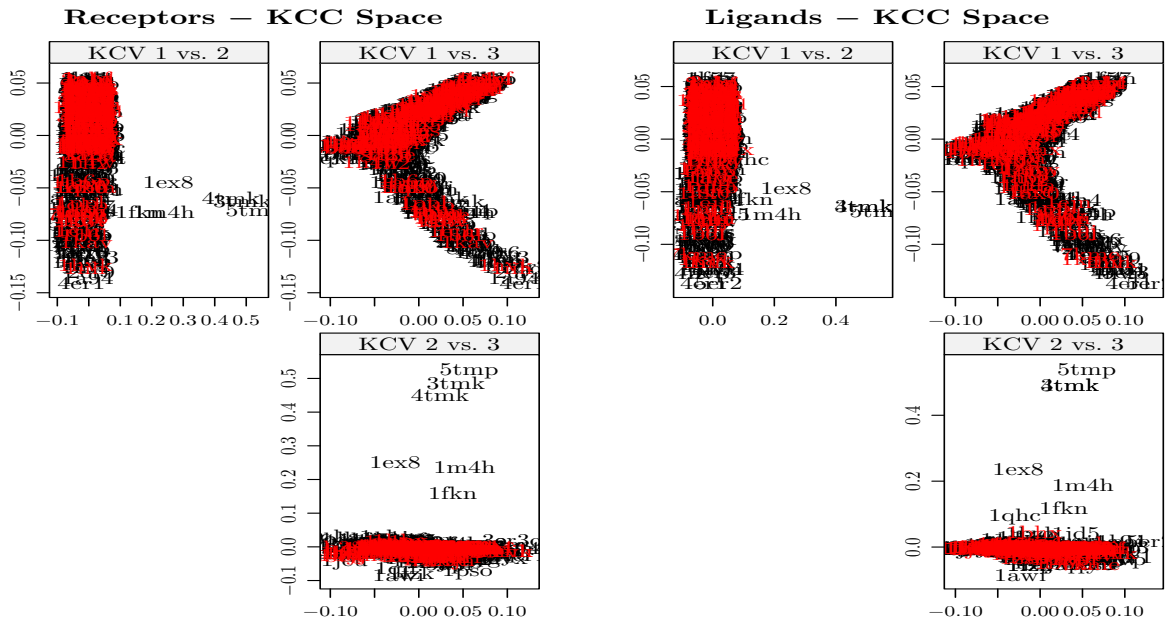

Figure 3.12: *A plot of the kernel canonical correlations from the RLP800 data set with the training data shown in black and the test data shown in red. From a visual assessment of the data it appears as though the two spaces are fairly well aligned.*

Figure 3.13 is similar to Figure 3.12 but with one test point highlighted (red), in both protein and ligand space, and its predicted value (green), in ligand space only. As can be seen the prediction is fairly accurate.

Figure 3.14 is a histogram of the ranks associated with our prediction using regularized KCCA. The average rank in this case was approximately 7.1, indicated by the vertical red line. For the purposes of comparison the average rank from linear RCCA (approximately 10) and the previous approach used in Oloff *et al.* (2006) (approximately 18.1) is also shown (blue and green respectively).

Figure 3.15 similarly summarizes the results from prediction on the WDI data set. The mean predicted rank using KCCA is approximately 56 (red line), using CCA it is approximately 67 (blue line) and using the previous method is approximately 310 (green line).

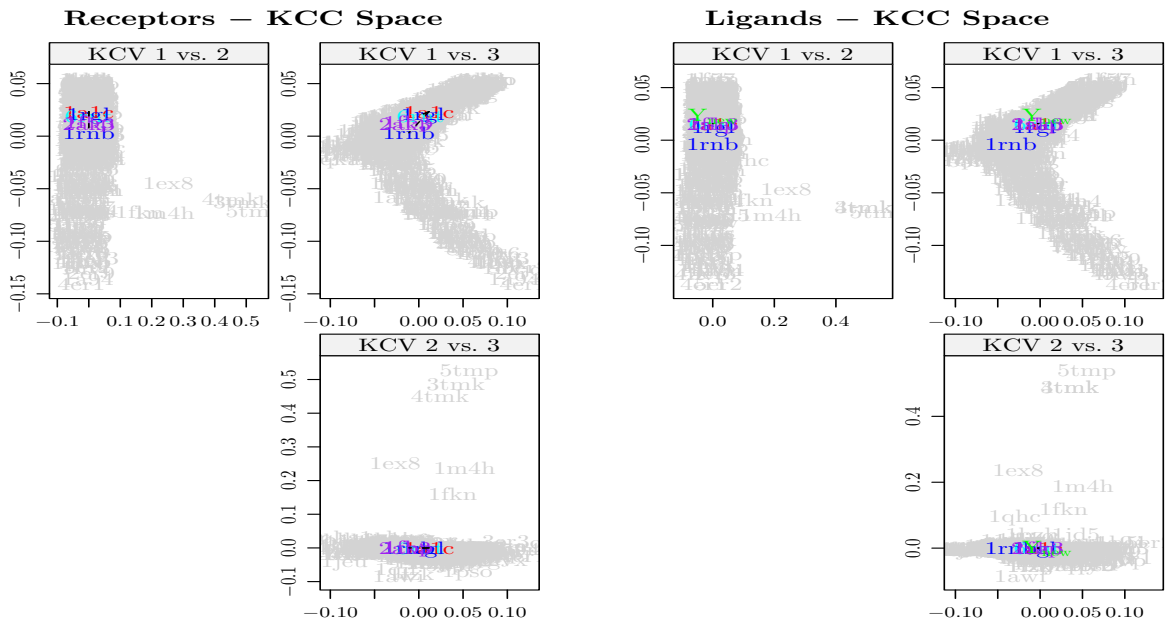

Figure 3.13: *Similar to Figure 3.12 but with one of the test points highlighted and only its three nearest neighbors. The color scheme is similar to that of the previous toy examples discussed in the linear case.*

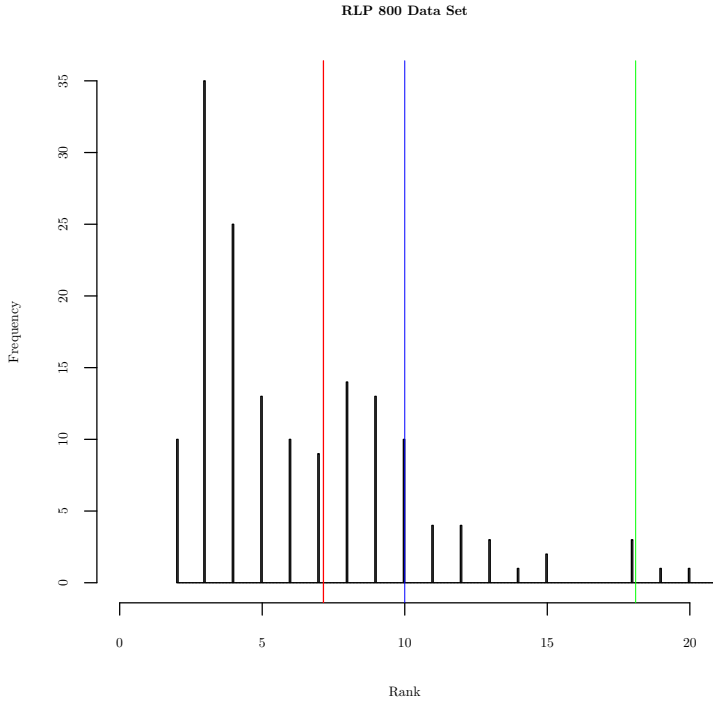

Figure 3.14: *A histogram showing the large improvement in rank resulting from KCCA prediction on the test data from the RLP800 dataset. The vertical red line indicates the average rank (approximately 7.1) using KCCA, the blue line shows the average rank using CCA (approximately 10) and the vertical green line the method implemented in Oloff* et al. *(2006) (approximately 18.1).*

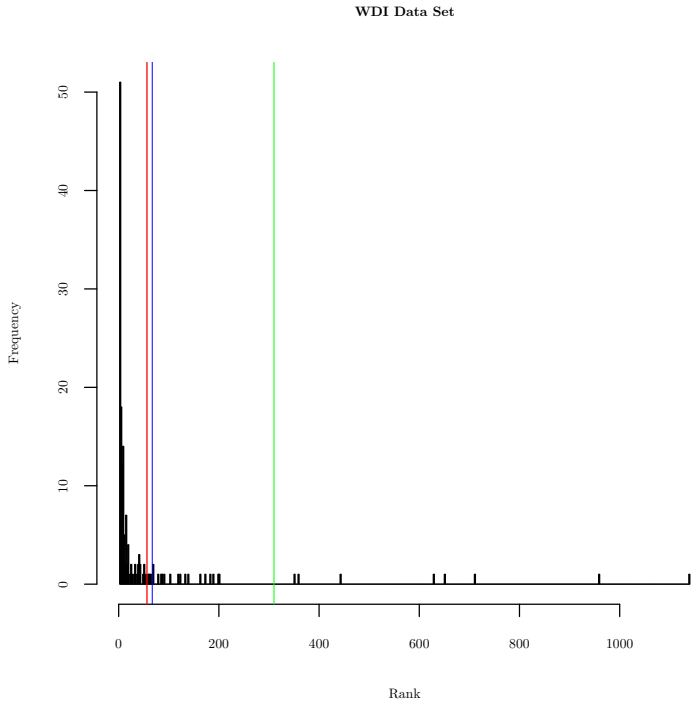

Figure 3.15: *Similar to the histogram above but using the WDI data. The mean rank using KCCA is approximately 56, RCCA is approximately 67 and the previous method is approximately 310.*

CHAPTER 4

Indefinite KCCA

A potential shortcoming of standard KCCA, that was illustrated in the example presented in Figure 3.7, is that standard positive definite kernels can be limited in their ability to capture non-standard heterogeneous behavior in the data. A general class of kernels which is better suited to handle this type of behavior takes the form

$$K(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} w(\mathbf{x}_i, \mathbf{x}_j) & \text{if } \mathbf{x}_j \in N(\mathbf{x}_i), \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

Here $N(\mathbf{x})$ denotes some neighborhood of the observation $\mathbf{x}$, such as a $k(\in \mathbb{Z}_+)$ or $\epsilon(> 0)$-neighborhood. Kernels of this form restrict attention to the local structure of the data and allow for a flexible definition of similarity. The problem encountered with this class of kernels is that they are frequently indefinite (see the discussion following Definition 4.1.1). Recalling our discussion from Section 3.2 one of the requirements on the function K is that it should be positive semi-definite. As a result of the indefiniteness many of the properties and optimality guarantees no longer hold.

Indefinite kernels have recently gained increased interest (Ong *et al.* (2004), Haasdonk (2005), Chen and Ye (2008), Luss and d'Aspremont (2008)), where rather than defining K to be a function defined in a RKHS K is defined in an space characterized by an *indefinite inner product* called a *Krein* space. In Section 4.1 we provide an overview of some of the definitions and theoretical results about Krein spaces (following the discussion of Ong

et al. (2004)). In Section 4.2 we formulate the IKCCA problem. In Section 4.3 we provide an overview of spectral clustering and in Section 4.4 we show a connection between IKCCA and LDA when a variant of the Normalized Graph Laplacian (NGL) kernel is used. In Section 4.5 we apply IKCCA to the non-standard data example introduced in Section 3.7. Finally in Section 4.6 we apply IKCCA to the protein-ligand matching problem.

4.1 Krein Spaces

In this section we provide some definitions and theorems as they relate to Krein spaces and connect these ideas to the IKCCA problem (more details can be found in Ong *et al.* (2004)).

Definition 4.1.1. *(Inner Product) Let $\mathcal{K}$ be a vector space on the scalar field. An inner product $\langle .,. \rangle_{\mathcal{K}}$ on $\mathcal{K}$ is a bilinear form where for all $f, g, h \in \mathcal{K}$, $\alpha \in \mathbb{R}$*

$$\langle f, g \rangle_{\mathcal{K}} = \langle g, f \rangle_{\mathcal{K}}$$

$$\langle \alpha f + g, h \rangle_{\mathcal{K}} = \alpha \langle f, h \rangle_{\mathcal{K}} + \langle g, g \rangle_{\mathcal{K}}$$

$$\langle f, g \rangle_{\mathcal{K}} = 0 \text{ for all } g \in \mathcal{K} \text{ implies } \Rightarrow f = 0.$$

The importance of $\mathcal{K}$ being a vector space on a *scalar field* is that it allows for a flexible definition of an inner product (i.e. the scalar in one of the dimensions could be complex or negative as we will see below). An inner product is said to be *positive* if for all $f \in \mathcal{K}$, $\langle f, f \rangle_{\mathcal{K}} \geq 0$. It is called a *negative* inner product, if for all $f \in \mathcal{K}$, $\langle f, f \rangle_{\mathcal{K}} \leq 0$. An inner product is called indefinite if it is neither strictly positive nor strictly negative.

Definition 4.1.2. *(Krein Space) An inner product space $(\mathcal{K}, \langle .,. \rangle_{\mathcal{K}})$ is a Krein space if there exist two Hilbert spaces $\mathcal{H}_+$, $\mathcal{H}_-$ spanning $\mathcal{K}$ such that*

1. *All $f \in \mathcal{K}$ can be decomposed into $f = f_+ + f_-$, where $f_+ \in \mathcal{H}_+$ and $f_- \in \mathcal{H}_-$.*
2. *$\forall f, g \in \mathcal{K}$, $\langle f, g \rangle_{\mathcal{K}} = \langle f_+, g_+ \rangle_{\mathcal{H}_+} - \langle f_-, g_- \rangle_{\mathcal{H}_-}$*

Definition 4.1.3. *(Associated Hilbert Space) Let $\mathcal{K}$ be a Krein space with decomposition into Hilbert spaces $\mathcal{H}_+$ and $\mathcal{H}_-$. Then we denote by $\bar{\mathcal{K}}$ the associated Hilbert space defined by*

$$\bar{\mathcal{K}} = \mathcal{H}_+ \oplus \mathcal{H}_- \text{ hence } \langle f, g\rangle_{\bar{\mathcal{K}}} = \langle f_+, g_+\rangle_{\mathcal{H}_+} + \langle f_-, g_-\rangle_{\mathcal{H}_-}.$$

Likewise we can introduce the symbol $\ominus$ to indicate that

$$\mathcal{K} = \mathcal{H}_+ \ominus \mathcal{H}_- \text{ hence } \langle f, g\rangle_{\mathcal{K}} = \langle f_+, g_+\rangle_{\mathcal{H}_+} - \langle f_-, g_-\rangle_{\mathcal{H}_-}.$$

The strong topology on $\mathcal{K}$ is defined as the Hilbert topology of $\bar{K}$. The topology does not depend on the decomposition chosen. Clearly $|\langle f, f\rangle|_{\mathcal{K}} \leq ||f||_{\bar{\mathcal{K}}}$ for all $f \in \mathcal{K}$. Note that we only have equality when $\langle f_-, g_-\rangle_{\mathcal{H}_-} = 0$, this, however, does not imply that the inner product, i.e. the kernel, is positive.

Let $\mathcal{X}$ be a non-empty set from which the data, $\mathbf{x}$ is sampled. Assuming K is an indefinite kernel and $\mathcal{K} \subset \mathbb{R}^{\mathcal{X}} := \{f : \mathcal{X} \to \mathbb{R}\}$ we have

$$\begin{aligned} \Phi : &\mathcal{X} \to \mathbb{R}^{\mathcal{X}} \\ &\mathbf{x} \mapsto K(., \mathbf{x}) = f(\mathbf{x}). \end{aligned}$$

Definition 4.1.4. *(Reproducing Kernel Krein Space) Let $\mathcal{X}$ be a nonempty set, $\mathcal{H}_+$ and $\mathcal{H}_-$ are RKHS (with kernels K_+ and K_-) and $\mathcal{K} = \mathcal{H}_+ \ominus \mathcal{H}_-$ a Krein space of functions $f : \mathcal{K} \to \mathbb{R}$ endowed with its strong topology $\bar{\mathcal{K}}$. Then $\mathcal{K}$ is called a reproducing kernel Krein space (Alpay (2001), Chapter 7) endowed with an inner product $\langle ., .\rangle_{\mathcal{K}}$ if Φ is continuous on $\mathcal{K}$ and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties*

1. $$\langle f, K(\mathbf{x}, .)\rangle_{\mathcal{K}} = f(\mathbf{x}) \text{ for all } f \in \mathcal{K}.$$

   In particular

   $$\langle K(\mathbf{x}, .), K(\mathbf{x}', .)\rangle_{\mathcal{K}} = K(\mathbf{x}, \mathbf{x}').$$

2. $K = K_+ - K_-$.

To illustrate how Krein spaces and indefinite inner products arise in the context of our problem consider the following. Suppose we have a symmetric kernel function K which is indefinite. The implication of this is that the resulting kernel matrix $\mathbf{K} = \{K_{ij}\}_{i=1}^n$ is indefinite and that it therefore contains positive *and* negative eigenvalues. Let $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ be the eigendecomposition of $\mathbf{K}$, where $\mathbf{U}$ are the eigenvectors and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues starting with the p positive eigenvalues, followed by the q negative ones and the $n - p - q \geq 1$ eigenvalues equal to 0. To see how $\mathbf{K}$ can be interpreted as a matrix composed of inner products in this indefinite inner product space consider the following representation of its eigendecomposition

$$\mathbf{K} = \mathbf{U}|\mathbf{\Lambda}|^{\frac{1}{2}}\text{diag}(\mathbf{1}_p, -\mathbf{1}_q, \mathbf{0}_{n-p-q})|\mathbf{\Lambda}|^{\frac{1}{2}}\mathbf{U}^T.$$

Let $\mathbf{M} = \text{diag}(\mathbf{1}_p, -\mathbf{1}_q)$ and $\mathbf{\Phi}$ be equal to the first $p + q$ columns of $\mathbf{U}|\mathbf{\Lambda}|^{\frac{1}{2}}$. Define the i^{th} row of $\mathbf{\Phi}$ be equal to

$$\Phi_i = (\underbrace{\phi_{i,1}, \ldots, \phi_{i,p}}_{=\Phi_i^+}, \underbrace{\phi_{i,p+1}, \ldots, \phi_{i,p+q}}_{=\Phi_i^-}).$$

We then have a kernel matrix composed of elements

$$\begin{aligned} K_{ij} &= \Phi_i^T \mathbf{M} \Phi_j \\ &= (\Phi_i^+)^T \Phi_j^+ - (\Phi_i^-)^T \Phi_j^- \\ &= \langle \Phi_i, \Phi_j \rangle_{\mathcal{H}_+} - \langle \Phi_i, \Phi_j \rangle_{\mathcal{H}_-} \\ &= \langle \Phi_i, \Phi_j \rangle_{\mathcal{K}}. \end{aligned}$$

Many of the properties that hold for reproducing kernel Hilbert spaces also hold for reproducing kernel Krein spaces. The key difference is that rather than minimizing (max-

imizing) a regularized risk functional the problem becomes that of finding a stationary point of a similar risk functional. In the statement of the optimization problem in Theorem 4.1.5, when we write "stabilize" it is meant to emphasize the fact that the solutions we are finding are not necessarily global or local minima and maxima (the solution could be a saddle point), but are stationary points.

Theorem 4.1.5. *(Ong* et al. *(2004)) Let $\mathcal{K}$ be a RKKS with kernel K. Denote by $L\{f, \mathcal{X}\}$ a continuous convex loss functional depending on $f \in \mathcal{K}$ only via its evaluation $f(\mathbf{x}_i)$ with $\mathbf{x}_i \in \mathcal{X}$, let $\Omega(\langle f, f\rangle_{\mathcal{K}})$ be a continuous stabilizer with strictly monotonic $\Omega : \mathbb{R} \to \mathbb{R}$ and let $C\{f, \mathcal{X}\}$ be a continuous functional imposing a set of constraints on f, that is $C : \mathcal{K} \times \mathcal{X}^m \to \mathbb{R}^n$. Then if the optimization problem*

$$\begin{aligned} &\textit{stabilize } L\{f, \mathcal{X}\} + \Omega(\langle f, f\rangle_{\mathcal{K}}) \\ &\textit{subject to } C\{f, \mathcal{X}\} \leq d \end{aligned} \tag{4.2}$$

has a stationary point f^*, it admits the expansion

$$f^* = \sum_i \alpha_i K(\mathbf{x}_i, \cdot) \textit{ where } \mathbf{x}_i \in \mathcal{X} \textit{ and } \alpha_i \in \mathbb{R}. \tag{4.3}$$

4.2 IKCCA

The results of Section 4.1 provide some insight into the challenges that arise from dealing with indefinite kernels. In particular the results of Theorem 4.1.5 point to the fact that the solution that we find may not be globally, or even locally optimal (as it may be a saddle point). The "stabilization" problem stated in (4.2) of Theorem 4.1.5 motivated the form of the Indefinite KCCA (IKCCA) problem we present in this section. In particular, the addition of the stabilizing function, Ω on the indefinite inner product, $\langle f, f\rangle_{\mathcal{K}}$ led us (in addition to results and discussion from Luss and d'Aspremont (2008)) to consider introducing a constraint on the behavior on the indefinite kernels matrix

itself.

In the following let $||\cdot||_F$ denote the Frobenius norm. Define $\mathbf{M} \succeq 0$ to mean that the matrix $\mathbf{M}$ is positive semi-definite and let $\lambda_X, \lambda_Y \in \mathbb{R}^+ \cup \infty$ be tuning parameters (discussed in more detail later this section). Here $\mathbf{K}_X^0$ and $\mathbf{K}_Y^0$ are the (potentially) indefinite kernels and $\mathbf{K}_X$ and $\mathbf{K}_Y$ will be the positive semi-definite approximations of these kernels. With these notations in mind we now define the IKCCA optimization problem,

$$\rho_{\mathcal{H}} = \max_{\mathbf{A}_X, \mathbf{A}_Y} \min_{\mathbf{K}_X, \mathbf{K}_Y} \text{Tr}(\mathbf{A}_X^T \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y) + \lambda_X ||\mathbf{K}_X - \mathbf{K}_X^0||_F^2 + \lambda_Y ||\mathbf{K}_Y - \mathbf{K}_Y^0||_F^2,$$

subject to,

$$\begin{aligned}
&\mathbf{A}_X^T \mathbf{K}_X^2 \mathbf{A}_X + \kappa \mathbf{A}_X^T \mathbf{A}_X = \mathbf{I}_n, \\
&\mathbf{A}_Y^T \mathbf{K}_Y^2 \mathbf{A}_Y + \kappa \mathbf{A}_Y^T \mathbf{A}_Y = \mathbf{I}_n, \\
&(\alpha_X^i)^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y^j = 0, \text{ for } i \neq j, i, j = 1, \ldots, n, \\
&\mathbf{K}_X \succeq \mathbf{0}, \\
&\mathbf{K}_Y \succeq \mathbf{0},
\end{aligned} \tag{4.4}$$

where, $\mathbf{A}_X = (\alpha_X^1, \ldots, \alpha_X^n)$ and $\mathbf{A}_Y = (\alpha_Y^1, \ldots, \alpha_Y^n)$. Note that the this optimization problem and the KCCA optimization problem (see (3.28) in Section 3.5 for details) are only equivalent when the kernel matrices $\mathbf{K}_X^0$ and $\mathbf{K}_Y^0$ are positive semi-definite, as will be shown in the proof of Theorem 4.2.2.

Theorem 4.2.1. *Letting $\lambda_X, \lambda_Y \to \infty$, the optimization problem in (4.4) is concave in α_X^i and α_Y^i, $i = 1, \ldots, n$ and convex in $\mathbf{K}_X$ and $\mathbf{K}_Y$.*

Proof. We begin by showing that the loss function

$$\text{Tr}(\mathbf{A}_X^T \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y) \tag{4.5}$$

is concave in α_X^i and α_Y^i, $i = 1, \ldots, n$. Note that (4.5) can be expressed as

$$\mathrm{Tr}(\mathbf{A}_X^T \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y) = \sum_{i=1}^{n} \alpha_X^{iT} \mathbf{K}_X \mathbf{K}_Y \alpha_Y^i.$$

It can be seen from this representation that if $\alpha_X^{iT} \mathbf{K}_X \mathbf{K}_Y \alpha_Y^i$ is concave in α_X^i and α_Y^i for all $i = 1, \ldots, n$ then (4.5) will also be concave. For the remainder of the proof we suppress the superscript i.

Suppose that $\mathbf{K}_X, \mathbf{K}_Y \succeq 0$. Recall that the solution for α_X in the KCCA optimization problem in (3.19) is

$$\alpha_X = \frac{1}{\rho_{\mathcal{H}}} (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_X \mathbf{K}_Y \alpha_Y.$$

Plugging this in we have

$$\alpha_X^T \mathbf{K}_X \mathbf{K}_Y \alpha_Y = \frac{1}{\rho_{\mathcal{H}}} \alpha_Y^T \mathbf{K}_Y \mathbf{K}_X (\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_Y \alpha_Y.$$

Note that $(\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_X$ is symmetric, this can be seen by looking at its eigendecomposition

$$\begin{aligned}
&(\mathbf{K}_X^2 + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_X \\
&= \mathbf{V}_X \begin{pmatrix} \frac{1}{(\lambda_X^1)^2+\kappa} & 0 & \cdots & 0 \\ 0 & \frac{1}{(\lambda_X^2)^2+\kappa} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{(\lambda_X^n)^2+\kappa} \end{pmatrix} \mathbf{V}_X^T \mathbf{V}_X \begin{pmatrix} \lambda_X^1 & 0 & \cdots & 0 \\ 0 & \lambda_X^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_X^n \end{pmatrix} \mathbf{V}_X^T \\
&= \mathbf{V}_X \begin{pmatrix} \frac{\lambda_X^1}{(\lambda_X^1)^2+\kappa} & 0 & \cdots & 0 \\ 0 & \frac{\lambda_X^2}{(\lambda_X^2)^2+\kappa} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\lambda_X^n}{(\lambda_X^n)^2+\kappa} \end{pmatrix} \mathbf{V}_X^T,
\end{aligned}$$

where $\mathbf{V}_X$ are the eigenvectors and $\lambda_X^i, i = 1, \ldots, n$ are the eigenvalues of the matrix

$\mathbf{K}_X$.

Now, if the kernel matrices $\mathbf{K}_X$ and $\mathbf{K}_Y$ are positive definite then $\mathbf{K}_Y\mathbf{K}_X(\mathbf{K}_X^2 + \kappa\mathbf{I}_n)^{-1}\mathbf{K}_Y$ must be positive definite. To see this, let $\mathbf{c} \in \mathbb{R}^n$ be a vector of constants, then

$$\begin{aligned}
&\mathbf{c}^T\mathbf{K}_Y\mathbf{K}_X(\mathbf{K}_X^2 + \kappa\mathbf{I}_n)^{-1}\mathbf{K}_Y\mathbf{c} \\
&= (\mathbf{c}^*)^T\mathbf{K}_X(\mathbf{K}_X^2 + \kappa\mathbf{I}_n)^{-1}\mathbf{c}^* \\
&\geq 0,
\end{aligned}$$

where $\mathbf{c}^* = \mathbf{K}_Y\mathbf{c}$. The last inequality holds since $(\mathbf{K}_X^2 + \kappa\mathbf{I}_n)^{-1}\mathbf{K}_X$ is positive definite. Therefore, since the terms $\lambda_X||\mathbf{K}_X - \mathbf{K}_X^0||_F^2$ and $\lambda_Y||\mathbf{K}_Y - \mathbf{K}_Y^0||_F^2$ do not depend on $\mathbf{A}_X$ and $\mathbf{A}_Y$, as will be shown in Theorem 4.2.2, the IKCCA loss function in (4.5) is concave, as we wanted to show.

Using the fact that the square of the Frobenius norm is strictly convex (Boyd and Vandeberghe (2004)) we then have that the inner minim

Putting this all together we have that the IKCCA problem is concave in α_X^i and α_Y^i, $i = 1, \ldots, n$ and it is convex in $\mathbf{K}_X$ and $\mathbf{K}_Y$, as we wanted to show.

□

Let $(\mathbf{X})_+$ denote the positive part of the matrix $\mathbf{X}$, i.e. $(\mathbf{X})_+ = \sum_i \max(0, \lambda_i)\mathbf{v}_i\mathbf{v}_i^T$, where λ_i and $\mathbf{v}_i$ are i^{th} eigenvalue-eigenvector pair of the matrix $\mathbf{X}$. With this in mind we following state theorem,

Theorem 4.2.2. *Letting $\lambda_X, \lambda_Y \to \infty$, and given the optimization problem in (4.4) the optimal values for $\mathbf{K}_X$ and $\mathbf{K}_Y$ are given by*

$$\begin{aligned}
\mathbf{K}_X &= (\mathbf{K}_X^0)_+, \\
\mathbf{K}_Y &= (\mathbf{K}_Y^0)_+.
\end{aligned} \tag{4.6}$$

Before proving Theorem 4.2.2 we will need to make use of the following lemma. Let $\mathbf{M}_0 \in \mathbb{R}^{n\times n}$ be a known, square, not necessarily positive-definite matrix, and $\mathbf{M} \in \mathbb{R}^{n\times n}$ a square, unknown matrix, then

Lemma 4.2.3. *The solution to the following optimization problem,*

$$arg\min_{\mathbf{M}\succeq 0} ||\mathbf{M} - \mathbf{M}_0||_F^2,$$

is

$$\mathbf{M} = (\mathbf{M}_0)_+ .$$

Proof. Let $\mathbf{\Lambda}_{M_0} = \text{diag}(\lambda_{M_0}^1, \ldots, \lambda_{M_0}^n)$ and $\mathbf{V}_{M_0}$, $i = 1, \ldots, n$ denote the eigenvalues and eigenvectors of $\mathbf{M}_0$. Note that for any real matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$ and orthonormal basis $\mathbf{V} \in \mathbb{R}^{q\times q}$ that

$$\begin{aligned} ||\mathbf{A}||_F^2 &= \text{Tr}(\mathbf{A}^T\mathbf{A}) \\ &= \text{Tr}(\mathbf{V}\mathbf{A}^T\mathbf{V}^T\mathbf{V}\mathbf{A}\mathbf{V}^T) \\ &= ||\mathbf{V}\mathbf{A}\mathbf{V}^T||_F^2. \end{aligned}$$

Keeping this identity in mind the optimization problem in (4.2.3) can be restated as

$$\begin{aligned} &\arg\min_{\mathbf{M}\succeq 0} ||\mathbf{M} - \mathbf{M}_0||_F^2 \\ &= \arg\min_{\mathbf{M}\succeq 0} ||\mathbf{V}_{M_0}^T(\mathbf{M} - \mathbf{M}_0)\mathbf{V}_{M_0}||_F^2 \\ &= \arg\min_{\mathbf{M}\succeq 0} ||\mathbf{V}_{M_0}^T\mathbf{M}\mathbf{V}_{M_0} - \mathbf{\Lambda}_{M_0}||_F^2. \end{aligned}$$

Note that since $\mathbf{\Lambda}_{M_0}$ is diagonal $\mathbf{V}_{M_0}^T\mathbf{M}\mathbf{V}_{M_0}$ should be diagonal in order to minimize the Frobenius norm. This implies that $\mathbf{V}_{M_0}$ must be the eigenvectors of $\mathbf{M}$. Thus we can

assume that the matrix $\mathbf{M}$ which minimizes the above problem has the form

$$\mathbf{M} = \mathbf{V}_{M_0} \mathbf{\Lambda}_M \mathbf{V}_{M_0}^T,$$

where $\mathbf{\Lambda}_M$ is a diagonal matrix with entries λ_M^i, $i = 1, \ldots, n$. The problem then becomes

$$\arg\min_{\mathbf{M} \succeq 0} ||\mathbf{\Lambda}_M - \mathbf{\Lambda}_{M_0}||_F^2 = \arg\min_{\lambda_M^i \geq 0} \sum_{i=1}^{n} (\lambda_M^i - \lambda_{M_0}^i)^2.$$

Clearly the quantity which minimizes this is $\lambda_M^i = \max(0, \lambda_{M_0})$. Thus we have that $\mathbf{M} = (\mathbf{M}_0)_+$ as we wanted to show. □

We now return to our proof of Theorem 4.2.2.

Proof. We begin by expanding out the terms in the objective function (4.4)

$$\begin{aligned}
\rho_{\mathcal{K}} &= \mathrm{Tr}(\mathbf{A}_X^T \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y) + \lambda_X ||\mathbf{K}_X - \mathbf{K}_X^0||_F^2 \\
&\quad + \lambda_Y ||\mathbf{K}_Y - \mathbf{K}_Y^0||_F^2 \\
&= \mathrm{Tr}(\mathbf{A}_X^T \mathbf{K}_X \mathbf{K}_Y \mathbf{A}_Y) + \lambda_X \mathrm{Tr}((\mathbf{K}_X - \mathbf{K}_X^0)^T (\mathbf{K}_X - \mathbf{K}_X^0)) + \lambda_Y \mathrm{Tr}((\mathbf{K}_Y - \mathbf{K}_Y^0)^T (\mathbf{K}_Y - \mathbf{K}_Y^0)) \\
&= \mathrm{Tr}(\mathbf{K}_Y \mathbf{A}_Y \mathbf{A}_X^T \mathbf{K}_X) + \lambda_X \mathrm{Tr}(\mathbf{K}_X \mathbf{K}_X - 2\mathbf{K}_X \mathbf{K}_X^0) + \lambda_Y \mathrm{Tr}(\mathbf{K}_Y \mathbf{K}_Y - 2\mathbf{K}_Y \mathbf{K}_Y^0) \\
&\quad + \lambda_X \mathrm{Tr}(\mathbf{K}_X^0 \mathbf{K}_X^0) + \lambda_Y \mathrm{Tr}(\mathbf{K}_Y^0 \mathbf{K}_Y^0).
\end{aligned}$$

Letting $C = \lambda_Y \mathrm{Tr}(\mathbf{K}_Y \mathbf{K}_Y - 2\mathbf{K}_Y \mathbf{K}_Y^0) + \lambda_Y \mathrm{Tr}(\mathbf{K}_Y^0 \mathbf{K}_Y^0) + \lambda_X \mathrm{Tr}(\mathbf{K}_X^0 \mathbf{K}_X^0)$ and $\mathbf{G}_{YX} = \mathbf{K}_Y \mathbf{A}_Y \mathbf{A}_X^T$ we have

$$\begin{aligned}
\rho_{\mathcal{H}} &= \mathrm{Tr}(\mathbf{G}_{YX} \mathbf{K}_X + \lambda_X \mathbf{K}_X \mathbf{K}_X - 2\lambda_X \mathbf{K}_X^0 \mathbf{K}_X) + C \\
&= \lambda_X \mathrm{Tr}\left(\left[\mathbf{K}_X - 2 \left(\frac{1}{2\lambda_X} \mathbf{G}_{YX} + \mathbf{K}_X^0 \right) \right] \mathbf{K}_X \right) + C.
\end{aligned}$$

Adding and subtracting $||\frac{1}{2\lambda_X}\mathbf{G}_{YX} + \mathbf{K}_X^0||_F^2$ we have

$$\rho_{\mathcal{K}} = \lambda_X \left|\left|\mathbf{K}_X - \left(\frac{1}{2\lambda_X}\mathbf{G}_{YX} + \mathbf{K}_X^0\right)\right|\right|_F^2 - \left|\left|\frac{1}{2\lambda_X}\mathbf{G}_{YX} + \mathbf{K}_X^0\right|\right|_F^2 + C.$$

Note that there is only one term involving $\mathbf{K}_X$. Thus the minim

$$\min_{\mathbf{K}_X} \left|\left|\mathbf{K}_X - \left(\frac{1}{2\lambda_X}\mathbf{G}_{YX} + \mathbf{K}_X^0\right)\right|\right|_F^2$$

subject to,

$$\mathbf{K}_X \succeq 0. \tag{4.7}$$

For the purpose of our application we only consider the case where $\lambda_X \to \infty$, forcing $\mathbf{K}_X$ to be the closest proxy of the matrix $\mathbf{K}_X^0$. This then becomes the projection of the matrix $\mathbf{K}_X^0$ on the cone of positive semidefinite matrices (Luss and d'Aspremont (2008)). The optimal solution to this problem is given by

$$\mathbf{K}_X = (\mathbf{K}_X^0)_+,$$

as we wanted to show. Similar results hold for $\mathbf{K}_Y$. □

Note that it is equivalent to solve the IKCCA problem by solving the regularized CCA optimization problem replacing the matrices $\mathbf{X}$ and $\mathbf{Y}$ with the matrices $\mathbf{C}_X$ and $\mathbf{C}_Y$, respectively, where

$$\mathbf{C}_X = \mathbf{K}_X^0 \mathbf{V}_X^+,$$
$$\mathbf{C}_Y = \mathbf{K}_Y^0 \mathbf{V}_Y^+.$$

The matrices $\mathbf{V}_X^+$ and $\mathbf{V}_Y^+$ are the matrices of eigenvectors corresponding to the positive eigenvalues in X and Y space respectively. A justification for this equivalency can be found in Kuss and Graepel (2003).

With this in mind, out-of-sample points $\mathbf{x} \in \mathbb{R}^{d_X}$ and $\mathbf{y} \in \mathbb{R}^{d_Y}$ are projected onto their first p canonical directions as follows: first compute their kernelization, using the indefinite kernel functions K_X^0 and K_Y^0

$$K_X^0(\mathbf{x}, .) = (K_X^0(\mathbf{x}, \mathbf{x}_1), \ldots, K_X^0(\mathbf{x}, \mathbf{x}_n))^T,$$
$$K_Y^0(\mathbf{x}, .) = (K_Y^0(\mathbf{y}, \mathbf{y}_1), \ldots, K_Y^0(\mathbf{y}, \mathbf{y}_n))^T.$$

Next, K_X^0 and K_Y^0 are projected onto the matrices of eigenvectors $\mathbf{V}_X^+$ and $\mathbf{V}_Y^+$, respectively, giving us

$$K_X(\mathbf{x}) = K_X^0 \mathbf{V}_X^0 \in \mathbb{R}^{p_X},$$
$$K_Y(\mathbf{y}) = K_Y^0 \mathbf{V}_Y^0 \in \mathbb{R}^{p_Y}.$$

Here p_X and p_Y correspond the the number of non-zero eigenvalues in X and Y space respectively. Finally, the projections onto the canonical directions are given by

$$f(\mathbf{x}) = \langle K_X(\mathbf{x}), \alpha_X \rangle = \sum_{i=1}^{p} \alpha_X^i K(\mathbf{x})_i,$$
$$f(\mathbf{y}) = \langle K_Y(\mathbf{y}), \alpha_Y \rangle = \sum_{i=1}^{p} \alpha_Y^i K(\mathbf{y})_i.$$

where the α_X^i's and α_Y^i's are the solutions from the IKCCA optimization problem in (4.4) (also note that $p \leq \min(p_X, p_Y)$).

In the following section we show that for the class of kernels in (4.1) an interesting and intuitive connection can be made between IKCCA and LDA.

In particular we study a class of kernels related to the normalized graph Laplacian (NGL) used in spectral clustering (Chung (1997), Shi and Malik (2000), Ng *et al.* (2002), Belkin and Niyogi (2003), Bengio *et al.* (2004), v. Luxburg (2007), v. Luxburg *et al.* (2008), Zelnik-Manor and Perona (2004)). In Section 4.3 we provide an overview of

spectral clustering and some associated properties. Then in Section 4.4 we show the connection between the NGL kernel for IKCCA and LDA.

4.3 Spectral Clustering

In this section we provide an overview of spectral clustering and its properties. Our discussion follows that of v. Luxburg (2007).

The intuitive goal of clustering can be summarized as follows: given a set of n data points, $\mathbf{x}_i \in \mathbb{R}^d$, and some measure of similarity between them, w_{ij} the goal is to divide the data points into several groups such that points in the same group are similar and points in different groups are dissimilar. A convenient way of representing the data in this context is in the form of a *similarity graph* $G = (V, E)$, $V = \{v_1, \ldots, v_n\}$, $E = \{w_{ij}\}$. The vertices $v_i \in V$ in this graph are the points $\mathbf{x}_i$. Two vertices are connected if the similarity, w_{ij} between the corresponding data points $\mathbf{x}_i$ and $\mathbf{x}_j$ is positive. The edge between them is given the weight w_{ij}.

The similarity graph provides a natural framework for clustering evidenced by the following restatement of the clustering problem: given the graph G the goal is to find a partition such that the weights of the edges within a group are large (i.e. that points which are similar to one another fall into the same cluster) and the weights of the edges between groups is small (i.e. that points which are dissimilar to one another are in different clusters). In the following section, we introduce some graph notation and briefly describe the types of graphs we are going to study.

4.3.1 Graph Notation

Let $G = (V, E)$ be an undirected graph with vertex set $V = \{v_1, \ldots, v_n\}$. We assume that the graph is weighted with non-negative edge weights $w_{ij} (\geq 0)$ between vertices v_i and v_j. The weighted adjacency matrix of a graph is a square, symmetric matrix $\mathbf{W} = (w_{ij})_{i,j=1}^n$. If $w_{ij} = 0$ this means that vertices v_i and v_j are not connected. The

degree of a vertex $v_i \in V$ is defined as

$$d_i = \sum_{j=1}^{n} w_{ij}.$$

Note that this sum runs over only those vertices which are adjacent to v_i, and tells us how well connected a vertex is. The *degree matrix* is defined as $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$. Given a subset of vertices $A \subset V$ we define its compliment as $\bar{A} = V - A$. Define the indicator vector, $\mathbf{f}_A = (f_1, \ldots, f_n)^T \in \mathbb{R}^n$ as the vector with entries $f_i = 1$ if $v_i \in A$ and $f_i = 0$ otherwise. Convenient shorthand is to write $i \in A$ to mean the set of indices $\{i | v_i \in A\}$. The two ways in which we measure the size of a set A is

$$\begin{aligned} |A| &:= \text{ the number of vertices in } A, \\ \text{vol}(A) &:= \sum_{i \in A} d_i. \end{aligned} \tag{4.8}$$

Intuitively we can think of $|A|$ as measuring the size of A by the number of vertices it contains and $\text{vol}(A)$ as measuring the size of A by the weights of its edges.

A subset $A \subset V$ is called *connected* if any two vertices A can be joined by a path such that all intermediate points also lie in A. A subset A is called a *connected component* if it is connected and if there are no connections between the vertices of A and $\bar{A}$.

In conventional set theory $A_1, \ldots, A_k$ form a *partition* when $A_i \cap A_j = \emptyset$ and $\cup_{i=1}^{k} A_i = V$. In graph theory there is a similar, stronger definition of a partition with the sets A_i, $i = 1, \ldots, k$ defined as connected components explicitly constructed from the similarity graph G.

4.3.2 Similarity Graphs

The goal in constructing a similarity graph, G, is to model the local distribution of the data, $v_i, i = 1, \ldots, n$. Below we list some of the similarity graphs that are frequently used in spectral clustering

1. **The ϵ-neighborhood graph**: Here all points (i.e. vertices v_i) which are in an ϵ-neighborhood of one another are connected by an edge. The potential shortcoming of this type of graph is that using a fixed ϵ may not capture the changes in the local scale of the data.

2. **The k-nearest neighbor graphs**: Here we connect the point v_j to the point v_i if v_j is within the k-neighborhood of v_i. However, care needs to be taken to avoid a graph that is not symmetric. There are two ways in which this is typically handled; the first is to put an edge between v_i and v_j if one is in the neighborhood of the other. The second is to only put an edge between v_i and v_j if they are both in the neighborhood of the other.

3. **The Fully Connected Graph**: Here all vertices in the graph are connected by a positive weight. In order to model the local behavior of the data typically a similarity function is used which can capture this type of information, e.g. the Gaussian similarity function $w_{ij} = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right)$.

4.3.3 Graph Laplacians

The main concept of spectral clustering revolves around the graph Laplacian matrix and its various representations (see Chung (1997) for a more detailed discussion). Here we provide an overview of some of the definitions and basic properties associated with the graph Laplacian. In the following, since we are dealing with generalized eigenvalue problems, when we speak of eigenvectors we do not assume that they have unit length. Additionally we assume that eigenvalues are ordered increasingly and when we speak of the first k eigenvectors we mean those eigenvectors associated with the k smallest eigenvalues.

The Unnormalized Graph Laplacian

The unnormalized graph Laplacian matrix is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{W}.$$

Recall membership in the connected component A is captured by the indicator vector $\mathbf{f}_A$. The quadratic form $\mathbf{f}^T\mathbf{L}\mathbf{f}$ will play the role of cluster index in spectral clustering.

The following proposition summarizes most of the important facts needed (see Mohar (1991) and Mohar and Juvan (1997) for further details)

Proposition 4.3.1. *The matrix* $\mathbf{L}$ *satisfies the following properties*

1. *For every vector* $\mathbf{g} \in \mathbb{R}^n$ *our cluster index can be computed as*

$$\mathbf{g}^T\mathbf{L}\mathbf{g} = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(g_i - g_j)^2.$$

2. $\mathbf{L}$ *is symmetric and positive semi-definite.*

3. *The smallest eigenvalue of* $\mathbf{L}$ *is 0, the corresponding eigenvector is the constant vector* $\mathbf{1} \in \mathbb{R}^n$.

4. $\mathbf{L}$ *has n non-negative, real-valued eigenvalues* $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.

For a proof see v. Luxburg (2007).

The unnormalized graph Laplacian and its eigenvalues and eigenvectors can be used to describe many properties of graphs. The following proposition is particularly important in spectral clustering:

Proposition 4.3.2. *(Number of Connected Components) Let* G *be an undirected graph with non-negative weights. Then the multiplicity* k *of the eigenvalue 0 of* $\mathbf{L}$ *equals the number of connected components* $A_1, \ldots, A_n$ *in the graph. The eigenspace of eigenvalue 0 is spanned by the indicator vectors* $\mathbf{f}_{A_1}, \ldots, \mathbf{f}_{A_n}$ *of those components.*

Remark 4.3.3. This proposition has been proven in v. Luxburg (2007). A similar proof is given here to highlight the way in which the graph Laplacian's eigenvectors behave as indicator (i.e. label) vectors.

Proof. For a fully connected graph, i.e. $k = 1$, we know from Proposition 4.3.1 that the smallest eigenvalue of $\mathbf{L}$ is $\lambda_1 = 0$ and the corresponding eigenvector is the constant vector $\mathbf{1}_n$.

For $k > 1$, assume without loss of generality that the vertices are ordered according to which connected component they belong to, the graph Laplacian then takes the block diagonal form

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{L}_k \end{pmatrix}.$$

The key observation to be made here is that each block $\mathbf{L}_i$, $i = 1, \ldots, k$ is itself a proper graph Laplacian. Therefore each of these blocks must have 0 as an eigenvalue and the constant vector $\mathbf{1}_{n_i}$ as an eigenvector, where n_i is the number of vertices contained in the i^{th} connected component. Thus, the matrix $\mathbf{L}$ has as many eigenvalues 0 as there are connected components, and the corresponding eigenvectors are the indicator vectors, $\mathbf{f}_{A_i}$ of the connected components. □

The Normalized Graph Laplacian

In this section we present some results on the normalized graph Laplacian. The normalized graph Laplacian is of particular interest to us as it has been shown by v. Luxburg *et al.* (2008) to have much stronger consistency properties, in terms of the convergence of its sample eigenvalues and eigenvectors to their population counterparts, then its unnormalized counterpart. For this reason in the discussion that follows we focus on the normalized graph Laplacian.

There are two closely related matrices which are referred to as normalized Graph Laplacians in the literature, these are

$$\mathbf{L}_{sym} := \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}},$$
$$\mathbf{L}_{rw} := \mathbf{D}^{-1}\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W}.$$

The first matrix is denoted as $\mathbf{L}_{sym}$ since it is symmetric. The second matrix is denoted by $\mathbf{L}_{rw}$ since it is closely related to the transition matrix of a random walk. The transition matrix in this case would be composed of transition probabilities of jumping in one step from vertex i to vertex j which would be equal to $p_{ij} := \frac{w_{ij}}{d_i}$.

Next we summarize some of the properties of these two matrices (see Chung (1997), Mohar (1991) and Mohar and Juvan (1997), for further details). The key properties associated with the normalized graph Laplacians are summarized below in Proposition 4.3.4. These properties are similar to those presented in the unnormalized case (Proposition 4.3.1).

Proposition 4.3.4. *(Properties of $\mathbf{L}_{sym}$ and $\mathbf{L}_{rw}$) The normalized Laplacians satisfy the following properties:*

1. *For every $\mathbf{g} \in \mathbb{R}^n$ we have*

$$\mathbf{g}^T\mathbf{L}_{sym}\mathbf{g} = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(\frac{g_i}{\sqrt{d_i}} - \frac{g_j}{\sqrt{d_j}}\right)^2.$$

2. *λ is an eigenvalue of $\mathbf{L}_{rw}$ with eigenvector $\mathbf{v}$ if and only if λ is an eigenvalue of $\mathbf{L}_{sym}$ with eigenvector $\mathbf{w} = \mathbf{D}^{\frac{1}{2}}\mathbf{v}$.*

3. *λ is an eigenvalue of $\mathbf{L}_{rw}$ with eigenvector $\mathbf{v}$ if and only if λ and $\mathbf{v}$ solve the generalized eigenvalue problem $\mathbf{L}\mathbf{v} = \lambda\mathbf{D}\mathbf{v}$.*

4. *0 is an eigenvalue of $\mathbf{L}_{rw}$ with the constant vector $\mathbf{1}_n$ as an eigenvector. 0 is an eigenvalue of $\mathbf{L}_{sym}$ with eigenvector $\mathbf{D}^{\frac{1}{2}}\mathbf{1}_n$.*

5. $\mathbf{L}_{sym}$ *and* $\mathbf{L}_{rw}$ *are positive semi-definite and have* n *non-negative real-valued eigenvalues* $0 = \lambda_1 \leq \ldots \leq \lambda_n$.

For a proof see v. Luxburg (2007).

The following proposition provides similar results to those discussed in Proposition 4.3.2 but for the normalized case.

Proposition 4.3.5. *Let* G *be an undirected graph with non-negative weights. Then the multiplicity* k *of the eigenvalue 0 of both* $\mathbf{L}_{sym}$ *and* $\mathbf{L}_{rw}$ *equals the number of connected components* $A_1, \ldots, A_n$ *in the graph. For* $\mathbf{L}_{sym}$ *the eigenspace of 0 is spanned by the vectors* $\mathbf{D}^{\frac{1}{2}}\mathbf{f}_{A_i}$. *For* $\mathbf{L}_{rw}$ *the eigenspace of 0 is spanned by the indicator vectors* $\mathbf{f}_{A_i}$.

Remark 4.3.6. Recall that in the protein-ligand matching problem we were primarily interested in predicting the binding between as of yet unobserved proteins and ligands. For this reason it is important that there be a direct way to compute the kernelization for out-of-sample observations. Because of this we use the weighted adjacency matrix

$$\mathbf{K} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}, \tag{4.9}$$

rather than the normalized graph Laplacian $\mathbf{L}_{sym}$, as there is no direct extension of the symmetric normalized graph Laplacian to out-of-sample observations. The lack of a direct out-of-sample extension can be seen from the following: by definition the i, j^{th} element of the symmetric normalized graph Laplacian is

$$(\mathbf{L}_{sym})_{ij} = \begin{cases} 1 - \frac{w_{ii}}{d_i} & \text{if } i = j, \text{ and } d_i \neq 0, \\ -\frac{w_{ij}}{\sqrt{d_i d_j}} & \text{if } i \text{ and } j \text{ are adjacent,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.10}$$

Thus, given a new observation, while it is possible to calculate its value in the last two cases of (4.10), it is not possible to calculate its value in the first case.

What is important to note is that we do not lose any relevant information by using $\mathbf{K}$ instead of $\mathbf{L}_{sym}$. The weighted adjacency matrix $\mathbf{K}$ has the same eigenvectors as $\mathbf{L}_{sym}$ and its eigenvalues are equal to $1-\lambda_i$, $i = 1, \ldots, n$, where λ_i are the eigenvalues of $\mathbf{L}_{sym}$.

In addition, the results stated in Proposition 4.3.5 still hold for $\mathbf{K}$ with the modification that the multiplicity of the eigenvalue 1 rather than 0 equals the number of connected components. This can be seen by noting that since the smallest eigenvalue of $\mathbf{L}_{sym} = \mathbf{I}_n - \mathbf{K}(\succeq 0)$ is 0, the largest eigenvalue of $\mathbf{K}$ (corresponding to the number of 0's in $\mathbf{L}_{sym}$) must be 1.

We can also establish a lower bound on the eigenvalues of $\mathbf{K}$, utilizing the following inequality

$$\left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}}\right)^2 \leq 2\left(\frac{f_i^2}{d_i} + \frac{f_j^2}{d_j}\right).$$

Keeping in mind that the eigenvectors $\mathbf{f}$ of $\mathbf{L}_{sym}$ have unit length (see Section 4.3.5) we have

$$\begin{aligned}
\mathbf{f}^T\mathbf{L}_{sym}\mathbf{f} &= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}}\right)^2 \\
&\leq \sum_{i,j=1}^{n} w_{ij}\left(\frac{f_i^2}{d_i} + \frac{f_j^2}{d_j}\right) \\
&= \sum_{i=1}^{n} f_i^2 + \sum_{j=1}^{n} f_j^2 \\
&= 2.
\end{aligned}$$

Therefore the smallest possible eigenvalue of $\mathbf{K}$ is $1 - \lambda_{max} = -1$. The consequence of this is that $\mathbf{K}$ is not strictly positive semi-definite, i.e. it may be indefinite. In order to be able to meaningfully use the weighted adjacency matrix with KCCA, conditions like those we discussed in Section 4.2, need to be introduced as otherwise there is no guarantee that the solutions we find will be meaningful.

Remark 4.3.7. In the following sections we refer to the weighted adjacency matrix as the

normalized graph Laplacian (NGL) kernel to emphasize its connection with the graph Laplacian. In Sections 4.4.1 and 4.4.2 the properties of the graph Laplacian discussed in Propositions 4.3.2 and 4.3.5 will be shown to connect IKCCA with LDA.

4.3.4 Spectral Clustering Algorithms

There are a number of spectral clustering algorithms used in practice. Here we state one algorithm which is commonly used in conjunction with the normalized symmetric graph Laplacian, $\mathbf{L}_{sym}$. Most spectral clustering algorithms have a similar structure.

Normalized Spectral Clustering according to Ng, Jordan and Weiss (2002)
Input: Similarity measure w_{ij}, number of clusters k

- Construct a similarity graph by one of the ways described in Section 4.3.2. Let $\mathbf{W}$ be its weighted adjacency matrix.
- Compute the normalized graph Laplacian $\mathbf{L}_{sym}$.
- Compute the first k eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of $\mathbf{L}_{sym}$.
- Let $\mathbf{V} \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ as columns.
- Form the matrix $\mathbf{U} \in \mathbb{R}^{n \times k}$ from $\mathbf{V}$ by normalizing the row sums to have norm 1, that is $u_{ij} = \frac{v_{ij}}{\left(\sum_{m=1}^{k} v_{im}^2\right)^{\frac{1}{2}}}$.
- For $i = 1, \ldots, n$, let $\mathbf{y}_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of $\mathbf{U}$.
- Cluster the points $\{\mathbf{y}_i\}_{i=1}^n$ with the k-means algorithm into clusters $C_1, \ldots, C_k$.

Output: Clusters $A_1, \ldots, A_k$ with $A_i = \{j | \mathbf{y}_j \in C_i\}$.

A note on the normalization step in the above algorithm. Recall from Proposition 4.3.5 that the eigenvectors corresponding to the smallest eigenvalue of each of the connected components of $\mathbf{L}_{sym}$ is equal to $\mathbf{D}^{\frac{1}{2}}\mathbf{f}_{A_i}$, where $\mathbf{f}_{A_i}$ is the indicator vector of the i^{th} connected

component. The purpose behind normalizing by $\frac{1}{\left(\sum_{m=1}^{k} v_{im}^2\right)^{\frac{1}{2}}}$ is therefore to retrieve the indicator vectors $\mathbf{f}_{A_i}$. In the more general setting where there is possible overlap between groups the purpose is to approximate the indicator vectors as closely as possible. In both cases this is meant to make the identification of the clusters in the k-means step easier.

4.3.5 Graph Cut Point of View

As stated at the beginning of this section, representing data in the form of a similarity graph provides a powerful approach to clustering. From a graph theoretic standpoint the clustering problem is typically formulated in terms of the *graph partitioning problem*. The objective of the graph partitioning problem is to divide a graph into k disjoint parts such that each of these parts is approximately equal in size and the sum of the edge weights is minimized. In this section we will show how spectral clustering can be derived as an approximate solution to the graph partitioning problem.

Given two disjoint subsets $A, B \subset V$ define

$$\text{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij}.$$

Given the adjacency matrix $\mathbf{W}$ the most straightforward way to construct a partition is to solve the *mincut problem*. This consists of finding a partition $A_1, \ldots, A_k$ which minimizes

$$\text{cut}(A_1, \ldots, A_k) := \sum_{i=1}^{k} \text{cut}(A_i, \bar{A}_i).$$

However, in practice this often does not lead to satisfactory partitions. The problem is that frequently the solution of the mincut problem results in one vertex being separated from the rest. This is of course not what we are usually interested in achieving. One way to avoid this issue is formulate the problem in such a way that the sets $A_1, \ldots, A_k$ are "reasonably large". The two most common objective functions which incorporate this are the RatioCut (Shi and Malik (2000)) and the normalized cut Ncut (Shi and Malik

(2000)). In RatioCut the size of a subset A is measured by the number of vertices, $|A|$, while in the Ncut the size of a subset of A is measured by weights of its edges $\text{vol}(A)$ (defined in (4.8)). The definitions are

$$\text{RatioCut}(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|},$$
$$\text{Ncut}(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{\text{vol}(A_i)}.$$

What both objective functions try to achieve is a balance in the clusters as measured by the number of vertices or edge weights, respectively. Unfortunately, by having these balancing conditions the mincut problem becomes NP hard (see Wagner and Wagner (1993) for details). What we will see is that spectral clustering is a way to solve "relaxed" versions of these problems. We focus here on the Ncut problem as this leads to the normalized spectral clustering problem, which is what we are primarily interested in (see v. Luxburg (2007) for a spectral clustering approach to the RatioCut problem).

Approximating Ncut

Following the discussion in v. Luxburg (2007), we begin with the case where the number of clusters k is 2. Define the cluster indicator vector $\mathbf{f}$ by

$$f_i = \begin{cases} \sqrt{\frac{\text{vol}(\bar{A})}{\text{vol}(A)}} & \text{if } i \in A, \\ -\sqrt{\frac{\text{vol}(A)}{\text{vol}(\bar{A})}} & \text{if } i \in \bar{A}. \end{cases} \tag{4.11}$$

One can check that $(\mathbf{Df})^T \mathbf{1}_n = 0$, $\mathbf{f}^T \mathbf{Df} = \text{vol}(V)$, and $\mathbf{f}^T \mathbf{Lf} = 2\text{vol}(V)\text{Ncut}(A, \bar{A})$. With this in mind an equivalent restatement of the Ncut problem is

$$\min_A \mathbf{f}^T \mathbf{Lf}$$

subject to,

$$\mathbf{f} \text{ as in (4.11) and } \mathbf{D}\mathbf{f}^T\mathbf{1}_n = 0, \mathbf{f}^T\mathbf{D}\mathbf{f} = \text{vol}(V). \tag{4.12}$$

This is an NP-hard discrete optimization problem (Wagner and Wagner (1993)) as the entries $\mathbf{f}$ are only allowed to take one of two values. The obvious relaxation in this setting is to remove the condition that the f_i's take one of two values and allow $f_i \in \mathbb{R}$. This leads to the relaxed optimization problem

$$\min_{\mathbf{f}} \mathbf{f}^T\mathbf{L}\mathbf{f}$$

subject to,

$$\mathbf{D}\mathbf{f}^T\mathbf{1}_n = 0, \mathbf{f}^T\mathbf{D}\mathbf{f} = \text{vol}(V).$$

Letting $\mathbf{g} = \mathbf{D}^{\frac{1}{2}}\mathbf{f}$ we have

$$\min_{\mathbf{g}} \mathbf{g}^T\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}\mathbf{g}$$

subject to,

$$\mathbf{g}^T\mathbf{D}^{\frac{1}{2}}\mathbf{1}_n = 0, ||\mathbf{g}||^2 = \text{vol}(V). \tag{4.13}$$

This is exactly the spectral clustering problem for $k = 2$ using the symmetric normalized graph Laplacian.

Generalizing to the case of $k > 2$ cluster, we begin by defining the indicator vectors $\mathbf{h}_i = (h_{1i}, \ldots, h_{ni})^T$, where

$$h_{ij} = \begin{cases} \frac{1}{\sqrt{\text{vol}(A_i)}} & \text{if } i \in A_j, \\ 0 & \text{otherwise .} \end{cases} \tag{4.14}$$

Letting $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_k) \in \mathbb{R}^{n \times k}$ we have that $\mathbf{H}^T\mathbf{H} = \mathbf{I}_k$, $\mathbf{h}_i^T\mathbf{D}\mathbf{h}_i = 1$, and $\mathbf{h}_i^T\mathbf{L}\mathbf{h}_i =$

$\frac{2\text{cut}(A_i,\bar{A}_i)}{\text{vol}(A_i)}$. We can then write the Ncut problem as

$$\min_{A_1,\ldots,A_k} \text{Tr}(\mathbf{H}^T\mathbf{L}\mathbf{H}) \text{ subject to } \mathbf{H}^T\mathbf{D}\mathbf{H} = \mathbf{I}_k.$$

As above, we relax the discreteness condition and substitute $\mathbf{U} = \mathbf{D}^{\frac{1}{2}}\mathbf{H}$, which then gives us

$$\min_{\mathbf{U}} \text{Tr}\left(\mathbf{U}^T\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}\mathbf{U}\right) \text{ subject to } \mathbf{U}^T\mathbf{U} = \mathbf{I}_k.$$

The solution to the latter is simply the eigendecomposition of $\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$.

4.4 Connecting the NGL Kernel for IKCCA with LDA

In the first part of this section we show that under some certain assumptions on the distribution of the data, when the NGL kernel is used, IKCCA finds the same directions as LDA. We also explore conditions under which the directions found by IKCCA deviate from those found by LDA. At the end of this section we extend these results to the more general setting by using the idea of "spectral relaxation" discussed in Section 4.3. The purpose behind this discussion is to provide a more concrete foundation for understanding how IKCCA behaves, when the NGL kernel is used.

4.4.1 IKCCA and LDA

We begin by describing the distribution of the data which we propose to study. We consider two scenarios, the first is an IKCCA setting which corresponds to standard LDA and the second scenario is the standard LDA setting.

1. As before we have a collection of pairs of observations $\mathbf{x}_i \in \mathbb{R}^{d_X}$ and $\mathbf{y}_i \in \mathbb{R}^{d_Y}$, $i = 1, \ldots, n$ which we will refer to as the data space and label space respectively. The $\mathbf{x}_i$'s (data space) fall into two distinct groups, highlighted in red and green and labeled by a "+" and "−" respectively in the left plot of Figure 4.4.1. The

$\mathbf{y}_i$'s (label space) also fall into two groups centered at $\mu_- = (\mu_-^1, \ldots, \mu_-^{d_Y})^T$ and $\mu_+ = (\mu_+^1, \ldots, \mu_+^{d_Y})^T$, shown in the right plot in Figure 4.4.1. The distribution of points within each of these groups follows the uniform distribution on a sphere with radius r. In the plot on the right in Figure 4.4.1 (label space) the means are connected by a dashed black line, the corresponding distance between the means is $\Delta = ||\mu_+ - \mu_-||$. The solid circles and lines correspond to the support type and radius of the support, respectively of the two groups ("+" in red and "−" in green). The dashed circles and connecting lines indicate the $2r$-neighborhoods of the two points in each group that are closest to the other group. Note that so long as $\Delta \geq 6r$ there will be no overlap in any of the $2r$-neighborhoods in each of the spaces.

2. The distribution of the $\mathbf{x}_i$'s are the same as described above but now the $\mathbf{y}_i$'s are label vectors, i.e. $y_{i1} = 1$ if $\mathbf{x}_i \in C_+$ and 0 otherwise and $y_{i2} = 1$ if $\mathbf{x}_i \in C_-$ and 0 otherwise, where C_+ and C_- correspond to the + and − group in the data space (note that $\mathbf{x}_i$ can only belong to one class). See Section 2.5.1 for details.

Recall from our discussion in Section 4.3.3 that given an adjacency matrix $\mathbf{W}$ the NGL kernel is defined as

$$\mathbf{K} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}},$$

where the ij^{th} element has the form

$$K_{ij} = \frac{w_{ij}}{\sqrt{\sum_{i'=1}^n w_{i'j}}\sqrt{\sum_{j'=1}^n w_{ij'}}}. \tag{4.15}$$

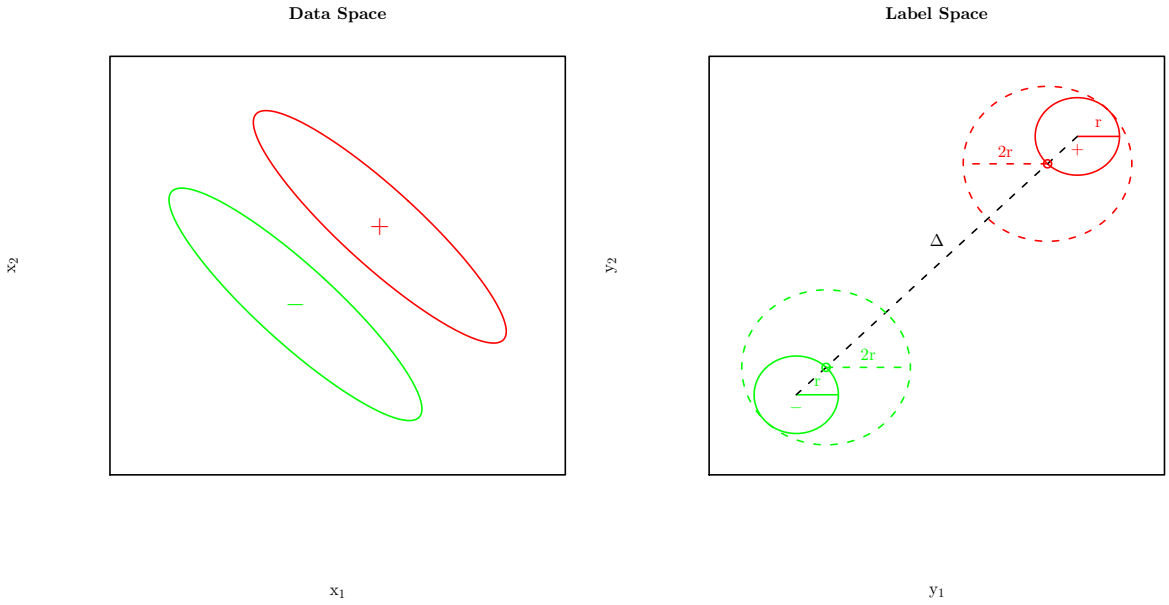

Figure 4.1: *A plot of the data as described in Scenario 1. In the Label Space plot the means are connected by a dashed black line, the corresponding distance between the means is* $\Delta = ||\mu_+ - \mu_-||$*. The solid circles and lines correspond to the support type and radius of the support, respectively of the two groups ("+" in red and "−" in green). The dashed circles and connecting lines indicate the* $2r$*-neighborhoods of the two points in each group that are closest to the other group.*

In this example we define the weights w_{ij} to be

$$w_{ij} = \begin{cases} 1 & \text{if } ||\mathbf{x}_i - \mathbf{x}_j|| \leq 2r \\ 0 & \text{otherwise.} \end{cases} \tag{4.16}$$

Theorem 4.4.1. *Given the distribution of the data as described in scenario (1) above with* n_1 *and* n_2 *observations in groups "+" and "−" respectively* ($n = n_1 + n_2$) *and the NGL kernel as represented in (4.15) and (4.16), if* $\Delta \geq 6r$ *then the directions found by IKCCA are identical to those found by LDA (i.e. in Scenario (2) described above).*

Note, in the following while the matrix $\mathbf{X}$ is assumed to be mean centered, we *do not* mean center the kernel matrix $\mathbf{K}_Y$. While we would normally center $\mathbf{K}_Y$, our primary interest in this example is to illustrate that the general behavior between IKCCA and LDA is similar, this is achieved more directly and clearly if $\mathbf{K}_Y$ is not mean centered.

Proof. We begin by writing down the exact form of the kernel matrix $\mathbf{K}_Y^0$ (using the notation from Section 4). The matrix of weights $\mathbf{W}_Y^0 = \{w_{ij}^Y\}_{i,j=1}^n$ is

$$\mathbf{W}_Y^0 = \left(\begin{array}{c|c} \mathbf{1}_{n_1}\mathbf{1}_{n_1}^T & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{1}_{n_2}\mathbf{1}_{n_2}^T \end{array} \right),$$

where $\mathbf{1}_n = (1, 1, \ldots, 1)_{(n\times 1)}^T$. Next define

$$\mathbf{D}_Y^0 = \text{diag}\left(\left\{ \sum_{j=1}^n w_{ij}^Y \right\}_{j=1}^n \right) = \text{diag}(\underbrace{n_1, n_1, \ldots, n_1}_{\times n_1}, \underbrace{n_2, n_2, \ldots, n_2}_{\times n_2}),$$

then

$$\mathbf{K}_Y^0 = (\mathbf{D}_Y^0)^{-\frac{1}{2}} \mathbf{W}_Y^0 (\mathbf{D}_Y^0)^{-\frac{1}{2}} = \left(\begin{array}{c|c} \frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T & \mathbf{0} \\ \hline \mathbf{0} & \frac{1}{n_2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^T \end{array} \right).$$

The expression for the positive part of the matrix $\mathbf{K}_Y$ is

$$\mathbf{K}_Y = (\mathbf{K}_Y^0)_+ = \begin{pmatrix} \frac{1}{\sqrt{n_1}}\mathbf{1}_{n_1} & 0 \\ 0 & \frac{1}{\sqrt{n_2}}\mathbf{1}_{n_2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{n_1}}\mathbf{1}_{n_1}^T & 0 \\ 0 & \frac{1}{\sqrt{n_2}}\mathbf{1}_{n_2}^T \end{pmatrix} = \mathbf{V}_Y \mathbf{V}_Y^T,$$

where $\mathbf{V}_Y = \begin{pmatrix} \frac{1}{\sqrt{n_1}}\mathbf{1}_{n_1} & 0 \\ 0 & \frac{1}{\sqrt{n_2}}\mathbf{1}_{n_2} \end{pmatrix}$. We know that only two of the eigenvalues are non-zero since the rank of the matrix $\mathbf{K}_Y$ is 2. The IKCCA optimization problem is

$$\begin{aligned} \rho_{\mathcal{K}} &= \arg\max_{\mathbf{w}_X, \alpha_Y} \mathbf{w}_X^T \mathbf{X}^T \mathbf{K}_Y \alpha_Y \\ &= \arg\max_{\mathbf{w}_X, \alpha_Y} \mathbf{w}_X^T \mathbf{X}^T \mathbf{V}_Y \mathbf{V}_Y^T \alpha_Y \\ &= \arg\max_{\mathbf{w}_X, \mathbf{w}_Y} \mathbf{w}_X^T \mathbf{X}^T \mathbf{V}_Y \mathbf{w}_Y \end{aligned} \tag{4.17}$$

subject to,

$$\mathbf{w}_X^T \mathbf{X}^T \mathbf{X} \mathbf{w}_X = \mathbf{w}_Y^T \mathbf{V}_Y^T \mathbf{V}_Y \mathbf{w}_Y = 1.$$

Since our primary interest here is to show that the discriminant direction, $\mathbf{w}_X$, is the same for IKCCA and LDA we solve for $\mathbf{w}_Y = \mathbf{V}_Y^T \alpha_Y$ rather than α_Y. In some sense this amounts to an inversion of the kernel trick. Note that our choice of $\mathbf{V}_Y$ in $\mathbf{w}_Y$ is arbitrary since we could select any $\mathbf{V}_Y^1$ and $\mathbf{V}_Y^2$, $\mathbf{V}_Y^1 \neq \mathbf{V}_Y^2$ such that

$$\mathbf{K}_Y = \mathbf{V}_Y^1 (\mathbf{V}_Y^2)^T$$

holds. However, as we will see this does not affect our results. Let

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1}_{n_1} & 0 \\ 0 & \mathbf{1}_{n_2} \end{pmatrix},$$

which we will refer to as the label matrix. We then have that

$$\mathbf{K}_Y = \mathbf{Y} \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \mathbf{Y}^T.$$

Next define

$$\mathbf{V}_Y^1 = \mathbf{Y} \begin{pmatrix} \frac{1}{n_1^a} & 0 \\ 0 & \frac{1}{n_2^b} \end{pmatrix} \text{ and } \mathbf{V}_Y^2 = \mathbf{Y} \begin{pmatrix} \frac{1}{n_1^{1-a}} & 0 \\ 0 & \frac{1}{n_2^{1-b}} \end{pmatrix},$$

for any $-\infty < a, b < \infty$. From here on we replace $\mathbf{K}_Y \alpha_Y$ in (4.17) by $\mathbf{V}_Y^1 (\mathbf{V}_Y^2)^T \alpha_Y = \mathbf{V}_Y^1 \mathbf{w}_Y$.

Recall from Section 2.1 that the optimization problem in (4.17) (solving for $\mathbf{w}_X$) results in the following generalized eigenvalue problem

$$\mathbf{X}^T \mathbf{V}_Y^1 ((\mathbf{V}_Y^1)^T \mathbf{V}_Y^1)^{-1} (\mathbf{V}_Y^1)^T \mathbf{X} \mathbf{w}_X = \rho_{\mathcal{H}}^2 \mathbf{X}^T \mathbf{X} \mathbf{w}_X. \tag{4.18}$$

Note that the left hand side of (4.18) is in fact the between-class sum of squares matrix,

$\mathbf{S}_B$, discussed in Section 2.5. In particular, note that

$$(\mathbf{V}_Y^1)^T\mathbf{X} = \begin{pmatrix} n_1^{1-a}\mathbf{m}_1^T \\ n_2^{1-b}\mathbf{m}_2^T \end{pmatrix},$$

and

$$((\mathbf{V}_Y^1)^T\mathbf{V}_Y^1)^{-1} = \begin{pmatrix} n_1^{2a} & 0 \\ 0 & n_2^{2b} \end{pmatrix}.$$

Putting this all together we have that

$$\mathbf{X}^T\mathbf{V}_Y^1((\mathbf{V}_Y^1)^T\mathbf{V}_Y^1)^{-1}(\mathbf{V}_Y^1)^T\mathbf{X} = n_1\mathbf{m}_1\mathbf{m}_1^T + n_2\mathbf{m}_2\mathbf{m}_2^T = \mathbf{S}_B.$$

Thus (4.18) becomes

$$\mathbf{S}_B\mathbf{w}_X = \rho_{\mathcal{K}}^2\mathbf{S}_T\mathbf{w}_X,$$

where $\mathbf{S}_T = \mathbf{S}_{XX}$. From here the same calculations done in Section (2.5.2) show us that the direction found by IKCCA is the same as that found by LDA when the label matrix $\mathbf{Y}$ is explicitly known. □

Next we consider the case where $\Delta \leq 6r$. Intuitively, with all points sharing the same neighborhood, i.e. the "+" and "−" populations are indistinguishable, the directions found by IKCCA should not provide any information with regard to the separation of these groups.

Theorem 4.4.2. *Using the same framework as in Theorem 4.4.1 when $\Delta \leq 6r$ the direction $\mathbf{w}_X$ is the null vector, $\mathbf{w}_X = (0, \ldots, 0)^T$.*

Proof. The NGL kernel matrix in this context is of the form

$$\mathbf{K}_Y^0 = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T.$$

The rank of this matrix is 1, thus there is at most 1 non-zero eigenvalue. The nearest positive approximation of $\mathbf{K}_Y^0$ is then

$$\mathbf{K}_X = (\mathbf{K}_X^0)_+ = \left(\frac{1}{\sqrt{n}}\mathbf{1}_n\right)\left(\frac{1}{\sqrt{n}}\mathbf{1}_n\right)^T = \mathbf{V}_Y\mathbf{V}_Y^T.$$

Following the same steps as in Theorem 4.4.1 we have

$$\mathbf{V}_Y^T\mathbf{X} = \sqrt{n}\mathbf{m} = 0,$$

since $\mathbf{X}$ is assumed to be mean centered. The generalized eigenvalue problem then reduces to

$$\mathbf{0} = \lambda\mathbf{S}_T\mathbf{w}_X.$$

So long as $\mathbf{S}_T = \mathbf{S}_{XX}$ is non-singular, the only possible solution is $\mathbf{w}_X = (0, \ldots, 0)^T$. □

Using the NGL kernel, Theorems 4.4.1 and 4.4.2 provide some insight into the behavior of IKCCA. Under a similar framework these results extend naturally to the case of more than two classes.

4.4.2 Spectral Relaxation

In this section we provide some discussion generalizing the results of Section 4.4.1. Consider a data set consisting of n multivariate vector pairs

$$\{(\mathbf{x}_i, \mathbf{y}_i) | \mathbf{x}_i \in \mathbb{R}^{d_X}, \mathbf{y}_i \in \mathbb{R}^{d_Y}\},$$

with $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$. Furthermore let us assume that the observations $\mathbf{y}_i$ fall into two distinct groups such that the NGL kernel representation of

the matrix $\mathbf{Y}$ has the block diagonal form

$$\mathbf{K}_Y = \begin{pmatrix} \mathbf{K}_{Y1} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{Y2} \end{pmatrix}.$$

From Remark 4.3.6 following Proposition 4.3.5 we know that the number of connected components is equal to the multiplicity of the eigenvalue 1. Suppose we modify the IKCCA optimization problem in (4.1.5) to include the constraint

$$\text{rank}(\mathbf{K}_Y) = \#\{\text{eig}(\mathbf{K}_Y^0) = 1\}, \tag{4.19}$$

where $\text{eig}(\mathbf{X})$ denotes the spectrum (the set of ordered eigenvalues) of the matrix $\mathbf{X}$ and $\mathbf{K}_Y^0$ is as defined in Section 4.2. The result of this additional constraint is that the best rank k (in this case 2) representation of the kernel matrix $\mathbf{K}_Y^0$ will be selected (see Lemma 4.2.3 in Section 4.2 for details). This corresponds to selecting the first k eigenvalue-eigenvector pairs.

With this in mind we now show that the resulting IKCCA generalized eigenvalue problem will look very similar to the LDA generalized eigenvalue problem. First, we introduce some notation: let $\mathbf{v}_{y1} = (v_{y11}, \ldots, v_{y1n})^T$ and $\mathbf{v}_{y2} = (v_{y21}, \ldots, v_{y2n})^T$ be the leading eigenvectors of $\mathbf{K}_{Y1}$ and $\mathbf{K}_{Y2}$ respectively. Define the $n \times 2$ matrix

$$\mathbf{V}_Y = \begin{pmatrix} \mathbf{v}_{y1} & \mathbf{0} \\ \mathbf{0} & \mathbf{v}_{y2} \end{pmatrix},$$

and let

$$\mathbf{N}_Y = \text{diag}(\underbrace{\sqrt{n_1}v_{y11}, \ldots, \sqrt{n_1}v_{y1n_1}}_{\times n_1}, \underbrace{\sqrt{n_2}v_{y21}, \ldots, \sqrt{n_2}v_{y2n_2}}_{\times n_2}),$$

where n_1 and n_2 are the number of observations in each of the two groups in Y space.

Let

$$\mathbf{C} = \begin{pmatrix} \frac{1}{n_1}\mathbf{J}_{n_1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\mathbf{J}_{n_2} \end{pmatrix},$$

where $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n^T$.

Next we apply the NGL kernel to the $\mathbf{y}_i$'s and leave the $\mathbf{x}_i$'s unchanged. Setting the regularization parameter $\kappa = 0$, we solve the IKCCA optimization problem with the addition of the rank constraint in (4.19). After some calculations this leads to the following generalized eigenvalue problem

$$\mathbf{X}^T\mathbf{V}_Y\mathbf{V}_Y^T\mathbf{V}_Y\mathbf{V}_Y^T\mathbf{X}\mathbf{w}_X = \rho_{\mathcal{K}}^2\mathbf{X}^T\mathbf{X}\mathbf{w}_X. \tag{4.20}$$

Focusing on the left hand side of (4.20) we have

$$\begin{aligned} &\mathbf{X}^T\mathbf{V}_Y\mathbf{V}_Y^T\mathbf{V}_Y\mathbf{V}_Y^T\mathbf{X}\mathbf{w}_X \\ &= \mathbf{X}^T\mathbf{N}_Y\mathbf{N}_Y^{-1}\mathbf{V}_Y\mathbf{V}_Y^T\mathbf{N}_Y^{-1}\mathbf{N}_Y\mathbf{X}\mathbf{w}_X \\ &= \mathbf{X}^{*T}\mathbf{C}\mathbf{X}^* \\ &= \mathbf{S}_B^*\mathbf{w}_X, \end{aligned}$$

where $\mathbf{X}^* = \mathbf{N}_Y\mathbf{X}$. Let $\mathbf{x}_i^*$, denote the i^{th} row of the matrix $\mathbf{X}^*$.

The key observation to be made here is that the matrix $\mathbf{S}_B^*$ is closely related to the between group sum of squares for the *uncentered* data matrix $\mathbf{X}^*$. To see how $\mathbf{S}_B^*$ is related to the between group sum of squares consider the following: let

$$\begin{aligned} \mathbf{m}_1^* &= \mathbf{X}^T\mathbf{N}_Y\mathbf{f}_1 = \mathbf{X}^{*T}\mathbf{f}_1, \\ \mathbf{m}_2^* &= \mathbf{X}^T\mathbf{N}_Y\mathbf{f}_2 = \mathbf{X}^{*T}\mathbf{f}_2, \end{aligned} \tag{4.21}$$

where $\mathbf{f}_1 = (f_{11}, \ldots, f_{1n})^T$ and $\mathbf{f}_2 = (f_{21}, \ldots, f_{2n})^T$, and

$$f_{1i} = \begin{cases} \frac{1}{n_1} & \text{if } \mathbf{y}_i \text{ is in cluster 1} \\ 0 & \text{otherwise.} \end{cases}$$

$$f_{2i} = \begin{cases} \frac{1}{n_2} & \text{if } \mathbf{y}_i \text{ is in cluster 2} \\ 0 & \text{otherwise.} \end{cases}$$

From (4.21) it can be seen that $\mathbf{m}_1^*$ and $\mathbf{m}_2^*$ are the group means of the $\mathbf{x}_i^*$'s corresponding to either the first or second cluster in Y space. Letting $\mathbf{m}^*$ be the overall mean of the $\mathbf{x}_i^*$'s we have

$$\begin{aligned} \mathbf{S}_B^* &= n_1 \mathbf{m}_1^* \mathbf{m}_1^{*T} + n_2 \mathbf{m}_2^* \mathbf{m}_2^{*T} \\ &= \frac{n_1 n_2}{n} (\mathbf{m}_1^* - \mathbf{m}_2^*)(\mathbf{m}_1^* - \mathbf{m}_2^*)^T + \frac{1}{n} \mathbf{m}^* \mathbf{m}^{*T}. \end{aligned} \tag{4.22}$$

From (4.22) it can be seen that the only difference between $\mathbf{S}_B^*$ and the standard definition of the between group sum of squares is the term $\frac{1}{n}\mathbf{m}^*\mathbf{m}^{*T}$. This additional term arises as a result of the fact that the $\mathbf{x}_i^*$'s are not centered.

Returning to our earlier discussion, we can rewrite (4.20) as

$$\mathbf{S}_B^* \mathbf{w}_X = \rho_{\mathcal{K}}^2 \mathbf{S}_T \mathbf{w}_X, \tag{4.23}$$

where $\mathbf{S}_T = \mathbf{X}^T\mathbf{X}$ is the total sum of squares, discussed in Section 2.5.2.

The generalized eigenvalue problem in (4.23) is closely related to the generalized eigenvalue problem associated with the Maximum Data Piling (MDP) problem (Ahn and Marron (2009)). In the MDP problem the eigenvector solving the generalized eigenvalue problem

$$\mathbf{S}_B \mathbf{w}_{MDP} = \lambda \mathbf{S}_T \mathbf{w}_{MDP}$$

can be shown to be

$$\mathbf{w}_{MDP} \propto \mathbf{S}_T^{-1}(\mathbf{m}_1 - \mathbf{m}_2).$$

If $d < n$ then the MDP direction vector and the LDA direction vector are the same.

In the IKCCA problem, if we assume that $\frac{1}{n}\mathbf{m}^*\mathbf{m}^{*T}$ in (4.22) is close to enough to zero that it is negligible (see Remark 4.4.4 for a discussion of when this may be a reasonable assumption). Then by similar methods used in proving Theorem 2.5.1 of Section 2.5 the leading eigenvector can be shown to be

$$\mathbf{w}_X^* = \frac{1}{\sqrt{\rho_{\mathcal{K}}^*}}\mathbf{S}_T^{-1}(\mathbf{m}_1^* - \mathbf{m}_2^*),$$

where $\rho_{\mathcal{K}}^* = \frac{n\rho_{\mathcal{K}}}{n_1 n_2}$. From this it can be seen that the IKCCA direction vector $\mathbf{w}_X^*$ will tend to behave quite similarly to the MDP direction vector $\mathbf{w}_{MDP}$.

Putting this all together, we can think of IKCCA, when the NGL kernel is used, as a spectral relaxation of the LDA problem.

Remark 4.4.3. Intuitively the diagonal matrix $\mathbf{N}_Y$ should, in some sense, impose the group structure of the points in Y space on the points in X space. The reason for this is that the elements of $\mathbf{N}_Y$, i.e. the eigenvectors $\mathbf{v}_{y1}$ and $\mathbf{v}_{y2}$, "code", as was discussed in Section 4.3, for the different groups. Thus, even if there is a different group structure in X space, the directions found by IKCCA should tend to cluster points in X space according to how they are distributed in Y space. Extending this line of reasoning one step further, if the NGL kernel is also used in X space then the directions found should incorporate group structure from both spaces. This phenomenon will be illustrated in Section 4.5.

Remark 4.4.4. An interesting observation can be made about the behavior of $\frac{1}{n}\mathbf{m}^*\mathbf{m}^{*T}$

as $n \to \infty$, that is

$$\left|\left|\frac{1}{n}\mathbf{m}^*\mathbf{m}^{*T}\right|\right|_F \to 0,$$

provided the distribution from which the $\mathbf{x}_i$'s are sampled has a finite second moment. This can be shown as follows

$$\begin{aligned}
&\left|\left|\frac{1}{n}\mathbf{m}^*\mathbf{m}^{*T}\right|\right|_F^2 \\
&= \frac{1}{n^2}\mathrm{Tr}(\mathbf{m}^*\mathbf{m}^{*T}\mathbf{m}^*\mathbf{m}^{*T}) \\
&= \frac{1}{n^2}\mathrm{Tr}(\mathbf{m}^{*T}\mathbf{m}^*\mathbf{m}^{*T}\mathbf{m}^*).
\end{aligned}$$

Taking a closer look at $\mathbf{m}^{*T}\mathbf{m}^*$ we have

$$\begin{aligned}
&\mathbf{m}^{*T}\mathbf{m}^* \\
&= \sum_{i=1}^{d}\left(\frac{1}{n}\left[\sum_{j=1}^{n_1}\sqrt{n_1}v_{y1j}x_{ji} + \sum_{j=n_1+1}^{n}\sqrt{n_2}v_{y2j}x_{ji}\right]\right)^2 \\
&\leq \frac{1}{n^2}\sum_{i=1}^{d}\left(n_1\sum_{j=1}^{n_1}v_{y1j}^2 + n_2\sum_{j=n_1+1}^{n}v_{y2j}^2\right)\left(\sum_{j=1}^{n_1}x_{ji}^2 + \sum_{j=n_1+1}^{n}x_{ji}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n}x_{ji}^2,
\end{aligned}$$

by the Cauchy-Schwartz inequality. Recall that the terms v_{yij}, $i = 1, 2$, $j = 1, \ldots, n$ are the elements of the leading eigenvectors of $\mathbf{K}_{Y1}$ and $\mathbf{K}_{Y2}$ respectively, therefore $\sum_{i=1}^{n_1} v_{y1i}^2 = \sum_{i=1}^{n_2} v_{y2i}^2 = 1$. Since the x_{ji}'s are mean centered, as $n \to \infty$, we have by the central limit theorem that

$$\frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n}x_{ji}^2 \to \sum_{i=1}^{d}\sigma_i^2, \tag{4.24}$$

where the σ_i's are the population standard deviations. Assuming that the $\mathbf{x}_i$'s have finite

second moments then each of the σ_i^2's in (4.24) will also be finite. Letting $s_i = \sum_{j=1}^{n} x_{ji}^2$ and $\mathbf{s} = (s_1, \ldots, s_d)^T$, provided that $\frac{d}{n} \to 0$ we then have that

$$\begin{aligned} &\frac{1}{n}||\mathbf{m}^*\mathbf{m}^{*T}||_F \\ &\leq \frac{1}{n}||\mathbf{s}^T\mathbf{s}||_F \\ &\to 0, \end{aligned}$$

as we wanted to show. What we can infer from this is that in the limit as $n \to \infty$ (subject to $\frac{d}{n} \to 0$), provided the group structure of the $\mathbf{y}_i$'s is preserved, it is reasonable to assume that $\frac{1}{n}\mathbf{m}^*\mathbf{m}^{*T}$ is negligible.

4.5 Toy Example: Non-standard Data

We now return to the example in Section 3.7 using the NGL kernel (4.15) with weights (4.25). From Figure 4.2 ti can be seen that we are now able to capture the underlying structure of the data, identifying each of the six subpopulations.

$$w_{ij} = \begin{cases} \exp\left\{-\frac{1}{2\sigma_{ij}}||\mathbf{x}_i - \mathbf{x}_j||^2\right\} & \text{if } \mathbf{x}_j \in N_k(\mathbf{x}_i) \\ 0 & \text{otherwise.} \end{cases} \tag{4.25}$$

Here $N_k(\mathbf{x}_i)$ is the symmetric k-neighborhood of the point $\mathbf{x}_i$ (i.e. if $\mathbf{x}_j \in N_k(\mathbf{x}_i)$ then $\mathbf{x}_i \in N_k(\mathbf{x}_j)$).

Looking at plots of the first four eigenvectors (Figures 4.3 and 4.4) in both the smiley face space and the cluster space we can see how the behavior of the eigenvectors causes the segmentation of the data that we observe in Figure 4.2. First we discuss how these figures are generated and then what it is they are telling us

1. Generating an equally spaced dimensional grid spanning the range of values in each space.

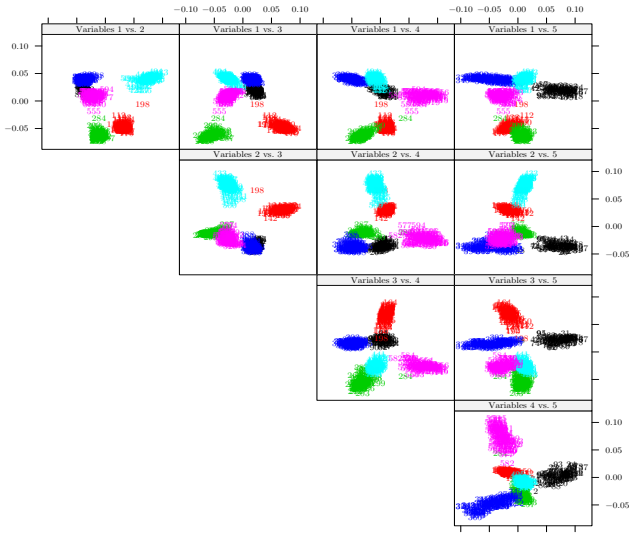

Figure 4.2: *Continuation from the example in Section 3.7. This is a scatterplot matrix of the projections onto the first five IKCCA directions using the kernel in (4.15). Unlike the projections shown in Figure 3.10 here we are able to separate out the six groups.*

2. Calculating the kernel representation and projection of each grid point into IKCC space.

3. Using the projected values to assign color intensities to each point in the grid of each space (blue for negative values, red for positive values).

4. Plotting the grid and for each point using the colors calculated from the previous step.

The important thing to note in both of these figures is the distribution of positive and negative projected values, and how these are driving the segmentation which we observe in Figure 4.2. For example in Figure 4.3 the first canonical variate segments out one of the faces (red) from the other (blue).

4.6 Performance on Real Data

Using the same kernel as in (4.15) we now look at the performance of IKCCA in the receptor ligand matching problem. Figure 4.5 shows the performance of our method which

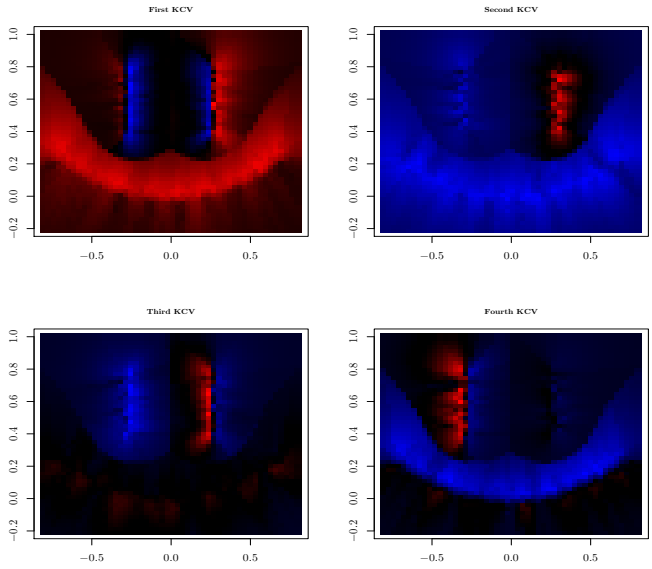

Figure 4.3: *A plot of the first four indefinite kernel canonical direction vectors in the smiley face space from the example in Section 3.7 using the kernel in (4.15). These plots allow us to visualize how the canonical vectors separate out each of the clusters.*

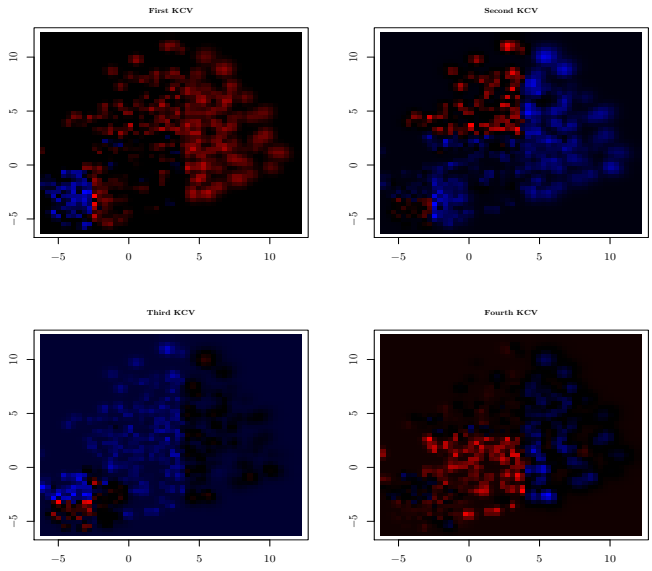

Figure 4.4: *A plot of the first four indefinite kernel canonical directions vectors in the cluster space from the example in Section 3.7 using the kernel in (4.15).*

has an average rank of approximately 4.5 (red vertical line) which is a large improvement over the previously described methods. Here the orange line corresponds to the RBF kernel (rank of 7.5), the blue line corresponds to standard CCA (rank of 10.1) and the

green line corresponds to the performance of the method from Oloff *et al.* (2006) (rank of 18.1).

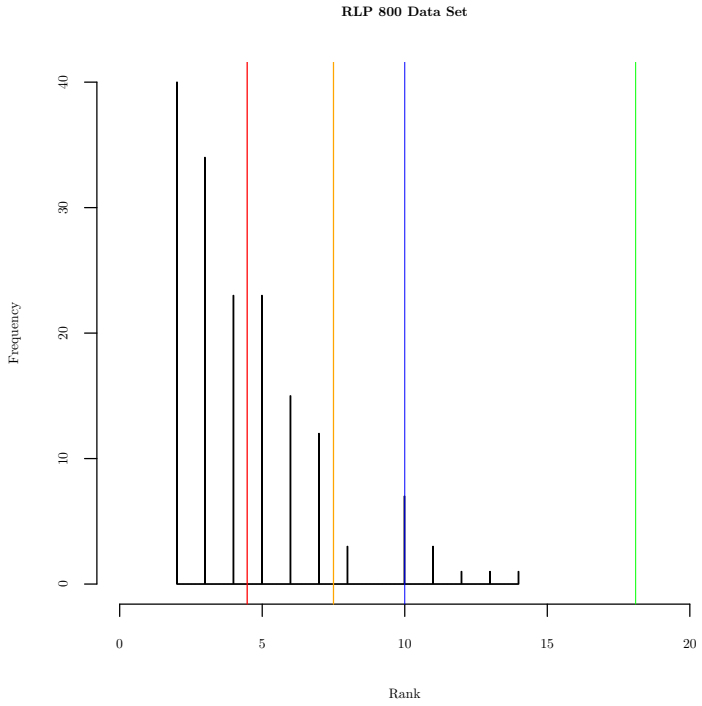

Figure 4.5: *The RLP 800 data set. The red line corresponds to IKCCA, the orange line corresponds to KCCA, the blue line corresponds to CCA and the green line corresponds to the method from Oloff* et al. *(2006).*

Figure 4.6 shows the extension from the RLP 800 data to the WDI. Once again our method, IKCCA, highlighted in red, has a much improved average performance, approximately 30, over previous methods. Standard KCCA has an average performance of 55, linear CCA has an average performance of 67, and the method from Oloff *et al.* (2006) has an average performance of 310.

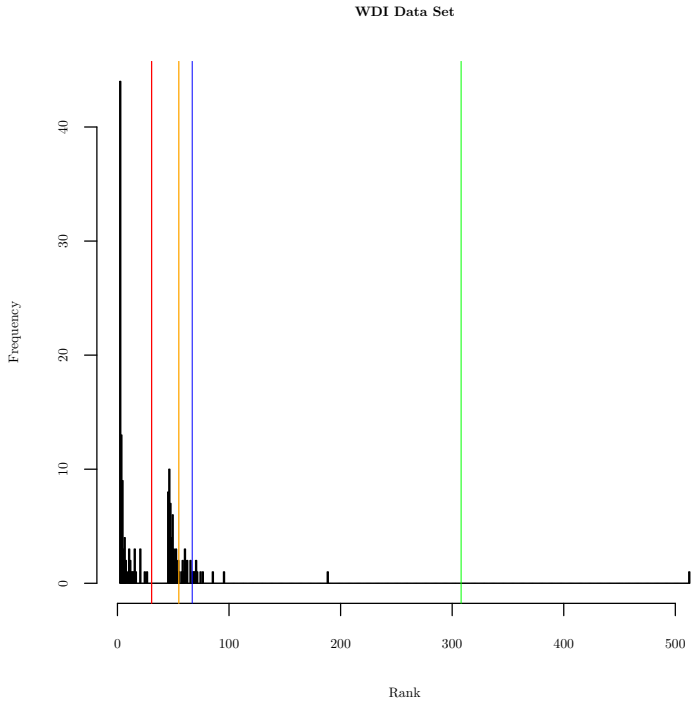

Figure 4.6: *The WDI data set. The red line corresponds to our method using IKCCA, the orange line corresponds to KCCA, the blue line corresponds to CCA and the green line corresponds to the method from Oloff* et al. *(2006).*

CHAPTER 5

HDLSS Asymptotics

A new challenge encountered in a large number of fields (including biology, signal processing and image analysis) is the relatively large number of covariates as compared to the number of observations. This is referred to as the high dimension low sample size (HDLSS) problem, Hall *et al.* (2005), Ahn *et al.* (2007), Lee (2007). The HDLSS problem has led to an interest in studying asymptotics from the standpoint of allowing the number of dimensions, d, to grow.

Amongst the many subjects studied in multivariate asymptotics, a great deal of work has focused on studying the eigenvalues and eigenvectors of sample covariance matrices, i.e. PCA (Anderson (2003), Muirhead (1982)). Classical asymptotics deals with the case where the sample size tends to infinity with the dimension fixed. In the case of the latter most of these studies make use of the fact that the sample covariance matrix is a good approximation of the population covariance. However, with $d \gg n$ this is usually no longer the case.

In studies where d is allowed to go to infinity there are three scenarios which are typically considered:

1. In the first case, which we refer to as the Low Dimension High Sample Size (LDHSS) problem, $d \ll n$, both d and n go to infinity, and $\frac{d}{n} \to 0$. These problems are similar to conventional asymptotics where $n \to \infty$.

2. In this case sample size and dimensionality grow together, in the sense that $\frac{d}{n} \to c$

for some constant c. Bai and Yin (1993), Paul (2005) and Johnstone and Lu (2004) have studied this type of asymptotic behavior. Some work has been done which looks at the case where d grows with some power of n. For example, Portnoy (1984) and Portnoy (1988) study the case where d grows as $\sqrt{n}$. This type of scenario will be referred to as High Dimension High Sample Size (HDHSS).

3. In this setting the sample size is fixed and the dimensionality is allowed to grow, in the sense that $\frac{d}{n} \to \infty$. In the case of n fixed and $d \to \infty$, Hall *et al.* (2005), explored the geometric structure of HDLSS data. In Ahn *et al.* (2007) conditions were found under which the first eigenvector of the sample covariance matrix converges consistently to its population counterpart. In this paper the population covariance matrix is structured such that the leading eigenvalue is considerably larger than the remaining eigenvalues. They also show that when the population covariance matrix does not have this extreme aspherical structure, the sample eigenvalues tend to behave as though they are from a spherical Gaussian distribution.

An important distinction needs to be made here between the aforementioned works and the work done in this dissertation. Here we turn our focus away from the eigen-analysis of the covariance matrix of a single set of variables (i.e. PCA) to the SVD-analysis of the correlation between *two* sets of variables, or the "cross-correlation" matrix. Similar work was done in Lee (2007) where the behavior of the covariance between two sets of variables was studied, also known as the "cross-covariance" matrix.

We also look to study the HDLSS problem in the context of KCCA where we show that high dimensionality can potentially lead to spurious results if not handled in an appropriate manner.

In Section 5.1 we begin with a review of previous work which primarily focuses on the HDHSS and HDLSS asymptotic behavior of the eigenvalues and eigenvectors of the sample covariance and cross-covariance matrices. Finally in Section 5.2 we turn our attention to the asymptotic behavior of CCA in the HDLSS setting. In this section we

discuss conditions under which we have consistent and strongly inconsistent convergence in the sample canonical correlations and vectors. We also present conditions where we have convergence in distribution in the canonical correlations.

5.1 Asymptotics of the Sample Covariance and Cross-Covariance Matrices

5.1.1 Asymptotics of the Sample Covariance Matrices

Suppose we have the data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times d}$, where the $\mathbf{x}_i$'s are $i.i.d.$ observations with mean 0 and covariance $\Sigma \in \mathbb{R}^{d \times d}$. Define the sample covariance matrix as

$$\mathbf{S} = \frac{1}{n}\mathbf{X}^T\mathbf{X}.$$

Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_r$ be the eigenvalues of $\mathbf{S}$ where $r = \text{rank}(\mathbf{S})$. Note that the data matrix $\mathbf{X}$ has not been mean centered, this form is commonly used in studying high-dimensional random matrices.

In the following sections we discuss some HDHSS and HDLSS asymptotic results primarily related to the eigenvalues and eigenvectors of the sample covariance matrix $\mathbf{S}$.

HDHSS Asymptotics

In this section we provide a summary of results analyzing the behavior of the sample covariance matrix when both the sample size and the dimensions are allowed to go to infinity so that $\frac{d}{n} \to \gamma \in (0, 1]$.

Spherical Distribution:

In this section we assume that the population covariance matrix $\Sigma = \mathbf{I}_d$. The empirical distribution of eigenvalues, frequently referred to as the Empirical Spectral Distribution,

is defined as

$$F_d(x) = \frac{1}{d} \times \{ \text{ number of } \lambda_i' s \leq x \}, \ i = 1, \ldots, d.$$

The limiting spectral distribution of F_d was first obtained by Marcenko and Pasture (1967). F_d converges the Marčenko and Pasture distribution F with probability density function

$$f(x) = F'(x) = \begin{cases} (2\pi\gamma x)^{-1}\sqrt{(x-a)(b-x)} & a < x < b \\ 0 & \text{otherwise,} \end{cases}$$

where $a = (1-\sqrt{\gamma})^2$ and $b = (1+\sqrt{\gamma})^2$. When $\gamma > 1$, this distribution has an additional Dirac measure at $x = 0$ of mass $1 - \frac{1}{c}$. The survey paper by Bai (1999) provides a comprehensive review on the spectral distribution.

Up to now our results have focused on the asymptotic behavior of the distribution of sample eigenvalues, λ_i, $i = 1, \ldots, n$. We now turn our attention to the asymptotic properties of each eigenvalue. Specifically we look at the behavior of the eigenvalues lying around the edge of the support of the distribution F, i.e. the largest and smallest eigenvalues.

Studies on the asymptotic behavior of the largest eigenvalue have been conducted by Geman (1980), Yin *et al.* (1988), Silverstein (1989) and Johnstone (2001). Geman (1980) show that for a spherical Gaussian distribution, the largest sample eigenvalue converges to the edge of the support of F,

$$\hat{\lambda}_1 \xrightarrow{a.s.} (1+\sqrt{\gamma})^2, \tag{5.1}$$

where $\xrightarrow{a.s.}$ denotes almost sure convergence. The smallest eigenvalue has also been studied extensively (Bai and Yin (1993) and Silverstein (1985)). Analogous to the largest sample eigenvalue, the smallest sample eigenvalue has been shown to converge to the lower edge

of the support of F,

$$\hat{\lambda}_{min} \overset{a.s.}{\to} (1 - \sqrt{\gamma})^2. \tag{5.2}$$

These results have been generalized to the non-Gaussian case by Yin *et al.* (1988) assuming finite fourth moments.

For the Gaussian case, Johnstone (2001) derived the limiting distribution of the largest sample eigenvalue, $\hat{\lambda}_1$. Specifically he showed that if $\hat{\lambda}_1$ was centered by

$$\mu_d = (\sqrt{n-1} + \sqrt{d})^2$$

and scaled by

$$\sigma_d = (\sqrt{n-1} + \sqrt{d}) \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{d}} \right)^{\frac{1}{3}},$$

then $\hat{\lambda}_1$ converges in distribution to the Tracy-Widom law of order 1 (Tracy and Widom (1996)).

Spiked Data

In many real world applications the assumption that the data follow a spherical distribution may not be accurate. Among the various approaches to studying non-spherical population models, the *spiked population* model, named by Johnstone (2001) is of particular interest. This is due in part to the observation that in many examples, such as speech recognition (Johnstone (2001), Buja *et al.* (1995)), wireless communication (Telatar (1999)), and statistical learning (Hoyle and Rattray (2004)) there are typically a few "larger" sample eigenvalues which are distinct from the rest. The spiked population model assumes that all but finitely many eigenvalues of the population covariance matrix

are one. The population covariance matrix is assumed to take the form

$$\Sigma = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M, 1, \ldots, 1), \tag{5.3}$$

where $\lambda_1 > \lambda_2 > \ldots > \lambda_M > 1$. The almost sure convergence of the largest eigenvalues in the spike population model was shown by Paul (2005) and Baik and Silverstein (2006). Paul (2005) examines the behavior of the eigenvalues assuming that the data are normally distributed and derives the asymptotic distribution of the largest eigenvalue and examines the behavior of the corresponding eigenvector. Baik and Silverstein (2006) provide results on the almost sure limits of the largest and smallest eigenvalues in both the real and complex non-Gaussian cases.

Baik and Silverstein (2006) and Paul (2005) also observed that under the spiked population model if $\lambda_1 < 1 + \sqrt{\gamma}$ then

$$\hat{\lambda}_1 \overset{a.s.}{\to} (1 + \sqrt{\gamma})^2 \tag{5.4}$$

and if $\lambda_{min} > 1 - \sqrt{\gamma}$ then

$$\hat{\lambda}_{min} \overset{a.s.}{\to} (1 - \sqrt{\gamma})^2, \tag{5.5}$$

provided $\gamma \in (0, 1)$. Note that here the limits in (5.4) and (5.5) are the same as the corresponding quantities in (5.1) and (5.2). In other words, when the largest (or smallest) population eigenvalue is not "different enough" from one, the corresponding sample eigenvalue, asymptotically, will behave as though it came from a population characterized by an identity covariance. This behavior, referred to as "phase transition", is an important observation made in both works. A similar phenomenon is also observed in the HDLSS setting (Ahn *et al.* (2007)).

The phase transition phenomenon is also observed in the sample eigenvectors (Paul

(2005)). Define $\mathbf{v}_1, \ldots, \mathbf{v}_d$ to be the eigenvectors of the population covariance $\Sigma = \mathbf{I}_d$, and $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_d$ the eigenvectors of the sample covariance matrix $\mathbf{S}$. It was shown in Paul (2005), that the following results hold when $\frac{d}{n} \to \gamma \in (0, 1)$:

If $\lambda_i \leq 1 + \sqrt{\gamma}$ then,

$$\langle \mathbf{v}_i, \hat{\mathbf{v}}_i \rangle \xrightarrow{a.s.} 0 \text{ as } n \to \infty.$$

If $\lambda_i > 1 + \sqrt{\gamma}$ and is of multiplicity one, then

$$|\langle \mathbf{v}_i, \hat{\mathbf{v}}_i \rangle| \xrightarrow{a.s.} \sqrt{\left(1 - \frac{\gamma}{(\lambda_i - 1)^2}\right) / \left(1 - \frac{\gamma}{\lambda_i - 1}\right)} \text{ as } n, d \to \infty. \tag{5.6}$$

The implication of (5.6) is that if the leading population eigenvalue is not much bigger than one than its corresponding eigenvector is *strongly inconsistent* to the population eigenvector in the sense that the two vectors are orthogonal.

HDLSS Asymptotics

We now turn our attention to the case where the number of observations n is fixed and d is allowed to go to infinity. In the following subsections we discuss the geometrical representation of HDLSS data and the HDLSS asymptotics associated with the sample covariance matrix.

Geometric Representation

The geometrical representation of HDLSS data was studied by Hall *et al.* (2005). Suppose $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independent random variables drawn from the Gaussian distribution with mean zero and covariance matrix $\mathbf{I}_d$. Since the sum of the squares of the entries of $\mathbf{z}_i$ has the Chi-square distribution with d degrees of freedom, it can be shown that

$$||\mathbf{z}_i - \mathbf{z}_j'|| = (2d)^{\frac{1}{2}} + O_p(1),$$

as $d \to \infty$. What this tells us is that for large enough d a sample of n standard normal random variables will tend to lie at the vertices of a regular n-simplex in $\mathbb{R}^d$. Note that the data vectors tend to have a deterministic distance apart. Hall *et al.* (2005) also studied the geometric representation of HDLSS data in the context of classification. In that study they obtained some insight into the limiting behavior of several classification methods such as support vector machines (Cristianini and Shawe-Taylor (2000)) and distance weighted discrimination (Marron *et al.* (2008)).

Dual Covariance Matrices

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ where $\mathbf{x}_i \sim N_d(0, \Sigma_d)$, for $d = 1, 2, \ldots$. The $n \times n$ dual sample covariance matrix is defined as

$$\mathbf{S}_D = \frac{1}{n}\mathbf{X}\mathbf{X}^T.$$

Ahn *et al.* (2007) studied conditions under which the dual sample covariance matrix converges to the identity, $\mathbf{I}_n$ as $d \to \infty$. These results were generalized to an arbitrary distribution under some general assumptions on the moments of the data by Jung and Marron (2009). In their analysis they also presented results on the consistency and strong inconsistency of the sample eigenvectors.

In our discussion of the HDLSS asymptotics of CCA we provide a more detailed discussion on the behavior of the dual sample covariance matrix which we refer to as the kernel matrix (Section 5.2.6).

5.1.2 HDLSS Asymptotics of the Sample Cross-Covariance Matrices

Consider a data set consisting of n paired multivariate vectors,

$$\{(\mathbf{x}_i, \mathbf{y}_i) | \mathbf{x}_i \in \mathbb{R}^{d_X}, \mathbf{y}_i \in \mathbb{R}^{d_Y}, i = 1, \ldots, n\},$$

where $(\mathbf{x}_i, \mathbf{y}_i) \sim N(\mathbf{0}, \Sigma)$. Here $\mathbf{0} \in \mathbb{R}^{d_X+d_Y}$ and

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{XX} & \Sigma_{XY} \\ \hline \Sigma_{YX} & \Sigma_{YY} \end{array} \right). \tag{5.7}$$

Define $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times d_X}$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T \in \mathbb{R}^{n \times d_Y}$ and the corresponding sample mean matrices as $\bar{X}$ and $\bar{Y}$. Note from here on we assume that the data matrices $\mathbf{X}$ and $\mathbf{Y}$ have been mean centered.

In Section 5.1.1 our discussion focused on the behavior of the eigenvalues and eigenvectors of the sample covariance matrix. In this section we turn our attention to the sample *cross-covariance* matrix

$$\mathbf{S}_{XY} = \frac{1}{n}\mathbf{X}^T\mathbf{Y}.$$

In Lee (2007), under specific assumptions on the structure of the covariance matrix Σ, the HDLSS asymptotics of the singular values and vectors of the sample cross-covariance matrix were studied. In this study conditions were established showing convergence in distribution of the largest sample singular value to a random quantity. In addition, consistency and strong inconsistency results were established for the leading sample singular vectors.

Remark 5.1.1. An important concept discussed in Lee (2007) is the construction of the population covariance matrix Σ. Because Σ potentially contains off-diagonal terms (i.e. Σ_{XY} in (5.7)), greater care needs to be taken in order to ensure that it is positive semi-definite. One way in which positive semi-definiteness can be guaranteed is to use the so-called factor matrices which are defined as

$$\mathbf{F} = \left(\begin{array}{c|c} \mathbf{F}_{XX} & \mathbf{F}_{XY} \\ \hline \mathbf{F}_{YX} & \mathbf{F}_{YY} \end{array} \right),$$

so that

$$\Sigma = \mathbf{F}^2.$$

Since $\mathbf{F}^2$ is positive semi-definite this ensures that Σ is positive semi-definite. The components of $\mathbf{F}$, i.e. $\mathbf{F}_{XX}$, $\mathbf{F}_{YY}$ and $\mathbf{F}_{XY}$ are meant to capture the type of joint structure which we would like to observe in Σ. This construction will be play a central role in our discussion of CCA in the HDLSS setting.

5.2 HDLSS Asymptotics of CCA

In Section 5.1 our discussion primarily focused on studying the asymptotic behavior of the sample covariance ($\mathbf{S}_{XX}$ and $\mathbf{S}_{YY}$) and cross-covariance ($\mathbf{S}_{XY}$) matrices and their eigenvalues and eigenvectors. In the following section we move our attention toward studying the population, sample and sample kernel cross-correlation matrices in the HDLSS setting. In Sections 5.2.1, 5.2.2 and 5.2.3 we introduce the population, sample and kernel sample cross-correlation matrices. In Sections 5.2.5 and 5.2.6 we study the asymptotic behavior of the sample and sample kernel cross-correlation matrices, respectively, in the HDLSS setting.

5.2.1 The Population Cross-Correlation Matrix

Recall from Section 2.1 that the canonical correlations and directions can be found by solving the generalized eigenvalue problem

$$\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{w}_X = \rho_{\mathcal{H}}^2\Sigma_{XX}\mathbf{w}_X.$$

An alternative representation of the above problem which is easier to study and allows us to solve for the canonical correlations and vectors simultaneously is the *cross-correlation* matrix which we now derive. Beginning with the generalized eigenvalue problem above

we have

$$\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}(\Sigma_{XX}^{\frac{1}{2}}\mathbf{w}_X) = \rho_{\mathcal{H}}^2\Sigma_{XX}^{\frac{1}{2}}(\Sigma_{XX}^{\frac{1}{2}}\mathbf{w}_X).$$

Letting $\mathbf{w}_X^* = \Sigma_{XX}^{\frac{1}{2}}\mathbf{w}_X$ and multiplying the left and right-hand sides by $\Sigma_{XX}^{-\frac{1}{2}}$ gives us

$$\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}\mathbf{w}_X^* = \rho_{\mathcal{H}}^2\mathbf{w}_X^*.$$

The matrix $\mathcal{R}_{XY} = \Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-\frac{1}{2}}$, is commonly referred to as the population cross-correlation matrix. Substituting in $\mathcal{R}_{XY}$ we have

$$\mathcal{R}_{XY}\mathcal{R}_{YX}\mathbf{w}_X^* = \rho_{\mathcal{H}}^2\mathbf{w}_X^*.$$

Put in this form it can be seen that the SVD of the cross-correlation matrix provides us with both the canonical correlation $\rho_{\mathcal{H}}$ and the *scaled*, canonical vectors $\mathbf{w}_X^*$ and $\mathbf{w}_Y^*$ (in contrast to the *unscaled* canonical vectors $\mathbf{w}_X$ and $\mathbf{w}_Y$). Both notions are useful for understanding the theory developed in Section 5.2.5.

Finding the sample counterpart of the cross-correlation matrix is not as straightforward. Because of the fact that we have $d \gg n$ the covariance matrices $\mathbf{S}_{XX}$ and $\mathbf{S}_{YY}$ are singular and therefore cannot be directly inverted, we deal with this by using an approach motivated by our previous discussion of regularized CCA as well as kernels and the kernel trick.

5.2.2 The Sample Cross-Correlation Matrix

Recall that the Lagrangian of the regularized CCA problem (see (2.21) in Section 2.3) is

$$\begin{aligned} L(\rho_X, \rho_Y, \hat{\mathbf{w}}_X, \hat{\mathbf{w}}_Y) =& \frac{1}{n}\hat{\mathbf{w}}_X^T\mathbf{X}^T\mathbf{Y}\hat{\mathbf{w}}_Y - \frac{\rho_X}{2}(\frac{1}{n}\hat{\mathbf{w}}_X^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_X + \kappa\hat{\mathbf{w}}_X^T\hat{\mathbf{w}}_X - 1) \\ & - \frac{\rho_Y}{2}(\frac{1}{n}\hat{\mathbf{w}}_Y^T\mathbf{Y}^T\mathbf{Y}\hat{\mathbf{w}}_Y + \kappa\hat{\mathbf{w}}_Y^T\hat{\mathbf{w}}_Y - 1), \end{aligned}$$

where ρ_X and ρ_Y are Lagrange multipliers and κ is the regularization parameter. The hats are meant to denote the respective variables sample counterpart.

Recall from our discussion in Section 2.3 on RCCA that the solution to the above optimization problem leads to the generalized eigenvalue problem

$$\mathbf{S}_{XY}(\mathbf{S}_{YY} + \kappa\mathbf{I}_n)^{-1}\mathbf{S}_{YX}\hat{\mathbf{w}}_X = \rho^2_{\mathcal{H}}(\mathbf{S}_{XX} + \kappa\mathbf{I}_n)\hat{\mathbf{w}}_X.$$

In a similar fashion to our calculation of the population cross-correlation matrix we have for the sample counterpart that

$$\mathbf{R}_{XY}\mathbf{R}_{YX}\hat{\mathbf{w}}^*_X = \hat{\rho}\hat{\mathbf{w}}^*_X,$$

where $\mathbf{R}_{XY} = (\mathbf{S}_{XX} + \kappa\mathbf{I}_n)^{-\frac{1}{2}}\mathbf{S}_{XY}(\mathbf{S}_{YY} + \kappa\mathbf{I}_n)^{-\frac{1}{2}}$ is the sample cross-correlation matrix and $\hat{\mathbf{w}}^*_X = (\mathbf{S}_{XX} + \kappa\mathbf{I}_n)^{\frac{1}{2}}\hat{\mathbf{w}}_X$ and $\hat{\mathbf{w}}^*_Y = (\mathbf{S}_{YY} + \kappa\mathbf{I}_n)^{\frac{1}{2}}\hat{\mathbf{w}}_Y$ are scaled sample canonical vectors.

5.2.3 The Sample Kernel Cross-Correlation Matrix

Because we are letting the number of dimensions d go to infinity, rather than only looking at the sample cross-correlation matrix it will also be useful to look at its kernelized variant since n in this setting is fixed.

Recall that because $\hat{\mathbf{w}}_X$ and $\hat{\mathbf{w}}_Y$ fall into the span of the column spaces of $\mathbf{X}$ and $\mathbf{Y}$ respectively, they can be re-written as

$$\hat{\mathbf{w}}_X = \mathbf{X}^T\alpha_X$$
$$\hat{\mathbf{w}}_Y = \mathbf{Y}^T\alpha_Y.$$

The Lagrangian is thus modified to be

$$\begin{aligned} L(\rho_X, \rho_Y, \alpha_X, \alpha_Y) &= \alpha_X^T \frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\alpha_Y - \frac{\rho_X}{2}(\alpha_X^T \frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\alpha_X + \kappa\alpha_X^T\mathbf{X}\mathbf{X}^T\alpha_X - 1) \\ &\quad - \frac{\rho_Y}{2}(\alpha_Y^T \frac{1}{n}\mathbf{Y}\mathbf{Y}^T\mathbf{Y}\mathbf{Y}^T\alpha_Y + \kappa\alpha_Y^T\mathbf{Y}\mathbf{Y}^T\alpha_Y - 1) \\ &= \alpha_X^T\mathbf{K}_X\mathbf{K}_Y\alpha_Y - \frac{\rho_X}{2}(\alpha_X^T\mathbf{K}_X^2\alpha_X + \kappa\alpha_X^T\mathbf{K}_X\alpha_X - 1) \\ &\quad - \frac{\rho_Y}{2}(\alpha_Y^T\mathbf{K}_Y^2\alpha_Y + \kappa\alpha_Y^T\mathbf{K}_T\alpha_Y - 1), \end{aligned}$$

where $\mathbf{K}_X = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ and $\mathbf{K} = \frac{1}{n}\mathbf{Y}\mathbf{Y}^T$ (in the HDLSS literature these matrices are sometimes referred to as the dual sample covariance matrices). Note that this particular representation of the Lagrangian corresponds to the regularized KCCA problem with $\alpha_X^T(\mathbf{K}_X^2 + \kappa\mathbf{K}_X)\alpha_X = 1$ as the constraint rather than $\alpha_X^T(\mathbf{K}_X^2 + \kappa\mathbf{I}_n)\alpha_X = 1$ which was discussed in Section 3.3. Continuing, we know that $\rho_{\mathcal{H}} = \rho_X = \rho_Y$ and that the solutions to α_X and α_Y are

$$\begin{aligned} \alpha_X &= \frac{1}{\rho_{\mathcal{H}}}(\mathbf{K}_X^2 + \kappa\mathbf{I}_n)^{-1}\mathbf{K}_Y\alpha_Y, \\ \alpha_Y &= \frac{1}{\rho_{\mathcal{H}}}(\mathbf{K}_Y^2 + \kappa\mathbf{I}_n)^{-1}\mathbf{K}_X\alpha_X. \end{aligned}$$

The derivative of the Lagrangian with respect to α_X is

$$\frac{\partial L}{\partial \alpha_X} = \mathbf{K}_X\mathbf{K}_Y\alpha_Y - \rho_{\mathcal{H}}(\mathbf{K}_X + \kappa\mathbf{I}_n)\mathbf{K}_X\alpha_X = 0.$$

Plugging the solution for α_Y into the above equation and re-arranging terms gives us

$$\mathbf{K}_X\mathbf{K}_Y(\mathbf{K}_Y + \kappa\mathbf{I}_n)^{-1}\mathbf{K}_X\alpha_X = \rho_{\mathcal{H}}^2(\mathbf{K}_X + \kappa\mathbf{I}_n)\mathbf{K}_X\alpha_X$$

Letting $\alpha_X^* = \mathbf{K}_X^{\frac{1}{2}}(\mathbf{K}_X + \kappa\mathbf{I}_n)^{\frac{1}{2}}\alpha_X$ and re-arranging terms we have

$$\mathbf{K}_X\mathbf{K}_Y^{\frac{1}{2}}(\mathbf{K}_Y + \kappa\mathbf{I}_n)^{-1}\mathbf{K}_Y^{\frac{1}{2}}\mathbf{K}_X^{\frac{1}{2}}(\mathbf{K}_X + \kappa\mathbf{I}_n)^{-\frac{1}{2}}\alpha_X^* = \rho_{\mathcal{H}}^2\mathbf{K}_X^{\frac{1}{2}}(\mathbf{K}_X + \kappa\mathbf{I}_n)^{\frac{1}{2}}\alpha_X^*.$$

Finally, multiplying both sides by $(\mathbf{K}_X + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{K}_X^{-\frac{1}{2}}$ gives us

$$(\mathbf{K}_X + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{K}_X^{\frac{1}{2}} \mathbf{K}_Y^{\frac{1}{2}} (\mathbf{K}_Y + \kappa \mathbf{I}_n)^{-1} \mathbf{K}_Y^{\frac{1}{2}} \mathbf{K}_X^{\frac{1}{2}} (\mathbf{K}_X + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \alpha_X^* = \rho_{\mathcal{H}}^2 \alpha_X^*. \tag{5.8}$$

Letting $\mathbf{R}_{YX}^K = (\mathbf{K}_Y + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{K}_Y^{\frac{1}{2}} \mathbf{K}_X^{\frac{1}{2}} (\mathbf{K}_X + \kappa \mathbf{I}_n)^{-\frac{1}{2}}$, we can re-write (5.8) as

$$\mathbf{R}_{XY}^K \mathbf{R}_{YX}^K \alpha_X^* = \rho_{\mathcal{H}}^2 \alpha_X^*.$$

From this we can see that the SVD of $\mathbf{R}_{XY}^K$ gives us the regularized canonical correlations and scaled kernel canonical vectors. We will refer to this matrix as the sample kernel cross-correlation matrix.

As we develop theoretical results for these various examples, in what follows we assume that the regularization parameter κ appears in the asymptotic form

$$\kappa \sim d^\gamma,$$

in the sense that $\frac{\kappa}{d^\gamma} \to c \in (0, \infty)$ as $d \to \infty$, where $\gamma \geq 0$. The regularization parameter plays a critical role in the consistency and strong inconsistency of the canonical correlations and vectors depending on the value of γ.

5.2.4 Population Models

In order to better understand the behavior of CCA in the HDLSS setting, we consider several population models meant to capture a broad range of behaviors in the marginal and joint distributions of the data. As in Lee (2007) we assume for the sake of notational simplicity that $d = d_X = d_Y$.

1. **Uncorrelated Spiked Covariance Model** (Model 1). The factor matrices (see Section 5.1.2, Remark 5.1.1 for a discussion on factor matrices) are $\mathbf{F}_{XX} = \mathbf{F}_{YY} = \text{diag}(d^\alpha, 1, \ldots, 1)$ and $\mathbf{F}_{XY} = \text{diag}(0, \ldots, 0)$. This model is meant to study the

behavior of CCA when there is no correlation between data sets.

2. **Spiked Covariance/Cross-Covariance Model** (Model 2). We consider two parameterizations of this model (which give different results) that we will refer to as $S1$ and $S2$. These models explore the effect of relative signal strength from the spike in the covariance matrix relative to the spike in the cross-covariance matrix.

   $S1$: $\mathbf{F}_{XX} = \mathbf{F}_{YY} = \text{diag}(d^{\alpha}, 1, \ldots, 1)$ and $\mathbf{F}_{XY} = \text{diag}(d^{\alpha\beta}, 0, \ldots, 0)$.

   $S2$: $\mathbf{F}_{XX} = \mathbf{F}_{YY} = \text{diag}(d^{\alpha} + d^{\beta}, 1, \ldots, 1)$ and $\mathbf{F}_{XY} = \text{diag}(d^{\beta}, 0, \ldots, 0)$.

3. **Constant Covariance/Spiked Cross-Covariance Model** (Model 3). $\mathbf{F}_{XX} = \mathbf{F}_{XX} = \text{diag}(1, \ldots, 1)$ and $\mathbf{F}_{XY} = \text{diag}(d^{\alpha}, 0, \ldots, 0)$. This model explores the effect of having a spike in only the cross-covariance matrix.

We now provide some details related to the eigenvalues, canonical correlations and canonical vectors for each of these population models.

For the purposes of our calculations we re-express the (centered) data matrices $\mathbf{X}$ and $\mathbf{Y}$ based on their joint distribution as

$$\begin{pmatrix} \mathbf{X} & \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_X & \mathbf{Z}_Y \end{pmatrix} \begin{pmatrix} \mathbf{F}_{XX} & \mathbf{F}_{XY} \\ \mathbf{F}_{YX} & \mathbf{F}_{YY} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_X\mathbf{F}_{XX} + \mathbf{Z}_Y\mathbf{F}_{YX} & \mathbf{Z}_Y\mathbf{F}_{YY} + \mathbf{Z}_X\mathbf{F}_{XY} \end{pmatrix},$$

where $\mathbf{Z}_X = (\mathbf{z}_{x1}, \ldots, \mathbf{z}_{xd})_{(n\times d)}$ and $\mathbf{Z}_Y = (\mathbf{z}_{y1}, \ldots, \mathbf{z}_{yd})_{(n\times d)}$.

1. *Model 1*

$$\mathbf{X} = \begin{pmatrix} d^{\alpha}z_{x1}^i & z_{x2}^i & \cdots & z_{xd}^i \end{pmatrix}_{i=1}^{n}$$
$$\mathbf{Y} = \begin{pmatrix} d^{\alpha}z_{y1}^i & z_{y2}^i & \cdots & z_{yd}^i \end{pmatrix}_{i=1}^{n}.$$

   In this model we have that $\Sigma_{XX} = \Sigma_{YY} = \text{diag}(d^{2\alpha}, 1, \ldots, 1)$ and $\Sigma_{XY} = \mathbf{0}$. The eigenvalues of Σ_{XX} and Σ_{YY} are $\lambda_X^1 = \lambda_Y^1 = d^{2\alpha}$ and $\lambda_X^i = \lambda_Y^i = 1, i = 2, \ldots, d$.

The corresponding cross-correlation matrix is

$$\mathcal{R}_{XY} = \text{diag}(0, \ldots, 0).$$

Under this model framework the canonical correlations are all 0 and the scaled canonical vectors are any orthonormal basis.

2. *Model 2*

   $S1$:

$$\mathbf{X} = \begin{pmatrix} d^{\alpha} z_{x1}^{i} + d^{\alpha\beta} z_{y1}^{i} & z_{x2}^{i} & \cdots & z_{xd}^{i} \end{pmatrix}_{i=1}^{n}$$
$$\mathbf{Y} = \begin{pmatrix} d^{\alpha} z_{y1}^{i} + d^{\alpha\beta} z_{x1}^{i} & z_{y2}^{i} & \cdots & z_{yd}^{i} \end{pmatrix}_{i=1}^{n}.$$

   In this model we have $\Sigma_{XX} = \Sigma_{YY} = \text{diag}(d^{2\alpha} + d^{2\alpha\beta}, 1, \ldots, 1)$ and $\Sigma_{XY} = \text{diag}(2d^{\alpha(1+\beta)}, 0, \ldots, 0)$. The eigenvalues of Σ_{XX} and Σ_{YY} are $\lambda_X^1 = \lambda_Y^1 = d^{2\alpha} + d^{2\alpha\beta}$ and $\lambda_X^i = \lambda_Y^i = 1, i = 2, \ldots, d$. The corresponding population cross-correlation matrix is

$$\mathcal{R}_{XY} = \text{diag}\left(\frac{2d^{\alpha(1+\beta)}}{d^{2\alpha} + d^{2\alpha\beta}}, 0, \ldots, 0\right).$$

   This is the same population model that was used by Lee (2007). The leading canonical correlation in this model converges to 1 if and only if $\beta = 1$ and it converges to 0 otherwise. If the leading canonical correlation converges to 1 then the leading scaled canonical vectors are $\mathbf{w}_X^{*1} = \mathbf{w}_Y^{*1} = (1, 0, \ldots, 0)^T$ and the remaining scaled canonical vectors are any orthonormal basis which is orthogonal to $\mathbf{w}_X^{*1}$ and $\mathbf{w}_Y^{*1}$. If the leading canonical correlation converges to 0 then the canonical vectors in both the X and Y spaces are any orthonormal basis.

In some sense β can be thought of as controlling the strength of the signal from the joint distribution relative to the marginal distributions. If $\beta = 1$ then the signal from the joint distribution is as strong as the signal of the marginal distributions. If $\beta \neq 1$ then, asymptotically, the signal of the joint distribution is dominated by the signal from one or both of the marginal distributions.

$S2$:

$$\mathbf{X} = \begin{pmatrix} (d^\alpha + d^\beta)z_{x1}^i + d^\beta z_{y1}^i & z_{x2}^i & \cdots & z_{xd}^i \end{pmatrix}_{i=1}^n$$
$$\mathbf{Y} = \begin{pmatrix} (d^\alpha + d^\beta)z_{y1}^i + d^\beta z_{x1}^i & z_{y2}^i & \cdots & z_{yd}^i \end{pmatrix}_{i=1}^n.$$

In this model we have $\Sigma_{XX} = \Sigma_{YY} = \text{diag}((d^\alpha + d^\beta)^2 + d^{2\beta}, 1, \ldots, 1)$ and $\Sigma_{XY} = \text{diag}(2d^\beta(d^\alpha + d^\beta), 0, \ldots, 0)$. The eigenvalues of Σ_{XX} and Σ_{YY} are $\lambda_X^1 = \lambda_Y^1 = (d^\alpha + d^\beta)^2 + d^{2\beta}$ and $\lambda_X^i = \lambda_Y^i = 1, i = 2, \ldots, d$. The corresponding cross-correlation matrix is

$$\mathcal{R}_{XY} = \text{diag}\left(\frac{2d^\beta(d^\alpha + d^\beta)}{(d^\alpha + d^\beta)^2 + d^{2\beta}}, 0, \ldots, 0\right).$$

Note that when $\alpha > \beta$ the leading population canonical correlation converges to 0 and if $\beta > \alpha$ then this value converges to 1. If $\alpha = \beta$ then the canonical correlation value is equal to $\frac{4}{5}$. In some sense α can be thought of as the noise parameter and β the signal parameter, i.e. if there is more signal than noise, as the dimensions go to infinity the signal can still be detected.

If $\beta > \alpha$, i.e. the leading canonical correlation converges to 1, then $\mathbf{w}_X^{*1} = \mathbf{w}_Y^{*1} = (1, 0, \ldots, 0)^T$ and the remaining canonical vectors are any orthonormal basis which is orthogonal to $\mathbf{w}_X^{*1}$ and $\mathbf{w}_Y^{*1}$. Otherwise if $\alpha > \beta$ then the canonical vectors are any orthonormal basis.

3. *Model 3*

$$\mathbf{X} = \begin{pmatrix} z^i_{x1} + d^\alpha z^i_{y1} & z^i_{x2} & \cdots & z^i_{xd} \end{pmatrix}_{i=1}^{n}$$
$$\mathbf{Y} = \begin{pmatrix} z^i_{y1} + d^\alpha z^i_{x1} & z^i_{y2} & \cdots & z^i_{yd} \end{pmatrix}_{i=1}^{n}.$$

In this model we have $\Sigma_{XX} = \Sigma_{YY} = \text{diag}(1+d^{2\alpha}, 1, \ldots, 1)$ and $\Sigma_{XY} = \text{diag}(2d^\alpha, 0, \ldots, 0)$. The eigenvalues of Σ_{XX} and Σ_{YY} are $\lambda^1_X = \lambda^1_Y = 1 + d^{2\alpha}$ and $\lambda^i_X = \lambda^i_Y = 2, \ldots, d$. The associated cross-covariance matrix for this population model is

$$\mathcal{R}_{XY} = \text{diag}\left(\frac{2d^\alpha}{1+d^{2\alpha}}, 0, \ldots, 0\right).$$

The leading canonical correlation is equal to 1 only when $\alpha = 0$, i.e. when there is no spike present, and is 0 otherwise. If the leading canonical correlation is equal to 1 then $\mathbf{w}^{*1}_X = \mathbf{w}^{*1}_Y = (1, 0, \ldots, 0)^T$ and the remaining canonical vectors are any orthonormal basis which is orthogonal to $\mathbf{w}^{*1}_X$ and $\mathbf{w}^{*1}_Y$. If the leading canonical correlation converges to 0 then the canonical vectors are any orthonormal basis.

Remark 5.2.1. We had also originally considered a "Spiked Covariance/Constant Cross-Covariance Model", where the factor matrices were structured as, $\mathbf{F}_{XX} = \mathbf{F}_{YY} = \text{diag}(d^\alpha, 1, \ldots, 1)$ and $\mathbf{F}_{XY} = \text{diag}(1, \ldots, 1)$. However, what we found was that this resulted in exactly the same joint covariance matrix Σ as in Model 3. This happened because the factor matrix $\mathbf{F}$ did *not* correspond to the matrix square root $\Sigma^{\frac{1}{2}}$, which is unique. This can be seen by the following: let

$$\Sigma = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T,$$

be the eigendecomposition of Σ. The matrix square root of Σ is defined as $\Sigma^{\frac{1}{2}} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^T$, which, provided Σ is a positive semi-definite matrix, is unique.

Let $\mathbf{B}$ be any orthonormal basis, we then have that

$$\begin{aligned}\Sigma &= \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{B}\mathbf{V}^T\mathbf{V}\mathbf{B}^T\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^T \\ &= \mathbf{F}\mathbf{F}^T,\end{aligned}$$

where $\mathbf{F} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{B}\mathbf{V}^T$. In general the matrix $\mathbf{F}$ will not be symmetric, however, if the matrix $\mathbf{B}$ is a permutation matrix, then the rows of $\mathbf{\Lambda}^{\frac{1}{2}}$ will be reordered and $\mathbf{F}$ will be symmetric. The result is that without closer inspection the matrix $\mathbf{F}$ may appear to be the matrix square root, while in reality it is not.

Consider the following: let $\mathbf{F}$ be the factor matrix associated with Model 3 and $\mathbf{P}$ a permutation matrix, then there exists a permutation such that

$$\Sigma = \left(\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & d^\alpha & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \hline d^\alpha & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{array}\right) \mathbf{P}\mathbf{P}^T \left(\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & d^\alpha & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \hline d^\alpha & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{array}\right)$$

$$= \left(\begin{array}{cccc|cccc} d^{\alpha} & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \hline 1 & 0 & \cdots & 0 & d^{\alpha} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{array}\right) \left(\begin{array}{cccc|cccc} d^{\alpha} & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \hline 1 & 0 & \cdots & 0 & d^{\alpha} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{array}\right).$$

From this it can be seen that rearranging the rows of $\mathbf{F}$ does not effect the structure of Σ. This is why the behavior of the Spiked Covariance/Constant Cross-Covariance Model does not differ from Model 3.

5.2.5 Asymptotics of the Sample Cross-Correlation Matrix

In this section we study the HDLSS asymptotics of the sample canonical correlations and vectors via the sample cross-correlation matrix discussed in Section 5.2.2. However, before we begin looking at the sample cross-correlation matrix it is necessary to first study the asymptotic behavior of the sample covariance and cross-covariance matrices, $\mathbf{S}_{XX}$, $\mathbf{S}_{YY}$ and $\mathbf{S}_{XY}$ as $d \to \infty$. Lemma 5.2.2 provides some results about the sample covariance and cross-covariance matrices that will be needed in order to study the asymptotic behavior of the sample cross-correlation matrix.

Let the ij^{th} entry of $\mathbf{S}_{XX(d)}$ be denoted by $s^{ij}_{xx(d)}$ and the ij^{th} entry of $\mathbf{S}_{XY(d)}$ be denoted by $s^{ij}_{xy(d)}$, where d is the dimension of the matrix. Define λ^1_X to be the leading eigenvalue of the population covariance matrix Σ_{XX}. The value of λ^1_X will depend on the population model (see Section 5.2.4 for details). Let

$$\tilde{\lambda}^1_X = \lim_{d\to\infty} \frac{1}{\lambda^1_X} s^{11}_{xx},$$

$$\tilde{\lambda}_Y^1 = \lim_{d\to\infty} \frac{1}{\lambda_X^1} s_{yy}^{11},$$
$$\tilde{\lambda}_{XY}^1 = \lim_{d\to\infty} \frac{1}{\lambda_X^1} s_{xy}^{11}.$$

The values of $\tilde{\lambda}_X^1$, $\tilde{\lambda}_Y^1$ and $\tilde{\lambda}_{XY}^1$ will depend on the population model. Note that $\tilde{\lambda}_X^1$, $\tilde{\lambda}_Y^1$ and $\tilde{\lambda}_{XY}^1$ correspond to the limiting eigenvalues or singular value, respectively of the matrices $\frac{1}{\lambda_X^1}\mathbf{S}_{XX}$, $\frac{1}{\lambda_X^1}\mathbf{S}_{YY}$ and $\frac{1}{\lambda_X^1}\mathbf{S}_{XY}$. From here on we suppress the subscript (d).

In Lemma 5.2.2 we will show that

$$\frac{1}{\lambda_X^1} s_{xx}^{11} \xrightarrow{F} \tilde{\lambda}_X^1,$$
$$\frac{1}{\lambda_X^1} s_{yy}^{11} \xrightarrow{F} \tilde{\lambda}_Y^1,$$
$$\frac{1}{\lambda_X^1} s_{xy}^{11} \xrightarrow{F} \tilde{\lambda}_{XY}^1,$$
$$\frac{1}{\lambda_X^1} s_{xx}^{ij} \xrightarrow{p} 0, \ i \neq j,$$
$$\frac{1}{\lambda_X^1} s_{yy}^{ij} \xrightarrow{p} 0, \ i \neq j,$$
$$\frac{1}{\lambda_X^1} s_{xy}^{ij} \xrightarrow{p} 0, \ i \neq j,$$

where $\xrightarrow{F}$ denotes convergence in distribution and $\xrightarrow{p}$ denotes convergence in probability.

Below we provide a summary of the values taken on by $\tilde{\lambda}_X^1$, $\tilde{\lambda}_Y^1$ and $\tilde{\lambda}_{XY}^1$ for each of the population models. Calculations showing the convergence to each of these quantities is given in the proof of Lemma 5.2.2. An interesting point is that in those circumstances where the population canonical correlation converges to 1 we have that $\tilde{\lambda}_X^1 = \tilde{\lambda}_Y^1 = \tilde{\lambda}_{XY}^1$. In contrast, when the population canonical correlation converges 0, these quantities are not necessarily equal.

Model 1:

$$\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1},$$

$$\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1},$$
$$\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}. \tag{5.9}$$

Model 2:

1. *S1*: We have three cases here: $\beta = 1$, $\beta > 1$ and $\beta < 1$, we will refer to these as S1 case I, S1 case II and S1 case III.

   S1 case I:

   $$\tilde{\lambda}_X^1 = \tilde{\lambda}_Y^1 = \tilde{\lambda}_{XY}^1 = \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1}). \tag{5.10}$$

   S1 case II:

   $$\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1},$$
   $$\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1},$$
   $$\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}. \tag{5.11}$$

   S1 case III:

   $$\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1},$$
   $$\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1},$$
   $$\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}. \tag{5.12}$$

2. *S2*: We have two cases here when $\alpha > \beta$ and when $\alpha < \beta$, we will refer to these as S2 case I and S2 case II.

   S2 case I:

   $$\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1},$$

$$\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1},$$
$$\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}. \tag{5.13}$$

S2 case II:

$$\tilde{\lambda}_X^1 = \tilde{\lambda}_Y^1 = \tilde{\lambda}_{XY}^1 = \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1}). \tag{5.14}$$

Model 3:

$$\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1},$$
$$\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1},$$
$$\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}. \tag{5.15}$$

Lemma 5.2.2. *Under the population models described in Section 5.2.4 we have the following behavior in the covariance and cross covariance matrices as* $d \to \infty$

1. $$\frac{1}{\lambda_X^1}\mathbf{S}_{XX} \xrightarrow{F} diag\left(\tilde{\lambda}_X^1, 0, \ldots, 0\right), \tag{5.16}$$

   and similar results hold for $\mathbf{S}_{YY}$. *The value of* $\tilde{\lambda}_X^1$ *and* $\tilde{\lambda}_Y^1$ *will depend on the population model.*

2. $$\frac{1}{\lambda_X^1}\mathbf{S}_{XY} \xrightarrow{F} diag\left(\tilde{\lambda}_{XY}^1, 0, \ldots, 0\right), \tag{5.17}$$

   where the value of $\tilde{\lambda}_{XY}^1$ *depends on the population model.*

Before going into the proof of Lemma 5.2.2 we will need the following results

Proposition 5.2.3. *Let* $\mathbf{u}$, $\mathbf{w}$, $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$, *where* $\mathbf{u}$, $\mathbf{w}$ *and* $\mathbf{z}$ *are independent of one another, then*

$$Cov(\mathbf{w}^T\mathbf{u}, \mathbf{z}^T\mathbf{u}) = n(n+2), \tag{5.18}$$

$$Cov(\mathbf{w}^T\mathbf{w}, \mathbf{w}^T\mathbf{z}) = 3n(n+4). \tag{5.19}$$

Proof. We begin by showing the equality in (5.18),

$$\begin{aligned}
\mathrm{Cov}(\mathbf{w}^T\mathbf{u}, \mathbf{z}^T\mathbf{u}) &= \mathrm{Cov}(\sum_{i=1}^{n} w_i u_i, \sum_{i=1}^{n} z_i u_i) \\
&= \sum_{i=1}^{n} \mathrm{Cov}(w_i u_i, z_i u_i) + 2\sum_{j<k} \mathrm{Cov}(w_j u_j, z_k u_k) \\
&= \sum_{i=1}^{n} \left(\mathrm{E}(w_i z_i u_i^2)^2 - [\mathrm{E}(w_i z_i u_i^2)]^2\right) + 2\sum_{j<k}^{n} \left(\mathrm{E}(w_j u_j z_k u_k)^2 - [\mathrm{E}(w_j u_j z_k u_k)]^2\right) \\
&= \sum_{i=1}^{n} \mathrm{E}w_i^2 \mathrm{E}z_i^2 \mathrm{E}u_i^4 + 2\sum_{j<k} \mathrm{E}w_j^2 \mathrm{E}u_j^2 \mathrm{E}z_k^2 \mathrm{E}u_k^2 \\
&= n(n+2).
\end{aligned}$$

Next we show the equality in (5.19)

$$\begin{aligned}
\mathrm{Cov}(\mathbf{w}^T\mathbf{w}, \mathbf{w}^T\mathbf{z}) &= \mathrm{Cov}\left(\sum_{i=1}^{n} w_i^2, \sum_{i=1}^{n} w_i z_i\right) \\
&= \sum_{i=1}^{n} \mathrm{Cov}(w_i^2, w_i z_i) + 2\sum_{j<k} \mathrm{Cov}(w_j^2, w_k z_k) \\
&= \sum_{i=1}^{n} \left(\mathrm{E}(w_i^3 z_i)^2 - \left[\mathrm{E}(w_i^3 z_i)\right]^2\right) + 2\sum_{j<k} \left(\mathrm{E}(w_j^2 w_k z_k)^2 - \left[\mathrm{E}(w_j^2 w_k z_k)\right]^2\right) \\
&= \sum_{i=1}^{n} \mathrm{E}w_i^6 \mathrm{E}z_i^2 + 2\sum_{j<k} \mathrm{E}w_j^4 \mathrm{E}w_k^2 \mathrm{E}z_k^2 \\
&= 3n(n+4)
\end{aligned}$$

□

We are now ready to prove Lemma 5.2.2.

Proof. For each of the population models described in Section 5.2.4 we first show

$$\frac{1}{\lambda_X^1}\mathbf{S}_{XX} \xrightarrow{F} \mathrm{diag}\left(\tilde{\lambda}_X^1, 0, \ldots, 0\right)$$

as $d \to \infty$ and then

$$\frac{1}{\lambda_X^1}\mathbf{S}_{XY} \xrightarrow{F} \text{diag}(\tilde{\lambda}_{XY}^1, 0, \ldots, 0),$$

as $d \to \infty$.

1. *Model 1*: Recall that in this population model we have no cross-covariance term and the leading population canonical correlation is always equal to 0. Furthermore, as we will see below, the limiting quantities $\tilde{\lambda}_X^1$, $\tilde{\lambda}_Y^1$ and $\tilde{\lambda}_{XY}^1$ are equal to different random variables. Under this model

$$\begin{aligned}\mathbf{X} &= (d^\alpha \mathbf{z}_{x1}, \mathbf{z}_{x2}, \ldots, \mathbf{z}_{xd}) \\ \mathbf{Y} &= (d^\alpha \mathbf{z}_{y1}, \mathbf{z}_{y2}, \ldots, \mathbf{z}_{yd}),\end{aligned}$$

from which we have

$$\begin{aligned}s_{xx}^{11} &= \frac{d^{2\alpha}}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1}, \\ s_{xx}^{1i} = s_{xx}^{i1} &= \frac{d^\alpha}{n}\mathbf{z}_{x1}^T\mathbf{z}_{xi}, i = 2, \ldots, d, \\ s_{xx}^{ij} = s_{xx}^{ji} &= \frac{1}{n}\mathbf{z}_{xi}^T\mathbf{z}_{xj}, i, j > 1\end{aligned}$$

and

$$\begin{aligned}s_{xy}^{11} &= \frac{d^{2\alpha}}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}, \\ s_{xy}^{1i} &= \frac{d^\alpha}{n}\mathbf{z}_{x1}^T\mathbf{z}_{yi}, i = 2, \ldots, d, \\ s_{xy}^{i1} &= \frac{d^\alpha}{n}\mathbf{z}_{xi}^T\mathbf{z}_{y1}, i = 2, \ldots, d, \\ s_{xy}^{ij} &= \frac{1}{n}\mathbf{z}_{xi}^T\mathbf{z}_{yj}, i, j > 1.\end{aligned}$$

We begin by looking at the behavior of the scaled covariance matrix $\frac{1}{\lambda_X^1}\mathbf{S}_{XX}$. Recalling that $\lambda_X^1 = d^{2\alpha}$ a direct calculation shows that

$$\begin{aligned}\frac{1}{\lambda_X^1}s_{xx}^{11} &= \frac{1}{d^{2\alpha}}d^{2\alpha}\mathbf{z}_{x1}^T\mathbf{z}_{x1} \\ &= \mathbf{z}_{x1}^T\mathbf{z}_{x1}.\end{aligned}$$

We now show that the remaining elements of $\frac{1}{\lambda_X^1}\mathbf{S}_{XX}$ converge to 0. We begin by looking at $\frac{1}{\lambda_X^1}s_{xx}^{1i}$.

$$\begin{aligned}P\left(\left|\frac{1}{\lambda_X^1}s_{xx}^{1i}\right| > \tau\right) &\leq \frac{\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{xi})}{\tau^2 n^2 d^{2\alpha}} \\ &= \frac{1}{\tau^2 n d^{2\alpha}} \\ &\to 0.\end{aligned}$$

Similar calculations show us that the remaining elements of the scaled covariance matrix converge to 0. Putting this together we have

$$\frac{1}{\lambda_X^1}\mathbf{S}_{XX} \xrightarrow{F} \mathrm{diag}(\mathbf{z}_{x1}^T\mathbf{z}_{x1}, 0, \ldots, 0). \tag{5.20}$$

Similar calculations give us

$$\frac{1}{\lambda_X^1}\mathbf{S}_{YY} \xrightarrow{F} \mathrm{diag}(\mathbf{z}_{y1}^T\mathbf{z}_{y1}, 0, \ldots, 0). \tag{5.21}$$

The behavior of the scaled cross-covariance matrix $\frac{1}{\lambda_X^1}\mathbf{S}_{XY}$ is quite similar to that of the scaled covariance matrix. The leading term, $\frac{1}{\lambda_X^1}s_{xy}^{11}$ can be calculated directly as

$$\frac{1}{\lambda_X^1}s_{xy}^{11} = \frac{1}{d^{2\alpha}}d^{2\alpha}\mathbf{z}_{x1}^T\mathbf{z}_{y1}$$

$$= \mathbf{z}_{x1}^T \mathbf{z}_{y1}.$$

Similar calculations as in the case of the scaled covariance matrix show that the remaining elements of $\frac{1}{\lambda_X^1}\mathbf{S}_{XY}$ converge to 0. Putting this together gives us

$$\frac{1}{\lambda_X^1}\mathbf{S}_{XY} \xrightarrow{F} \text{diag}(\mathbf{z}_{x1}^T \mathbf{z}_{y1}, 0, \ldots, 0). \tag{5.22}$$

Finally, from (5.20), (5.21) and (5.22) we have

$$\begin{aligned}
\tilde{\lambda}_X^1 &= \frac{1}{n}\mathbf{z}_{x1}^T \mathbf{z}_{x1}, \\
\tilde{\lambda}_Y^1 &= \frac{1}{n}\mathbf{z}_{y1}^T \mathbf{z}_{y1}, \\
\tilde{\lambda}_{XY}^1 &= \frac{1}{n}\mathbf{z}_{x1}^T \mathbf{z}_{y1},
\end{aligned}$$

as in (5.9).

2. *Model 2*: Under this population model we have two scenarios which we consider, models $S1$ and $S2$.

   S1: Recall that under this population model the population canonical correlation converges 1 only when $\beta = 1$ and converges to 0 otherwise. An interesting observation is that when $\beta = 1$ we have that $\tilde{\lambda}_X^1 = \tilde{\lambda}_Y^1 = \tilde{\lambda}_{XY}^1$. When $\beta \neq 1$ we see that these quantities all correspond to different random variables. In this population model we have that

$$\begin{aligned}
\mathbf{X} &= \left(d^{\alpha}\mathbf{z}_{x1} + d^{\alpha\beta}\mathbf{z}_{y1}, \mathbf{z}_{x2}, \ldots, \mathbf{z}_{xd}\right), \\
\mathbf{Y} &= \left(d^{\alpha}\mathbf{z}_{y1} + d^{\alpha\beta}\mathbf{z}_{x1}, \mathbf{z}_{y2}, \ldots, \mathbf{z}_{yd}\right).
\end{aligned}$$

From which it can be seen that

$$
\begin{aligned}
s_{xx}^{11} &= \frac{1}{n}(d^{\alpha}\mathbf{z}_{x1} + d^{\alpha\beta}\mathbf{z}_{y1})^T(d^{\alpha}\mathbf{z}_{x1} + d^{\alpha\beta}\mathbf{z}_{y1}),\\
s_{xx}^{1i} &= s_{xx}^{i1} = \frac{1}{n}(d^{\alpha}\mathbf{z}_{x1} + d^{\alpha\beta}\mathbf{z}_{y1})^T\mathbf{z}_{xi}, i = 2,\ldots,d,\\
s_{xx}^{ij} &= s_{xx}^{ji} = \frac{1}{n}\mathbf{z}_{xi}^T\mathbf{z}_{xj},
\end{aligned}
$$

and

$$
\begin{aligned}
s_{xy}^{11} &= \frac{1}{n}(d^{\alpha}\mathbf{z}_{x1} + d^{\alpha\beta}\mathbf{z}_{y1})^T(d^{\alpha}\mathbf{z}_{y1} + d^{\alpha\beta}\mathbf{z}_{x1}),\\
s_{xy}^{1i} &= \frac{1}{n}(d^{\alpha}\mathbf{z}_{x1} + d^{\alpha\beta}\mathbf{z}_{y1})^T\mathbf{z}_{yi}, i = 2,\ldots,d,\\
s_{xy}^{i1} &= \frac{1}{n}(d^{\alpha}\mathbf{z}_{y1} + d^{\alpha\beta}\mathbf{z}_{x1})^T\mathbf{z}_{xi}, i = 2,\ldots,d,\\
s_{xy}^{ij} &= \frac{1}{n}\mathbf{z}_{xi}^T\mathbf{z}_{yj}, i,j > 1.
\end{aligned}
$$

Now we will show the convergence of the components of the covariance and cross-covariance matrices to their respective quantities as $d \to \infty$. We consider three cases, $\beta = 1$, $\beta < 1$ and $\beta > 1$ (recall that $\lambda_X^1 = d^{2\alpha} + d^{2\alpha\beta}$).

$\beta = 1$: In this case we can calculate directly the value of $\tilde{\lambda}_X^1 = \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})$

$$
\begin{aligned}
\frac{1}{\lambda_X^1}s_{xx}^{11} &= \frac{1}{n(d^{2\alpha} + d^{2\alpha\beta})}(d^{\alpha}\mathbf{z}_{x1} + d^{\alpha\beta}\mathbf{z}_{y1})^T(d^{\alpha}\mathbf{z}_{x1} + d^{\alpha\beta}\mathbf{z}_{y1})\\
&= \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})\\
&\overset{F}{=} \frac{\chi_n^2}{n},
\end{aligned}
$$

where χ_n^2 denotes the Chi-squared distribution with n degrees of freedom. Similar calculations give us $\tilde{\lambda}_Y^1 = \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})$.

$\beta < 1$: Here we show that $\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1}$ (similar calculations give us $\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1}$)

$$
\begin{aligned}
&P\left(\left|\frac{1}{\lambda_X^1}s_{xx}^{11} - \frac{\mathbf{z}_{x1}^T\mathbf{z}_{x1}}{n}\right| > \tau\right) \\
&\leq \frac{1}{\tau^2}\mathrm{Var}\left(\frac{1}{d^{2\alpha}+d^{2\alpha\beta}}s_{xx}^{11} - \frac{\mathbf{z}_{x1}^T\mathbf{z}_{x1}}{n}\right) \\
&= \frac{1}{\tau^2 n^2(d^{2\alpha}+d^{2\alpha\beta})^2}\mathrm{Var}(d^{2\alpha}\mathbf{z}_{x1}^T\mathbf{z}_{x1} - (d^{2\alpha}+d^{2\alpha\beta})\mathbf{z}_{x1}^T\mathbf{z}_{x1} + d^{2\alpha\beta}\mathbf{z}_{y1}^T\mathbf{z}_{y1} \\
&\quad + 2d^{\alpha(1+\beta)}\mathbf{z}_{x1}^T\mathbf{z}_{y1}) \\
&= \frac{1}{\tau^2 n^2(d^{2\alpha}+d^{2\alpha\beta})^2}\left[d^{4\alpha\beta}\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{x1}) + d^{4\alpha\beta}\mathrm{Var}(\mathbf{z}_{y1}^T\mathbf{z}_{y1}) + 4d^{2\alpha(1+\beta)}\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{y1})\right. \\
&\quad \left. -4d^{2\alpha\beta+\alpha(1+\beta)}\mathrm{Cov}(\mathbf{z}_{x1}^T\mathbf{z}_{x1}, \mathbf{z}_{x1}^T\mathbf{z}_{y1}) + 4d^{2\alpha\beta+\alpha(1+\beta)}\mathrm{Cov}(\mathbf{z}_{y1}^T\mathbf{z}_{y1}, \mathbf{z}_{x1}^T\mathbf{z}_{y1})\right] \\
&= \frac{4(d^{4\alpha\beta}+d^{2\alpha(1+\beta)})}{\tau^2 n(d^{2\alpha}+d^{2\alpha\beta})^2} \\
&\to 0.
\end{aligned}
$$

Where we have convergence to 0 as the order of the terms in the numerator, $d^{4\alpha\beta}$ and $d^{2\alpha(1+\beta)}$ are less than $d^{4\alpha}$ since $\beta < 1$.

$\beta > 1$: The calculation in this case is similar to when $\beta < 1$ but now $\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1}$ and $\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1}$ Next we show that $\frac{1}{\lambda_X^1}s_{xx}^{1i} \to 0$, $i = 2, \ldots, d$.

$$
\begin{aligned}
&P\left(\left|\frac{1}{\lambda_X^1}s_{xx}^{1i}\right| > \tau\right) \\
&\leq \frac{1}{\tau^2(d^{2\alpha}+d^{2\alpha\beta})^2}\mathrm{Var}(s_{xx}^{1i}) \\
&= \frac{1}{\tau^2(d^{2\alpha}+d^{2\alpha\beta})^2}\mathrm{Var}(d^{\alpha}\mathbf{z}_{x1}^T\mathbf{z}_{xi} + d^{\alpha\beta}\mathbf{z}_{y1}^T\mathbf{z}_{xi}) \\
&= \frac{1}{\tau^2(d^{2\alpha}+d^{2\alpha\beta})^2}\left[d^{2\alpha}\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{xi}) + d^{2\alpha\beta}\mathrm{Var}(\mathbf{z}_{y1}^T\mathbf{z}_{xi})\right.
\end{aligned}
$$

$$
\begin{aligned}
&+2d^{\alpha(1+\beta)}\mathrm{Cov}(\mathbf{z}_{x1}^T\mathbf{z}_{xi}, \mathbf{z}_{y1}^T\mathbf{z}_{xi})\big] \\
&= \frac{1}{\tau^2 n(d^{2\alpha}+d^{2\alpha\beta})} + \frac{2(n+2)}{\tau^2 n(d^{\frac{\alpha}{2}(3-\beta)}+d^{\frac{\alpha}{2}(3\beta-1)})^2} \\
&\to 0
\end{aligned}
$$

Similar calculations give us $\frac{1}{\lambda_X^1}\mathbf{z}_{xi}^T\mathbf{z}_{xj} \to 0$, as $d \to \infty$ for $i, j > 1$. Putting these results together we have

$$
\frac{1}{\lambda_X^1}\mathbf{S}_{XX} \xrightarrow{F} \mathrm{diag}(\tilde{\lambda}_X^1, 0, \ldots, 0),
$$

where $\tilde{\lambda}_X^1$ depends on the value β.

Next we look at the behavior of the elements of $\mathbf{S}_{XY}$ when $\beta = 1$ and when $\beta \neq 1$.

$\beta = 1$: As was the case with the covariance matrix the cross-covariance matrix can be calculated directly giving us $\tilde{\lambda}_{XY}^1 = \frac{1}{2n}(\mathbf{z}_{x1}+\mathbf{z}_{y1})^T(\mathbf{z}_{x1}+\mathbf{z}_{y1})$

$$
\begin{aligned}
\frac{1}{\lambda_X^1}s_{xy}^{11} &= \frac{1}{n(d^{2\alpha}+d^{2\alpha\beta})}(d^{\alpha}\mathbf{z}_{x1}+d^{\alpha\beta}\mathbf{z}_{y1})^T(d^{\alpha}\mathbf{z}_{y1}+d^{\alpha\beta}\mathbf{z}_{x1}) \\
&= \frac{1}{2n}(\mathbf{z}_{x1}+\mathbf{z}_{y1})^T(\mathbf{z}_{x1}+\mathbf{z}_{y1}) \\
&\overset{d}{=} \frac{\chi_n^2}{n}.
\end{aligned}
$$

$\beta \neq 1$: When $\beta \neq 1$ we have that $\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}$.

$$
\begin{aligned}
&P\left(\left|\frac{1}{\lambda_X^1}s_{xy}^{11} - \frac{\mathbf{z}_{x1}^T\mathbf{z}_{y1}}{n}\right| > \tau\right) \\
&\leq \frac{1}{\tau^2}\mathrm{Var}\left(\frac{1}{d^{2\alpha}+d^{2\alpha\beta}}s_{xy}^{11} - \frac{\mathbf{z}_{x1}^T\mathbf{z}_{y1}}{n}\right) \\
&= \frac{1}{n^2\tau^2(d^{2\alpha}+d^{2\alpha\beta})^2}\mathrm{Var}(d^{\alpha(1+\beta)}\mathbf{z}_{x1}^T\mathbf{z}_{x1} + d^{\alpha(1+\beta)}\mathbf{z}_{y1}^T\mathbf{z}_{y1}
\end{aligned}
$$

$$
\begin{aligned}
&+ (d^{2\alpha} + d^{2\alpha\beta})\mathbf{z}_{x1}^T\mathbf{z}_{y1} - (d^{2\alpha} + d^{2\alpha\beta})\mathbf{z}_{x1}^T\mathbf{z}_{y1}) \\
&= \frac{4nd^{2\alpha(1+\beta)}}{n^2\tau^2(d^{2\alpha} + d^{2\alpha\beta})^2} \\
&= \frac{4}{n\tau^2(d^{\alpha(1-\beta)} + d^{\alpha(\beta-1)})^2} \\
&\to 0
\end{aligned}
$$

Similar calculations for $\beta > 1$ gives us $\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}$. Next we show that $\frac{1}{\lambda_X^1}s_{xy}^{1i} \to 0$,

$$
\begin{aligned}
&P\left(\left|\frac{1}{\lambda_X^1}s_{xy}^{1i}\right| > \tau\right) \\
&\leq \frac{1}{(d^{2\alpha} + d^{2\alpha\beta})^2}\mathrm{Var}(s_{xy}^{1i}) \\
&= \frac{1}{\tau^2n^2(d^{2\alpha} + d^{2\alpha\beta})^2}\mathrm{Var}(d^{\alpha}\mathbf{z}_{x1}^T\mathbf{z}_{yi} + d^{\alpha\beta}\mathbf{z}_{y1}^T\mathbf{z}_{yi}) \\
&= \frac{1}{\tau^2n^2(d^{2\alpha} + d^{2\alpha\beta})^2}\left[d^{2\alpha}\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{yi}) + d^{2\alpha\beta}\mathrm{Var}(\mathbf{z}_{y1}^T\mathbf{z}_{yi})\right. \\
&\quad \left.+2d^{\alpha(1+\beta)}\mathrm{Cov}(\mathbf{z}_{x1}^T\mathbf{z}_{yi}, \mathbf{z}_{y1}^T\mathbf{z}_{yi})\right] \\
&= \frac{1}{\tau^2n(d^{2\alpha} + d^{2\alpha\beta})} + \frac{2(n+2)}{\tau^2n(d^{\frac{\alpha}{2}(3-\beta)} + d^{\frac{\alpha}{2}(3\beta-1)})^2} \\
&\to 0.
\end{aligned}
$$

Similar calculations are used to show that $\frac{1}{\lambda_X^1}s_{xy}^{i1}, \frac{1}{\lambda_X^1}s_{xy}^{ij} \to 0$. From this we see that

$$
\frac{1}{\lambda_X^1}\mathbf{S}_{XY} \xrightarrow{F} \mathrm{diag}(\tilde{\lambda}_{XY}^1, 0, \ldots, 0),
$$

where the value of $\tilde{\lambda}_{XY}^1$ depends on the value of α and β. Summarizing the results of Model 2 $S1$ we have

$\beta = 1$:

$$
\tilde{\lambda}_X^1 = \tilde{\lambda}_Y^1 = \tilde{\lambda}_{XY}^1 = \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1}).
$$

$\beta > 1$:

$$\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1},$$
$$\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1},$$
$$\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}.$$

$\beta < 1$:

$$\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1},$$
$$\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1},$$
$$\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}.$$

S2: Recall that in this population model we have that the populations canonical correlation converges to 1 when $\alpha < \beta$ and it converges to 0 when $\alpha > \beta$. As was discussed in Model 2 *S*1 the values of $\tilde{\lambda}_X^1$, $\tilde{\lambda}_Y^1$ and $\tilde{\lambda}_{XY}^1$ will be equal or not depending on whether the population canonical correlation converges to 1 or 0. In this population model we have that

$$\mathbf{X} = ((d^\alpha + d^\beta)\mathbf{z}_{x1} + d^\beta\mathbf{z}_{y1}, \mathbf{z}_{x2}, \ldots, \mathbf{z}_{xd}),$$
$$\mathbf{Y} = ((d^\alpha + d^\beta)\mathbf{z}_{y1} + d^\beta\mathbf{z}_{x1}, \mathbf{z}_{y2}, \ldots, \mathbf{z}_{yd}).$$

From which it can be seen that

$$s_{xx}^{11} = \frac{1}{n}((d^\alpha + d^\beta)\mathbf{z}_{x1} + d^\beta\mathbf{z}_{y1})^T((d^\alpha + d^\beta)\mathbf{z}_{x1} + d^\beta\mathbf{z}_{y1}),$$
$$s_{xx}^{1i} = s_{xx}^{i1} = \frac{1}{n}((d^\alpha + d^\beta)\mathbf{z}_{x1} + d^\beta\mathbf{z}_{y1})^T\mathbf{z}_{xi}, i = 2, \ldots, d,$$
$$s_{xx}^{ij} = s_{xx}^{ji} = \frac{1}{n}\mathbf{z}_{xi}^T\mathbf{z}_{xj},$$

and

$$
\begin{aligned}
s_{xy}^{11} &= \frac{1}{n}((d^\alpha + d^\beta)\mathbf{z}_{x1} + d^\beta \mathbf{z}_{y1})^T((d^\alpha + d^\beta)\mathbf{z}_{y1} + d^\beta \mathbf{z}_{x1}),\\
s_{xy}^{1i} &= \frac{1}{n}((d^\alpha + d^\beta)\mathbf{z}_{x1} + d^\beta \mathbf{z}_{y1})^T\mathbf{z}_{yi}, i = 2, \ldots, d,\\
s_{xy}^{i1} &= \frac{1}{n}((d^\alpha + d^\beta)\mathbf{z}_{y1} + d^\beta \mathbf{z}_{x1})^T\mathbf{z}_{xi}, i = 2, \ldots, d,\\
s_{xy}^{ij} &= \frac{1}{n}\mathbf{z}_{xi}^T\mathbf{z}_{yj}, i, j > 1.
\end{aligned}
$$

We begin by showing the convergence of $\frac{1}{\lambda_X^1}s_{xx}^{11}$ for the case $\alpha > \beta$ and $\alpha < \beta$ (recall that $\lambda_X^1 = (d^\alpha + d^\beta)^2 + d^{2\beta}$).

$\alpha > \beta$: Here we show that $\tilde{\lambda}_X^1 = \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1}$ (and $\tilde{\lambda}_Y^1 = \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1}$).

$$
\begin{aligned}
&P\left(\left|\frac{1}{\lambda_X^1}s_{xx}^{11} - \frac{\mathbf{z}_{x1}^T\mathbf{z}_{x1}}{n}\right| > \tau\right)\\
&\leq \frac{1}{\tau^2}\mathrm{Var}\left(\frac{1}{(d^\alpha + d^\beta)^2 + d^{2\beta}}s_{xx}^{11} - \frac{\mathbf{z}_{x1}^T\mathbf{z}_{x1}}{n}\right)\\
&= \frac{1}{\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\mathrm{Var}((d^\alpha + d^\beta)^2\mathbf{z}_{x1}^T\mathbf{z}_{x1} + d^{2\beta}\mathbf{z}_{y1}^T\mathbf{z}_{y1}\\
&\quad + 2d^\beta(d^\alpha + d^\beta)\mathbf{z}_{x1}^T\mathbf{z}_{y1} - ((d^\alpha + d^\beta)^2 + d^{2\beta})\mathbf{z}_{x1}^T\mathbf{z}_{x1})\\
&= \frac{1}{\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\left[d^{4\beta}(\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{x1}) + \mathrm{Var}(\mathbf{z}_{y1}^T\mathbf{z}_{y1}))\right.\\
&\quad +4d^{2\beta}(d^\alpha + d^\beta)^2\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{y1}) - 4d^{3\beta}(d^\alpha + d^\beta)\mathrm{Cov}(\mathbf{z}_{x1}^T\mathbf{z}_{x1}, \mathbf{z}_{x1}^T\mathbf{z}_{y1})\\
&\quad \left.+4d^{3\beta}(d^\alpha + d^\beta)\mathrm{Cov}(\mathbf{z}_{y1}^T\mathbf{z}_{y1}, \mathbf{z}_{x1}^T\mathbf{z}_{y1})\right]\\
&= \frac{4n(d^{4\beta} + d^{2\beta}(d^\alpha + d^\beta))}{\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\\
&= \frac{4(d^{4(\beta-\alpha)} + d^{2(\beta-\alpha)}(1 + d^{\beta-\alpha})^2)}{\tau^2 n((1 + d^{\beta-\alpha})^2 + d^{2(\beta-\alpha)})^2}\\
&\to 0.
\end{aligned}
$$

$\alpha < \beta$: Here we show that when $\alpha < \beta$ that $\tilde{\lambda}_X^1 = \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})$

(and $\tilde{\lambda}_Y^1 = \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})$).

$$
\begin{aligned}
&P\left(\left|\frac{1}{\lambda_X^1}s_{xx}^{11} - \frac{(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})}{2n}\right| > \tau\right) \\
&\leq \frac{1}{\tau^2}\mathrm{Var}\left(\frac{1}{\lambda_X^1}s_{xx}^{11} - \frac{(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})}{2n}\right) \\
&= \frac{1}{\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\mathrm{Var}\left((d^\alpha + d^\beta)^2\mathbf{z}_{x1}^T\mathbf{z}_{x1} + d^{2\beta}\mathbf{z}_{y1}^T\mathbf{z}_{y1}\right. \\
&\quad \left. +2d^\beta(d^\alpha + d^\beta)\mathbf{z}_{x1}^T\mathbf{z}_{y1} - \frac{((d^\alpha + d^\beta)^2 + d^{2\beta})(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})}{2}\right) \\
&= \frac{1}{4\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\mathrm{Var}\left(((d^\alpha + d^\beta)^2 - d^{2\beta})\mathbf{z}_{x1}^T\mathbf{z}_{x1}\right. \\
&\quad \left. +(d^{2\beta} - (d^\alpha + d^\beta)^2)\mathbf{z}_{y1}^T\mathbf{z}_{y1} - 2d^{2\alpha}\mathbf{z}_{x1}^T\mathbf{z}_{y1}\right) \\
&= \frac{1}{4\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\left[((d^\alpha + d^\beta)^2 - d^{2\beta})^2\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{x1})\right. \\
&\quad +(d^{2\beta} - (d^\alpha + d^\beta)^2)^2\mathrm{Var}(\mathbf{z}_{y1}^T\mathbf{z}_{y1}) + 4d^{4\alpha}\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{y1}) \\
&\quad -4d^{2\alpha}((d^\alpha + d^\beta)^2 - d^{2\beta})\mathrm{Cov}(\mathbf{z}_{x1}^T\mathbf{z}_{x1}, \mathbf{z}_{x1}^T\mathbf{z}_{y1}) \\
&\quad \left. -4d^{2\alpha}(d^{2\beta} - (d^\alpha + d^\beta)^2)\mathrm{Cov}(\mathbf{z}_{y1}^T\mathbf{z}_{y1}, \mathbf{z}_{x1}^T\mathbf{z}_{y1})\right] \\
&= \frac{2n[((d^\alpha + d^\beta)^2 - d^{2\beta})^2 + (d^{2\beta} - (d^\alpha + d^\beta)^2)^2 + 2d^{4\alpha}]}{4\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2} \\
&= \frac{((d^{\alpha-\beta} + 1)^2 - 1)^2 + (1 - (d^{\alpha-\beta} + 1)^2)^2 + 2d^{4(\alpha-\beta)}}{2\tau^2 n((d^{\alpha-\beta} + 1)^2 + 1)^2} \\
&\to 0.
\end{aligned}
$$

Next we show that the remaining elements of $\frac{1}{\lambda_X^1}\mathbf{S}_{XX}$ go to zero as $d \to \infty$. We begin with s_{xx}^{1i}

$$
\begin{aligned}
&P\left(\left|\frac{1}{\lambda_X^1}s_{xx}^{1i}\right| > \tau\right) \\
&\leq \frac{1}{\tau^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\mathrm{Var}(s_{xx}^{1i}) \\
&= \frac{1}{\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\mathrm{Var}((d^\alpha + d^\beta)\mathbf{z}_{x1}^T\mathbf{z}_{xi} + d^\beta\mathbf{z}_{y1}^T\mathbf{z}_{xi}) \\
&= \frac{1}{\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\left[(d^\alpha + d^\beta)^2\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{xi}) + d^{2\beta}\mathrm{Var}(\mathbf{z}_{y1}^T\mathbf{z}_{xi})\right.
\end{aligned}
$$

$$
\begin{aligned}
&+2d^{\beta}(d^{\alpha}+d^{\beta})\mathrm{Cov}(\mathbf{z}_{x1}^{T}\mathbf{z}_{xi},\mathbf{z}_{y1}^{T}\mathbf{z}_{xi})\big] \\
&= \frac{1}{\tau^2 n((d^{\alpha}+d^{\beta})^2+d^{2\beta})} + \frac{2(n+2)}{\tau^2 n((d^{\alpha-\frac{1}{3}\beta}+d^{\frac{2}{3}\beta})^{\frac{3}{2}}+(d^{\alpha-3\beta}+d^{-2\beta})^{-\frac{1}{2}})^2} \\
&\to 0.
\end{aligned}
$$

The proof showing $\frac{1}{\lambda_X^1}s_{xx}^{ij} \xrightarrow{p} 0$, $i,j>1$ is similar. Putting these all together we have that

$$
\frac{1}{\lambda_X^1}\mathbf{S}_{XX} \xrightarrow{F} \mathrm{diag}(\tilde{\lambda}_X^1,0,\ldots,0),
$$

where $\tilde{\lambda}_X^1$ depends on the value α and β.

Next we study the behavior of the scaled cross-covariance matrix $\frac{1}{\lambda_X^1}\mathbf{S}_{XY}$ when $\alpha>\beta$ and when $\alpha<\beta$.

$\alpha>\beta$: We begin by showing that when $\alpha>\beta$ we have $\tilde{\lambda}_{XY}^1 = \frac{1}{n}\mathbf{z}_{x1}^{T}\mathbf{z}_{y1}$.

$$
\begin{aligned}
&P\left(\left|\frac{1}{\lambda_X^1}s_{xy}^{11}-\frac{1}{n}\mathbf{z}_{x1}^{T}\mathbf{z}_{y1}\right|>\tau\right) \\
&\leq \frac{1}{\tau^2}\mathrm{Var}\left(\frac{1}{\lambda_X^1}s_{xy}^{11}-\frac{1}{n}\mathbf{z}_{x1}^{T}\mathbf{z}_{y1}\right) \\
&= \frac{1}{\tau^2 n^2((d^{\alpha}+d^{\beta})^2+d^{2\beta})^2}\mathrm{Var}(d^{\beta}(d^{\alpha}+d^{\beta})(\mathbf{z}_{x1}^{T}\mathbf{z}_{x1}+\mathbf{z}_{y1}^{T}\mathbf{z}_{y1}) \\
&\quad +((d^{\alpha}+d^{\beta})^2+d^{2\beta})\mathbf{z}_{x1}^{T}\mathbf{z}_{y1}-((d^{\alpha}+d^{\beta})^2+d^{2\beta})\mathbf{z}_{x1}^{T}\mathbf{z}_{y1}) \\
&= \frac{1}{\tau^2 n^2((d^{\alpha}+d^{\beta})^2+d^{2\beta})^2}(d^{\beta}(d^{\alpha}+d^{\beta}))^2(\mathrm{Var}(\mathbf{z}_{x1}^{T}\mathbf{z}_{x1})+\mathrm{Var}(\mathbf{z}_{y1}^{T}\mathbf{z}_{y1})) \\
&= \frac{4n(d^{\beta}(d^{\alpha}+d^{\beta}))^2}{\tau^2 n^2((d^{\alpha}+d^{\beta})^2+d^{2\beta})^2} \\
&= \frac{4(d^{\beta-\alpha}(1+d^{\beta-\alpha}))^2}{\tau^2 n((1+d^{\beta-\alpha})^2+d^{2(\beta-\alpha)})^2} \\
&\to 0.
\end{aligned}
$$

$\alpha < \beta$: Next we show that for $\alpha < \beta$ $\tilde{\lambda}^1_{XY} = \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})$.

$$
\begin{aligned}
& P\left(\left|\frac{1}{\lambda^1_X}s^{11}_{xy} - \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})\right| > \tau\right) \\
& \leq \frac{1}{\tau^2}\mathrm{Var}\left(\frac{1}{\lambda^1_X}s^{11}_{xy} - \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1})\right) \\
& = \frac{1}{4\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\mathrm{Var}\left(2d^\beta(d^\alpha + d^\beta)(\mathbf{z}^T_{x1}\mathbf{z}_{x1} + \mathbf{z}^T_{y1}\mathbf{z}_{y1})\right. \\
& \quad +2((d^\alpha + d^\beta)^2 + d^{2\beta})\mathbf{z}^T_{x1}\mathbf{z}_{y1} - ((d^\alpha + d^\beta)^2 + d^{2\beta})(\mathbf{z}^T_{x1}\mathbf{z}_{x1} + \mathbf{z}^T_{y1}\mathbf{z}_{y1}) \\
& \quad \left.-2((d^\alpha + d^\beta)^2 + d^{2\beta})\mathbf{z}^T_{x1}\mathbf{z}_{y1}\right) \\
& = \frac{1}{4\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}d^{4\alpha}(\mathrm{Var}(\mathbf{z}^T_{x1}\mathbf{z}_{x1}) + \mathrm{Var}(\mathbf{z}^T_{y1}\mathbf{z}_{y1})) \\
& = \frac{4nd^{4\alpha}}{4\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2} \\
& = \frac{d^{4(\alpha-\beta)}}{\tau^2 n((d^{\alpha-\beta} + 1)^2 + 1)^2} \\
& \to 0.
\end{aligned}
$$

Next we show that the remaining elements of $\frac{1}{\lambda^1_X}\mathbf{S}_{XY}$ converge to 0 as d goes infinity. We begin by looking at s^{1i}_{xy}, $i = 2, \ldots, d$. Proof of convergence to 0 for the remaining elements follow along similar lines.

$$
\begin{aligned}
& P\left(\left|\frac{1}{\lambda^1_X}s^{1i}_{xy}\right| > \tau\right) \\
& \leq \frac{1}{\tau^2}\mathrm{Var}\left(\frac{1}{\lambda^1_X}s^{1i}_{xy}\right) \\
& = \frac{1}{\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\mathrm{Var}((d^\alpha + d^\beta)\mathbf{z}^T_{x1}\mathbf{z}_{yi} + d^\beta\mathbf{z}^T_{y1}\mathbf{z}_{yi}) \\
& = \frac{1}{\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\left[(d^\alpha + d^\beta)^2\mathrm{Var}(\mathbf{z}^T_{x1}\mathbf{z}_{yi}) + d^{2\beta}\mathrm{Var}(\mathbf{z}^T_{y1}\mathbf{z}_{yi})\right. \\
& \quad \left.2d^\beta(d^\alpha + d^\beta)\mathrm{Cov}(\mathbf{z}^T_{x1}\mathbf{z}_{yi}, \mathbf{z}^T_{y1}\mathbf{z}_{yi})\right] \\
& = \frac{1}{\tau^2 n^2((d^\alpha + d^\beta)^2 + d^{2\beta})^2}\left[n((d^\alpha + d^\beta)^2 + d^{2\beta}) + 2n(n+2)d^\beta(d^\alpha + d^\beta)\right] \\
& = \frac{1}{\tau^2 n((d^\alpha + d^\beta)^2 + d^{2\beta})} + \frac{2(n+2)}{\tau^2 n((d^{\alpha-\frac{1}{3}\beta})^{\frac{3}{2}} + (d^{\alpha-3\beta} + d^{-2\beta}))^2}
\end{aligned}
$$

$\to 0$.

Putting all these results together we have

$$\frac{1}{\lambda_X^1}\mathbf{S}_{XY} \xrightarrow{F} \text{diag}(\tilde{\lambda}_{XY}^1, 0, \ldots, 0),$$

where the value of $\tilde{\lambda}_{XY}^1$ depends on α and β. We summarize the results from Model 2 $S2$ below

$\alpha > \beta$:

$$\begin{aligned}
\tilde{\lambda}_X^1 &= \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{x1}, \\
\tilde{\lambda}_Y^1 &= \frac{1}{n}\mathbf{z}_{y1}^T\mathbf{z}_{y1}, \\
\tilde{\lambda}_{XY}^1 &= \frac{1}{n}\mathbf{z}_{x1}^T\mathbf{z}_{y1}.
\end{aligned}$$

$\alpha < \beta$:

$$\tilde{\lambda}_X^1 = \tilde{\lambda}_Y^1 = \tilde{\lambda}_{XY}^1 = \frac{1}{2n}(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T(\mathbf{z}_{x1} + \mathbf{z}_{y1}).$$

3. *Model 3*: Recall that in this population model the population canonical correlation always converges to 0 (with the exception of the trivial case where the parameter $\alpha = 0$). In this population model we have that

$$\begin{aligned}
\mathbf{X} &= (\mathbf{z}_{x1} + d^\alpha\mathbf{z}_{y1}, \mathbf{z}_{x2}, \ldots, \mathbf{z}_{xd}), \\
\mathbf{Y} &= (\mathbf{z}_{y1} + d^\alpha\mathbf{z}_{x1}, \mathbf{z}_{y2}, \ldots, \mathbf{z}_{yd}).
\end{aligned}$$

From which it can be seen that

$$s_{xx}^{11} = \frac{1}{n}(\mathbf{z}_{x1} + d^\alpha\mathbf{z}_{y1})^T(\mathbf{z}_{x1} + d^\alpha\mathbf{z}_{y1}),$$

$$
\begin{aligned}
s_{xx}^{1i} = s_{xx}^{i1} &= \frac{1}{n}(\mathbf{z}_{x1} + d^{\alpha}\mathbf{z}_{y1})^T\mathbf{z}_{xi}, i = 2, \ldots, d, \\
s_{xx}^{ij} = s_{xx}^{ji} &= \frac{1}{n}\mathbf{z}_{xi}^T\mathbf{z}_{xj}, i, j > 1,
\end{aligned}
$$

and

$$
\begin{aligned}
s_{xy}^{11} &= \frac{1}{n}(\mathbf{z}_{x1} + d^{\alpha}\mathbf{z}_{y1})^T(\mathbf{z}_{y1} + d^{\alpha}\mathbf{z}_{x1}), \\
s_{xy}^{1i} &= \frac{1}{n}(\mathbf{z}_{x1} + d^{\alpha}\mathbf{z}_{y1})^T\mathbf{z}_{yi}, i = 2, \ldots, d, \\
s_{xy}^{i1} &= \frac{1}{n}(\mathbf{z}_{y1} + d^{\alpha}\mathbf{z}_{x1})^T\mathbf{z}_{xi}, i = 2, \ldots, d, \\
s_{xy}^{ij} &= \frac{1}{n}\mathbf{z}_{xi}^T\mathbf{z}_{yj}, i, j > 1.
\end{aligned}
$$

We begin by looking at the scaled covariance matrix $\frac{1}{\lambda_X^1}\mathbf{S}_{XX}$. Since the case for $\alpha = 0$ does not depend on the number of dimensions d we do not consider it in this example. Turning our attention to $\alpha > 0$ we show that $\tilde{\lambda}_X^1 = \mathbf{z}_{y1}^T\mathbf{z}_{y1}$.

$$
\begin{aligned}
& P\left(\left|\frac{1}{\lambda_X^1}s_{xx}^{11} - \mathbf{z}_{y1}^T\mathbf{z}_{y1}\right| > \tau\right) \\
& \leq \frac{1}{\tau^2}\mathrm{Var}\left(\frac{1}{\lambda_X^1}s_{xx}^{11} - \mathbf{z}_{y1}^T\mathbf{z}_{y1}\right) \\
& = \frac{1}{\tau^2 n^2(1 + d^{2\alpha})^2}\mathrm{Var}(\mathbf{z}_{x1}^T\mathbf{z}_{x1} + 2d^{\alpha}\mathbf{z}_{x1}^T\mathbf{z}_{y1}) \\
& = \frac{2n(1 + 2d^{2\alpha} + 6(n+4)d^{\alpha})}{\tau^2 n^2(1 + d^{2\alpha})^2} \\
& \to 0.
\end{aligned}
$$

Next we show that the remaining elements of $\frac{1}{\lambda_X^2}\mathbf{S}_{XX}$ converge to 0. We only show $\frac{1}{\lambda_X^1}s_{xx}^{1i} \to 0$, $i = 2, \ldots, d$, as the proofs for the remaining elements is quite similar.

$$
\begin{aligned}
& P\left(\left|\frac{1}{\lambda_X^1}s_{xx}^{1i}\right|\right) \\
& \leq \frac{1}{\tau^2}\mathrm{Var}\left(\frac{1}{\lambda_X^1}s_{xx}^{1i}\right)
\end{aligned}
$$

$$= \frac{1}{\tau^2 n^2 (1+d^{2\alpha})^2} \mathrm{Var}(\mathbf{z}_{x1}^T \mathbf{z}_{xi} + d^{\alpha} \mathbf{z}_{y1}^T \mathbf{z}_{xi})$$
$$\frac{n(1+d^{2\alpha}+2(n+2)d^{\alpha})}{\tau^2 n^2 (1+d^{2\alpha})^2}$$
$$\to 0.$$

Putting these all together we have that

$$\frac{1}{\lambda_X^1} \mathbf{S}_{XX} \xrightarrow{F} \mathrm{diag}(\mathbf{z}_{y1}^T \mathbf{z}_{y1}, 0, \ldots, 0).$$

We now turn our focus to the scaled cross-covariance matrix $\frac{1}{\lambda_X^1}\mathbf{S}_{XY}$. As the proof for the cross-covariance matrix are quite similar to that of the covariance matrix the details are omitted. We have that

$$\frac{1}{\lambda_X^1} \mathbf{S}_{XY} \xrightarrow{F} \mathrm{diag}(\mathbf{z}_{x1}^T \mathbf{z}_{y1}, 0, \ldots, 0).$$

Summarizing the results for Model 3 we have

$$\tilde{\lambda}_X^1 = \frac{1}{n} \mathbf{z}_{y1}^T \mathbf{z}_{y1},$$
$$\tilde{\lambda}_Y^1 = \frac{1}{n} \mathbf{z}_{x1}^T \mathbf{z}_{x1},$$
$$\tilde{\lambda}_{XY}^1 = \frac{1}{n} \mathbf{z}_{x1}^T \mathbf{z}_{y1}.$$

This completes our proof. □

We now return to our discussion of the sample cross-correlation matrix $\mathbf{R}_{XY}$. With the results of Lemma 5.2.2 we can now prove the following theorem.

Theorem 5.2.4. *Under all the population models described in Section 5.2.4 we have*

$$\mathbf{R}_{XY} \xrightarrow{d \to \infty} \mathit{diag}(r, 0, \ldots, 0),$$

where

$$r = \frac{\tilde{\lambda}^1_{XY}}{\sqrt{\tilde{\lambda}^1_X + c}\sqrt{\tilde{\lambda}^1_Y + c}},$$

and $c = \lim_{d\to\infty} \frac{\kappa}{\lambda^1_X} \sim \lim_{d\to\infty} \frac{d^\gamma}{\lambda^1_X}$. *The values of* $\lambda^1_X, \tilde{\lambda}^1_X, \tilde{\lambda}^1_Y$, $\tilde{\lambda}^1_{XY}$ *and subsequently* r *depend on the population model and the parameters* α *and* β.

Corollary 5.2.5 follows directly from Theorem 5.2.4. Letting f_R denote the density of the sample correlation coefficient when the correlation is 0 (Anderson (2003))

$$f_R = \frac{\Gamma\left[\frac{1}{2}n\right]}{\Gamma\left[\frac{1}{2}(n-1)\right]\sqrt{\pi}}(1-R^2)^{\frac{1}{2}(n-3)},$$

we have

Corollary 5.2.5. *Let* ρ_1 *and* $\hat{\rho}_1$ *denote the population and sample canonical correlations, we then have*

$$\hat{\rho}_{1(d)} \to \begin{cases} 1 & \text{if } \rho_1 \overset{d\to\infty}{\to} 1 \text{ and } \frac{\kappa}{\lambda^1_X} \overset{d\to\infty}{\to} 0, \\ 0 & \text{if } \rho_1 \overset{d\to\infty}{\to} 0 \text{ or } 1 \text{ and } \frac{\kappa}{\lambda^1_X} \overset{d\to\infty}{\to} \infty, \\ R & \text{if } \rho_1 \overset{d\to\infty}{\to} 0 \text{ and } \frac{\kappa}{\lambda^1_X} \overset{d\to\infty}{\to} 0, R \sim f_R. \end{cases}$$

The results from Corollary 5.2.5 are summarized in Remark 5.2.8.

Remark 5.2.6. Note that in Corollary 5.2.5 when $\rho \overset{d\to\infty}{\to} 1$ we can make the additional statement that $\hat{\rho}_1$ is either *consistent* or *strongly inconsistent* depending on the behavior of $\frac{\kappa}{\lambda^1_X}$. A similar statement cannot be made about the remaining cases since the support of R contains 0.

The results from Corollary 5.2.5 lead to the following theorem

Theorem 5.2.7. *Assuming* $\rho_1 \overset{d\to\infty}{\to} 1$ *and* $\frac{\kappa}{\lambda^1_X} \overset{d\to\infty}{\to} 0$ *then*

$$angle(\mathbf{w}^1_X, \hat{\mathbf{w}}^1_X) \xrightarrow{p} 0,$$

$$angle(\mathbf{w}^1_Y, \hat{\mathbf{w}}^1_Y) \xrightarrow{p} 0.$$

In Theorem 5.2.7 we consider only the case where ρ_1 converges to 1 since in all other cases no conclusive statement can be made about the convergence of the angle between the population and sample canonical vectors. The reason for this is that the directions found when ρ_1 does not converge to 1 will be random in either the population and/or sample canonical vectors. This can be seen by noting that either the population or sample cross-correlation matrix will be equal to the matrix of 0's. Therefore the set of left and right singular vectors resulting from an SVD of the cross-correlation matrix can be any orthonormal basis.

We now prove Theorem 5.2.4.

Proof. From Lemma 5.2.2 we have

$$\begin{aligned}
\frac{1}{\lambda_X^1}\mathbf{S}_{XX} &\overset{d\to\infty}{\to} \operatorname{diag}(\tilde{\lambda}_X^1, 0, \ldots, 0),\\
\frac{1}{\lambda_X^1}\mathbf{S}_{YY} &\overset{d\to\infty}{\to} \operatorname{diag}(\tilde{\lambda}_Y^1, 0, \ldots, 0),\\
\frac{1}{\lambda_X^1}\mathbf{S}_{XY} &\overset{d\to\infty}{\to} \operatorname{diag}(\tilde{\lambda}_{XY}^1, 0, \ldots, 0).
\end{aligned}$$

With this in mind we have

$$\begin{aligned}
\mathbf{R}_{XY} &= (\mathbf{S}_{XX} + \kappa\mathbf{I}_n)^{-\frac{1}{2}}\mathbf{S}_{XY}(\mathbf{S}_{YY} + \kappa\mathbf{I}_n)^{-\frac{1}{2}}\\
&= (\mathbf{S}_{XX} + \kappa\mathbf{I}_n)^{-\frac{1}{2}}\left(\frac{1}{\lambda_X^1}\right)^{-\frac{1}{2}}\left(\frac{1}{\lambda_X^1}\right)^{\frac{1}{2}}\mathbf{S}_{XY}\left(\frac{1}{\lambda_X^1}\right)^{\frac{1}{2}}\left(\frac{1}{\lambda_X^1}\right)^{-\frac{1}{2}}(\mathbf{S}_{YY} + \kappa\mathbf{I}_n)^{-\frac{1}{2}}\\
&= \left(\frac{1}{\lambda_X^1}\mathbf{S}_{XX} + \frac{\kappa}{\lambda_X^1}\mathbf{I}_n\right)^{-\frac{1}{2}}\left(\frac{1}{\lambda_X^1}\mathbf{S}_{XY}\right)\left(\frac{1}{\lambda_X^1}\mathbf{S}_{YY} + \frac{\kappa}{\lambda_X^1}\mathbf{I}_n\right)^{-\frac{1}{2}}\\
&\overset{d\to\infty}{\to} \begin{pmatrix} \tilde{\lambda}_X^1 + c & 0 & \cdots & 0\\ 0 & c & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & c \end{pmatrix}^{-\frac{1}{2}} \begin{pmatrix} \tilde{\lambda}_{XY}^1 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \tilde{\lambda}_Y^1 + c & 0 & \cdots & 0\\ 0 & c & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & c \end{pmatrix}^{-\frac{1}{2}}
\end{aligned}$$

$$= \text{diag}\left(\frac{\tilde{\lambda}_{XY}}{\sqrt{\tilde{\lambda}_X^1 + c}\sqrt{\tilde{\lambda}_X^1 + c}}, 0, \ldots, 0\right),$$

as we wanted to show. □

Remark 5.2.8. We consider each of the population models separately.

Model 1: Note that under Model 1 $\rho_1 = 0$ and does not depend on the value of d or α. From (5.9) we have

$$r = \frac{\mathbf{z}_{x1}^T \mathbf{z}_{y1}}{\sqrt{\mathbf{z}_{x1}^T \mathbf{z}_{x1} + c}\sqrt{\mathbf{z}_{y1}^T \mathbf{z}_{y1} + c}},$$

and so

$$\hat{\rho}_1 \overset{d\to\infty}{\to} \begin{cases} 0 & \text{if } c = \infty, \\ R & \text{if } c = 0. \end{cases}$$

Model 2:

1. *S1 case I*: Note that this case corresponds to conditions under which $\rho_1 \overset{d\to\infty}{\to} 1$. From (5.10) we have

$$r = \frac{(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T (\mathbf{z}_{x1} + \mathbf{z}_{y1})}{(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T (\mathbf{z}_{x1} + \mathbf{z}_{y1}) + c},$$

   and so

$$\hat{\rho}_1 \overset{d\to\infty}{\to} \begin{cases} 0 & \text{if } c = \infty, \\ 1 & \text{if } c = 0. \end{cases}$$

2. *S1 case II and case III*: These cases correspond to conditions under which $\rho_1 \overset{d\to\infty}{\to} 0$. From (5.11) and (5.12) we have

$$r = \frac{\mathbf{z}_{x1}^T \mathbf{z}_{y1}}{\sqrt{\mathbf{z}_{x1}^T \mathbf{z}_{x1} + c}\sqrt{\mathbf{z}_{y1}^T \mathbf{z}_{y1} + c}},$$

and so

$$\hat{\rho}_1 \overset{d\to\infty}{\to} \begin{cases} 0 & \text{if } c = \infty, \\ R & \text{if } c = 0. \end{cases}$$

3. *S2 case I*: This case correspond to conditions under which $\rho_1 \overset{d\to\infty}{\to} 0$. From (5.13) we have

$$r = \frac{\mathbf{z}_{x1}^T \mathbf{z}_{y1}}{\sqrt{\mathbf{z}_{x1}^T \mathbf{z}_{x1} + c}\sqrt{\mathbf{z}_{y1}^T \mathbf{z}_{y1} + c}},$$

and so

$$\hat{\rho}_1 \overset{d\to\infty}{\to} \begin{cases} 0 & \text{if } c = \infty, \\ R & \text{if } c = 0. \end{cases}$$

4. *S2 case II*: This case corresponds to conditions under which $\rho_1 \overset{d\to\infty}{\to} 1$. From (5.14) we have

$$r = \frac{(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T (\mathbf{z}_{x1} + \mathbf{z}_{y1})}{(\mathbf{z}_{x1} + \mathbf{z}_{y1})^T (\mathbf{z}_{x1} + \mathbf{z}_{y1}) + c},$$

and so

$$\hat{\rho}_1 \overset{d\to\infty}{\to} \begin{cases} 0 & \text{if } c = \infty, \\ 1 & \text{if } c = 0. \end{cases}$$

Model 3: Note that under Model 3 $\rho_1 = 0$ regardless of the value of d or α. From (5.9) we have

$$r = \frac{\mathbf{z}_{x1}^T \mathbf{z}_{y1}}{\sqrt{\mathbf{z}_{y1}^T \mathbf{z}_{y1} + c}\sqrt{\mathbf{z}_{x1}^T \mathbf{z}_{x1} + c}},$$

and so

$$\hat{\rho}_1 \overset{d\to\infty}{\to} \begin{cases} 0 & \text{if } c = \infty, \\ R & \text{if } c = 0. \end{cases}$$

We now turn out attention to the proof of Theorem 5.2.7.

Proof. The only population models which satisfy the assumptions of the theorem are Model 2 S1 and Model 2 S2. Under these models and the assumptions made in the

statement of theorem we have that the cross-correlation matrix

$$\mathbf{R}_{XY} \overset{d\to\infty}{\to} \text{diag}(1, 0, \ldots, 0).$$

Let $\hat{\mathbf{w}}_X^{*1} = (1, 0, \ldots, 0)^T \in \mathbb{R}^{d\times 1}$ and $\mathbf{W}_X^{*(d-1)} \in \mathbb{R}^{d\times d-1}$ be any orthonormal basis orthogonal to $\hat{\mathbf{w}}_X^{*1}$. Similarly, let $\hat{\mathbf{w}}_Y^{*1} = (1, 0, \ldots, 0)^T \in \mathbb{R}^{d\times 1}$ and $\mathbf{W}_Y^{*(d-1)} \in \mathbb{R}^{d\times d-1}$ be any orthonormal basis orthogonal to $\hat{\mathbf{w}}_Y^{*1}$. Letting $\mathbf{R} = \text{diag}(1, 0, \ldots, 0) \in \mathbb{R}^{d\times d}$ we then have that the SVD of $\mathbf{R}_{XY}$ is

$$\mathbf{W}_X^* \mathbf{R} \mathbf{W}_Y^{*T},$$

where $\mathbf{W}_X^* = \begin{pmatrix} \hat{\mathbf{w}}_X^{*1} & \mathbf{W}_X^{*(d-1)} \end{pmatrix}$ and $\mathbf{W}_Y^* = \begin{pmatrix} \hat{\mathbf{w}}_Y^{*1} & \mathbf{W}_Y^{*(d-1)} \end{pmatrix}$. Thus we have that the leading scaled canonical vectors are $\mathbf{w}_X^{*1}$ and $\mathbf{w}_Y^{*1}$ and the corresponding unscaled canonical vectors are

$$\mathbf{w}_X^1 = (\mathbf{S}_{XX} + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{w}_X^{*1},$$
$$\mathbf{w}_Y^1 = (\mathbf{S}_{YY} + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{w}_Y^{*1}.$$

Next in order to show that the sample canonical vectors converge to their population counterpart as $d \to \infty$ we show that the angle between them goes to 0. We begin by calculating the cosine of the angle between the sample and population canonical vectors

$$\begin{aligned}
\langle \mathbf{w}_X^1, \hat{\mathbf{w}}_X^1 \rangle &= \frac{\mathbf{w}_X^{*1T} \Sigma_{XX}^{-\frac{1}{2}} (\mathbf{S}_{XX} + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \hat{\mathbf{w}}_X^{*1}}{\sqrt{\mathbf{w}_X^{*1T} \Sigma_{XX}^{-1} \mathbf{w}_X^{*1}} \sqrt{\hat{\mathbf{w}}_X^{*1T} (\mathbf{S}_{XX} + \kappa \mathbf{I}_n)^{-1} \hat{\mathbf{w}}_X^{*1}}} \\
&= \frac{\mathbf{w}_X^{*1T} \left(\frac{1}{\lambda_X^1} \Sigma_{XX} \right)^{-\frac{1}{2}} \left(\frac{1}{\lambda_X^1} \mathbf{S}_{XX} + \frac{\kappa}{\lambda_X^1} \mathbf{I}_n \right)^{-\frac{1}{2}} \hat{\mathbf{w}}_X^{*1}}{\sqrt{\mathbf{w}_X^{*1T} \left(\frac{1}{\lambda_X^1} \Sigma_{XX} \right)^{-1} \mathbf{w}_X^{*1}} \sqrt{\hat{\mathbf{w}}_X^{*1T} \left(\frac{1}{\lambda_X^1} \mathbf{S}_{XX} + \frac{\kappa}{\lambda_X^1} \mathbf{I}_n \right)^{-1} \hat{\mathbf{w}}_X^{*1}}} \\
&\overset{d\to\infty}{\to} \frac{(1, 0, \ldots, 0) \text{diag} \left(\left(\frac{1}{2n} (\mathbf{z}_{x1} + \mathbf{z}_{y1})^T (\mathbf{z}_{x1} + \mathbf{z}_{y1}) \right)^{-\frac{1}{2}}, 0, \ldots 0 \right) \hat{\mathbf{w}}_X^{*1}}{\left(\frac{1}{2n} (\mathbf{z}_{x1} + \mathbf{z}_{y1})^T (\mathbf{z}_{x1} + \mathbf{z}_{y1}) \right)^{-\frac{1}{2}}} \\
&= (1, 0, \ldots, 0)(1, 0, \ldots, 0)^T
\end{aligned}$$

$$= 1.$$

Thus we have that

$$\text{angle}(\mathbf{w}_X^1, \hat{\mathbf{w}}_X^1) = \arccos\left(\langle \mathbf{w}_X^1, \hat{\mathbf{w}}_X^1 \rangle\right) \overset{d\to\infty}{\to} 0$$

as we wanted to show. □

5.2.6 Asymptotics of the Sample Kernel Cross-Correlation Matrix

Using the sample kernel cross-correlation representation for analyzing the behavior of CCA in the HDLSS setting has appeal in that the matrices $\mathbf{K}_X$ and $\mathbf{K}_Y$ are composed of the inner products between observations. We can exploit certain asymptotic properties such as independence between observations or utilize other assumptions which we place on the distribution of the data by letting $d \to \infty$.

An important component in our analysis of the HDLSS behavior of CCA are the measures of sphericity

$$\epsilon_X \equiv \frac{\text{Tr}^2(\Sigma_{XX})}{d\text{Tr}(\Sigma^2)} = \frac{(\sum_{i=1}^d \lambda_X^i)^2}{d\sum_{i=1}^d (\lambda_X^i)^2}$$

and

$$\epsilon_Y \equiv \frac{\text{Tr}^2(\Sigma_{YY})}{d\text{Tr}(\Sigma^2)} = \frac{(\sum_{i=1}^d \lambda_Y^i)^2}{d\sum_{i=1}^d (\lambda_Y^i)^2},$$

proposed by John John (1971) and John (1972) as the basis of a hypothesis test for the equality of eigenvalues. Here λ_X^i and $\lambda_Y^i, i = 1, \ldots, d$ are the eigenvalues of Σ_{XX} and Σ_{YY}. Note that for ϵ denoting either ϵ_X or ϵ_Y, the following inequalities always hold

$$\frac{1}{d} \leq \epsilon \leq 1.$$

Also note that $\epsilon = 1$ only when all the eigenvalues are all equal.

Following the discussion of Jung and Marron (2009) we assume the ϵ-condition, i.e. that $\epsilon_X, \epsilon_Y \gg \frac{1}{d}$, in the sense that

$$\begin{aligned}(d\epsilon_X)^{-1} &= \frac{\sum_{i=1}^{d}(\lambda_X^i)^2}{(\sum_{i=1}^{d}\lambda_X^i)^2} \to 0 \text{ as } d \to \infty \\ (d\epsilon_Y)^{-1} &= \frac{\sum_{i=1}^{d}(\lambda_Y^i)^2}{(\sum_{i=1}^{d}\lambda_Y^i)^2} \to 0 \text{ as } d \to \infty. \end{aligned} \tag{5.23}$$

In the following lemma we will show that the following conditions are necessary in order for the ϵ-condition to hold,

1. Model 1: $0 \leq \alpha < \frac{1}{2}$.
2. Model 2:

   $S1$: If $\beta \leq 1$ then $0 \leq \alpha < \frac{1}{2}$, and if $\beta > 1$ then $0 < \alpha < \frac{1}{2\beta}$.

   $S2$: Either $0 \leq \alpha < \beta < \frac{1}{2}$ or $0 \leq \beta < \alpha < \frac{1}{2}$.
3. Model 3: $0 \leq \alpha < \frac{1}{2}$.

With this in mind we have the following lemma

Lemma 5.2.9. *Assuming the ϵ-condition holds and letting λ_X^i and $\lambda_Y^i, i = 1, \ldots, n$ be the eigenvalues of the population covariance matrices Σ_{XX} and Σ_{YY}. Then the off diagonal elements of the scaled kernel matrices, $\frac{n}{\sum_{i=1}^{d}\lambda_X^i}\mathbf{K}_X$ and $\frac{n}{\sum_{i=1}^{d}\lambda_Y^i}\mathbf{K}_Y$ converge to 0 and the diagonal elements converge to 1 as $d \to \infty$.*

Proof. We consider each of the models described in Section 5.2.4 and the conditions under which they satisfy the ϵ-condition separately. We present results for $\frac{1}{\sum_{i=1}^{n}\lambda_X^i}\mathbf{K}_X$ only as the proof for $\frac{1}{\sum_{i=1}^{n}\lambda_Y^i}\mathbf{K}_Y$ is exactly the same.

Model 1:

We begin by showing that the off diagonal elements of the scaled kernel matrix converge to zero in probability which we denote by $\xrightarrow{p}$. We have for $i \neq j$, and by Chebyshev's inequality, that

$$\begin{aligned}
&P\left(\left|\frac{1}{d^{2\alpha}+d-1}\mathbf{x}_i^T\mathbf{x}_j\right| > \tau\right) \\
&\leq \frac{\mathrm{Var}\left(\frac{1}{d^{2\alpha}+d-1}\mathbf{x}_i^T\mathbf{x}_j\right)}{\tau^2} \\
&= \frac{1}{\tau^2(d^{2\alpha}+d-1)^2}\left[\mathrm{Var}\left(d^{2\alpha}z_{x1}^i z_{x1}^j\right) + \mathrm{Var}\left(\sum_{k=2}^{d} z_{xk}^i z_{xk}^j\right)\right] \\
&= \frac{d^{4\alpha}+d-1}{\tau^2(d^{2\alpha}+d-1)^2} \xrightarrow{p} 0,
\end{aligned}$$

as $d \to \infty$ for $0 \leq \alpha < \frac{1}{2}$. Note that it is necessary that $0 \leq \alpha < \frac{1}{2}$ in order for the ϵ-condition to hold. Also, in the inequality above $\mathrm{Var}(\mathbf{x}_i^T\mathbf{x}_j) = \mathrm{E}(\mathbf{x}_i^T\mathbf{x}_j)^2$ since $\left(\mathrm{E}(\mathbf{x}_i^T\mathbf{x}_j)\right)^2 = 0$. Variance exploits the fact that we have independent components in the $z_{xk}^i, k = 1, \ldots d, i = 1, \ldots, n$ so that the variance of the sums is the sum of the variances.

Also note that in the last equality above, the ratio, excluding the $1/\tau^2$ is the sum of the squared eigenvalues over the sum of the eigenvalues squared, which is the exact form of the ϵ-condition. As we will see, this behavior holds true throughout this proof.

Next we show that the diagonal elements converge to 1 as $d \to \infty$

$$\begin{aligned}
&P\left(\left|\frac{1}{d^{2\alpha}+d-1}\mathbf{x}_i^T\mathbf{x}_i - 1\right| > \tau\right) \\
&\leq \frac{\mathrm{E}\left(\frac{1}{d^{2\alpha}+d-1}\mathbf{x}_i^T\mathbf{x}_i - 1\right)^2}{\tau^2} \\
&= \frac{1}{\tau^2}\left[\mathrm{Var}\left(\frac{1}{d^{2\alpha}+d-1}\mathbf{x}_i^T\mathbf{x}_i - 1\right) + \left(\mathrm{E}\left(\frac{1}{d^{2\alpha}+d-1}\mathbf{x}_i^T\mathbf{x}_i - 1\right)\right)^2\right] \\
&= \frac{1}{\tau^2}\left[\frac{1}{(d^{2\alpha}+d-1)^2}\left(d^{4\alpha}\mathrm{Var}\left((z_{x1}^i)^2\right) + \sum_{k=2}^{d}\mathrm{Var}\left((z_{xk}^i)^2\right)\right)\right] \\
&= \frac{d^{4\alpha}+d-1}{\tau^2(d^{2\alpha}+d-1)^2} \xrightarrow{p} 0,
\end{aligned}$$

as $d \to \infty$ for $0 \leq \alpha < \frac{1}{2}$.

Model 2 (S1):

In a similar fashion to Model 1, we begin by showing that the off-diagonal elements converge to zero

$$
\begin{aligned}
&P\left(\left|\frac{1}{d^{2\alpha}+d^{2\alpha\beta}+d-1}\mathbf{x}_i^T\mathbf{x}_j\right| > \tau\right) \\
&\leq \frac{\mathrm{Var}(\frac{1}{d^{2\alpha}+d^{2\alpha\beta}+d-1}\mathbf{x}_i^T\mathbf{x}_j)}{\tau^2} \\
&= \frac{1}{\tau^2(d^{2\alpha}+d^{2\alpha\beta}+d-1)^2}\left[\mathrm{Var}\left((d^\alpha z_{x1}^i + d^{\alpha\beta}z_{y1}^i)(d^\alpha z_{x1}^j + d^{\alpha\beta}z_{y1}^j)\right) + \mathrm{Var}\left(\sum_{k=2}^{d} z_{xk}^i z_{xk}^j\right)\right] \\
&= \frac{1}{\tau^2(d^{2\alpha}+d^{2\alpha\beta}+d-1)^2}\left[d^{4\alpha}\mathrm{Var}(z_{x1}^i z_{x1}^j) + d^{4\alpha\beta}\mathrm{Var}(z_{y1}^i z_{y1}^j)\right. \\
&\quad \left. + d^{\alpha(1+\beta)}\mathrm{Var}(z_{x1}^i z_{y1}^j) + d^{\alpha(1+\beta)}\mathrm{Var}(z_{y1}^i z_{x1}^j) + \sum_{k=2}^{d}\mathrm{Var}(z_{xk}^i z_{xk}^j)\right] \\
&= \frac{d^{4\alpha}+d^{4\alpha\beta}+2d^{\alpha(1+\beta)}+d-1}{\tau^2(d^{2\alpha}+d^{2\alpha\beta}+d-1)^2} \xrightarrow{p} 0.
\end{aligned}
$$

Recall that in the context of this population model the ϵ-condition requires that if $\beta \leq 1$ then $\alpha < \frac{1}{2}$ or if $\beta > 1$ then $\alpha < \frac{1}{2\beta}$, which in both cases allows for convergence to 0.

Next we show the convergence of the diagonal elements to 1.

$$
\begin{aligned}
&P\left(\left|\frac{1}{d^{2\alpha}+d^{2\alpha\beta}+d-1}\mathbf{x}_i^T\mathbf{x}_i - 1\right| > \tau\right) \\
&\leq \frac{\mathrm{E}\left(\frac{1}{d^{2\alpha}+d^{2\alpha\beta}+d-1}\mathbf{x}_i^T\mathbf{x}_i - 1\right)^2}{\tau^2} \\
&= \frac{1}{\tau^2}\mathrm{E}\left[\frac{1}{d^{2\alpha}+d^{2\alpha\beta}+d-1}\left((d^\alpha z_{x1}^i + d^{\alpha\beta}z_{y1}^i)^2 + \sum_{k=2}^{d}(z_{xk}^i)^2\right) - 1\right]^2 \\
&= \frac{1}{\tau^2}\left[\mathrm{Var}\left(\frac{1}{d^{2\alpha}+d^{2\alpha\beta}+d-1}\left((d^\alpha z_{x1}^i + d^{\alpha\beta}z_{y1}^i)^2 + \sum_{k=2}^{d}(z_{xk}^i)^2\right) - 1\right)\right. \\
&\quad \left. + \left(\mathrm{E}\frac{1}{d^{2\alpha}+d^{2\alpha\beta}+d-1}\left(\left(d^\alpha z_{x1}^i + d^{\alpha\beta}z_{y1}^i\right)^2 + \sum_{k=2}^{d}(z_{xk}^i)^2\right) - 1\right)^2\right]
\end{aligned}
$$

$$= \frac{1}{\tau^2}\left[\frac{d^{4\alpha}+d^{4\alpha\beta}+2d^{2\alpha(1+\beta)}+d-1}{(d^{2\alpha}+d^{2\alpha\beta}+d-1)^2}+\frac{(d^{2\alpha}+d^{2\alpha\beta}+d-1)^2}{(d^{2\alpha}+d^{2\alpha\beta}+d-1)^2}-1\right] \xrightarrow{p} 0,$$

where the above convergence holds provided the ϵ-condition is satisfied.

Model 2 (S2):

As before we begin by showing the convergence of the off-diagonal elements to zero

$$\begin{aligned}
&P\left(\left|\frac{1}{(d^\alpha+d^\beta)^2+d^{2\beta}+d-1}\mathbf{x}_i^T\mathbf{x}_j\right|>\tau\right)\\
&\leq \frac{\mathrm{Var}\left(\frac{1}{(d^\alpha+d^\beta)^2+d^{2\beta}+d-1}\mathbf{x}_i^T\mathbf{x}_j\right)}{\tau^2}\\
&= \frac{1}{\tau^2((d^\alpha+d^\beta)^2+d^{2\beta}+d-1)^2}\left[\mathrm{Var}\left(((d^\alpha+d^\beta)z_{x1}^i+d^\beta z_{y1}^i)((d^\alpha+d^\beta)z_{x1}^j+d^\beta z_{y1}^j)\right.\right.\\
&\quad\left.\left.+\sum_{k=2}^{d}z_{xk}^i z_{xk}^j\right)\right]\\
&= \frac{1}{\tau^2((d^\alpha+d^\beta)^2+d^{2\beta}+d-1)^2}\left[\mathrm{Var}\left((d^\alpha+d^\beta)^2 z_{x1}^i z_{x1}^j+d^{2\beta}z_{y1}^i z_{y1}^j\right.\right.\\
&\quad\left.\left.+d^{2\beta}(d^\alpha+d^\beta)z_{x1}^i z_{y1}^j+d^\beta(d^\alpha+d^\beta)z_{y1}^i z_{x1}^j\right)+\sum_{k=2}^{d}\mathrm{Var}(z_{xk}^i z_{xk}^j)\right]\\
&= \frac{(d^\alpha+d^\beta)^4+d^{4\beta}+2d^{2\beta}(d^\alpha+d^\beta)^2+d-1}{\tau^2((d^\alpha+d^\beta)^2+d^{2\beta}+d-1)^2}\xrightarrow{p}0,
\end{aligned}$$

provided the ϵ-condition hold, i.e. that either $0\leq\alpha<\beta<\frac{1}{2}$ or $0\leq\beta<\alpha<\frac{1}{2}$.

Next we show that the diagonal elements converge to 1.

$$\begin{aligned}
&P\left(\left|\frac{1}{(d^\alpha+d^\beta)^2+d^{2\beta}+d-1}\mathbf{x}_i^T\mathbf{x}_i-1\right|>\tau\right)\\
&\leq \frac{\mathrm{E}\left(\frac{1}{(d^\alpha+d^\beta)^2+d^{2\beta}+d-1}\mathbf{x}_i^T\mathbf{x}_i-1\right)^2}{\tau^2}\\
&= \frac{1}{\tau^2}\mathrm{E}\left[\frac{1}{(d^\alpha+d^\beta)^2+d^{2\beta}+d-1}\left(((d^\alpha+d^\beta)z_{x1}^i+d^\beta z_{y1}^i)^2+\sum_{k=2}^{d}(z_{xk}^i)^2\right)-1\right]^2\\
&= \frac{1}{\tau^2}\left[\frac{1}{(d^\alpha+d^\beta)^2+d^{2\beta}+d-1}\mathrm{Var}\left(((d^\alpha+d^\beta)z_{x1}^i+d^\beta z_{y1}^i)^2+\sum_{k=2}^{d}(z_{xk}^i)^2\right)\right.
\end{aligned}$$

$$+\left(\frac{1}{(d^\alpha+d^\beta)^2+d^{2\beta}+d-1}\mathrm{E}\left(((d^\alpha+d^\beta)z_{x1}^i+d^\beta z_{y1}^i)^2+\sum_{k=2}^{d}(z_{xk}^i)^2\right)-1\right)^2\Bigg]$$

$$=\frac{1}{\tau^2}\left[\frac{(d^\alpha+d^\beta)^4+d^{4\beta}+2d^{2\beta}(d^\alpha+d^\beta)^2+d-1}{((d^\alpha+d^\beta)^2+d^{2\beta}+d-1)^2}+\left(\frac{(d^\alpha+d^\beta)^2+d^{2\beta}+d-1}{(d^\alpha+d^\beta)^2+d^{2\beta}+d-1}-1\right)^2\right]\xrightarrow{p}0,$$

In order for the ϵ-condition to be satisfied we need to have either $0\leq\alpha<\beta<\frac{1}{2}$ or $0\leq\beta<\alpha<\frac{1}{2}$, from which the above convergence to 0 follows.

Model 3:

In a similar fashion to the previous proofs we begin by showing convergence of the off-diagonal elements to 0

$$\begin{aligned}
&P\left(\left|\frac{1}{d^{2\alpha}+d}\mathbf{x}_i^T\mathbf{x}_j\right|>\tau\right)\\
&\leq\frac{\mathrm{Var}\left(\frac{1}{d^{2\alpha}+d}\mathbf{x}_i^T\mathbf{x}_j\right)}{\tau^2}\\
&=\frac{1}{\tau^2(d^{2\alpha}+d)^2}\left[\mathrm{Var}\left((z_{x1}^i+d^\alpha z_{y1}^i)(z_{x1}^j+d^\alpha z_{y1}^j)\right)+\mathrm{Var}\left(\sum_{k=2}^{d}z_{xk}^i z_{xk}^j\right)\right]\\
&=\frac{d^{4\alpha}+2d^{2\alpha}+d}{\tau^2(d^{2\alpha}+d)^2}\xrightarrow{p}0.
\end{aligned}$$

The above convergence to 0 holds provided the ϵ-condition is satisfied, i.e. that $0\leq\alpha<\frac{1}{2}$.

Next we show convergence of the diagonal elements to 1.

$$\begin{aligned}
&P\left(\left|\frac{1}{d^{2\alpha}+d}\mathbf{x}_i^T\mathbf{x}_j-1\right|>\tau\right)\\
&\leq\frac{\mathrm{E}\left(\frac{1}{d^{2\alpha}+d}\mathbf{x}_i^T\mathbf{x}_j-1\right)^2}{\tau^2}\\
&=\frac{1}{\tau^2}\left[\frac{1}{(d^{2\alpha}+d)^2}\mathrm{E}\left((z_{x1}^i+d^\alpha z_{y1})^2+\sum_{k=2}^{d}(z_{xk}^i)^2\right)-1\right]^2\\
&=\frac{1}{\tau^2}\left[\mathrm{Var}\left(\frac{1}{d^{2\alpha}+d}\left((z_{x1}^i+d^\alpha z_{y1})^2+\sum_{k=2}^{d}(z_{xk}^i)^2\right)-1\right)\right.
\end{aligned}$$

$$+\left(\mathrm{E}\left(\frac{1}{d^{2\alpha}+d}\left((z_{x1}^i+d^\alpha z_{y1})^2+\sum_{k=2}^{d}(z_{xk}^i)^2\right)\right)-1\right)^2\Bigg]$$
$$=\frac{1}{\tau^2}\left[\frac{d^{4\alpha}+2d^{2\alpha}+d}{(d^{2\alpha}+d)^2}+\left(\frac{d^{2\alpha}+d}{d^{2\alpha}+d}-1\right)^2\right]\xrightarrow{p}0.$$

Putting all these results together we have that all the population models presented in Section 5.2.4 that

$$\frac{1}{\sum_{i=1}^n\lambda_X^i}\mathbf{K}_X\xrightarrow{p}\mathbf{I}_n$$
$$\frac{1}{\sum_{i=1}^n\lambda_Y^i}\mathbf{K}_Y\xrightarrow{p}\mathbf{I}_n.$$

□

The following theorems present conditions under which we have consistency or strong-inconsistency in the canonical correlations. As we will see these results depend heavily on the behavior of the regularization parameter $\kappa\sim d^\gamma$. Let ρ_i denote the population canonical correlation, $\hat{\rho}_i$ the sample canonical correlation, $i=1,\ldots,n$.

Theorem 5.2.10. *Assume for each of the population models described in Section 5.2.4 that the parameters α and β satisfy the ϵ-condition (discussed in conjunction with Lemma 5.2.9. Based on the population models described above we have the following behavior in the canonical correlations*

$$\lim_{d\to\infty}\hat{\rho}_i=\begin{cases}0 & \text{if } \gamma>1,\\ 1 & \text{if } \gamma<1.\end{cases}$$

1. *Model 1: If $\gamma>1$ then $\hat{\rho}_1$ is consistent and $\hat{\rho}_i$, $i=2,\ldots,n$ are strongly inconsistent. If $\gamma<1$ then all $\hat{\rho}_i$'s are strongly inconsistent.*

2. *Model 2:*

   S1: If $\beta=1$ and $0\le\alpha<\frac{1}{2}$, then $\hat{\rho}_1$ is consistent if $\gamma<1$ and $\hat{\rho}_i$, $i=2,\ldots,n$ are strongly inconsistent. If $\beta\neq1$ and $0\le\alpha<\frac{1}{2\beta}$ then $\hat{\rho}_i$, $i=1,\ldots,d$ are

consistent if $\gamma > 1$ and are strongly inconsistent otherwise.

S2: If $0 \leq \alpha < \beta < \frac{1}{2}$ then $\hat{\rho}_1$ is consistent if $\gamma < 1$, $\hat{\rho}_i$, $i = 2, \ldots, n$ are strongly inconsistent. If $0 \leq \beta < \alpha < \frac{1}{2}$ then all $\hat{\rho}_i$'s are consistent if $\gamma > 1$ and are all strongly-inconsistent otherwise.

3. *Model 3: If $\alpha = 0$, then $\hat{\rho}_1$ is consistent if $\gamma < 1$ and $\hat{\rho}_i$, $i = 2, \ldots, n$ are strongly inconsistent. If $\alpha > 0$ then all $\hat{\rho}_i$'s will be consistent if $\gamma > 1$ and will be strongly inconsistent otherwise.*

Proof. We begin by looking at the behavior of the dual cross-correlation matrix as $d \to \infty$,

$$\begin{aligned}
\mathbf{R}_{XY}^{K} &= (\mathbf{K}_X + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \mathbf{K}_X^{\frac{1}{2}} \mathbf{K}_Y^{\frac{1}{2}} (\mathbf{K}_Y + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \\
&= (\mathbf{K}_X + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \left(\sum_{i=1}^{d} \lambda_X^i \right)^{\frac{1}{2}} \left(\sum_{i=1}^{d} \lambda_X^i \right)^{-\frac{1}{2}} \mathbf{K}_X^{\frac{1}{2}} \\
&\quad \times \mathbf{K}_Y^{\frac{1}{2}} \left(\sum_{i=1}^{d} \lambda_Y^i \right)^{-\frac{1}{2}} \left(\sum_{i=1}^{d} \lambda_Y^i \right)^{\frac{1}{2}} (\mathbf{K}_Y + \kappa \mathbf{I}_n)^{-\frac{1}{2}} \\
&= \left(\frac{1}{\sum_{i=1}^{d} \lambda_X^i} \mathbf{K}_X + \frac{\kappa}{\sum_{i=1}^{d} \lambda_X^i} \mathbf{I}_n \right)^{-\frac{1}{2}} \left(\frac{1}{\sum_{i=1}^{d} \lambda_X^i} \mathbf{K}_X \right)^{\frac{1}{2}} \\
&\quad \times \left(\frac{1}{\sum_{i=1}^{d} \lambda_Y^i} \mathbf{K}_Y \right)^{\frac{1}{2}} \left(\frac{1}{\sum_{i=1}^{d} \lambda_Y^i} \mathbf{K}_Y + \frac{\kappa}{\sum_{i=1}^{d} \lambda_Y^i} \mathbf{I}_n \right)^{-\frac{1}{2}} \\
&\to \frac{1}{1+c} \mathbf{I}_n \text{ as } d \to \infty,
\end{aligned}$$

by Lemma 5.2.9. Of interest to us is to study the behavior of $c = \lim_{d \to \infty} \frac{d^\gamma}{\sum_{i=1}^{d} \lambda_X^i} = \lim_{d \to \infty} \frac{d^\gamma}{\sum_{i=1}^{d} \lambda_Y^i}$. Here c converges to 0 or 1 depending on the value of γ relative to the highest order of the sum of the eigenvalues λ_X^i and λ_Y^i, $i = 1, \ldots, d$. We will now look at the behavior of c under each of the above population models,

1. *Model 1*: Recall that the sum of the eigenvalues under this model is,

$$\sum_{i=1}^{d} \lambda_X^i = d^{2\alpha} + d - 1.$$

From Lemma 5.2.9 we know that in order for the ϵ-condition to hold we must have that $0 \leq \alpha < \frac{1}{2}$. We then have that

$$c = \lim_{d\to\infty} \frac{1}{\sum_{i=1}^{d} \lambda_X^i = d^{2\alpha-\gamma} + d^{1-\gamma} - d^{-\gamma}} = \begin{cases} \infty & \text{if } \gamma > 1, \\ 0 & \text{if } \gamma < 1. \end{cases}$$

From this we then have

$$\frac{1}{1+c} = \begin{cases} 0 & \text{if } \gamma > 1, \\ 1 & \text{if } \gamma < 1. \end{cases}$$

Conditions for consistency and strong inconsistency are described in the statement of the theorem.

2. *Model 2*:

   $S1$: Under this population model the sum of the eigenvalues is

$$\sum_{i=1}^{d} \lambda_X^i = d^{2\alpha} + d^{2\alpha\beta} + d - 1.$$

   Under the constraints of the ϵ-condition we have that if $\beta \leq 1$ then $0 \leq \alpha < \frac{1}{2}$ or if $\beta > 1$ then $0 \leq \alpha < \frac{1}{2\beta}$. Thus

$$c = \lim_{d\to\infty} \frac{1}{d^{2\alpha-\gamma} + d^{2\alpha\beta-\gamma} + d^{1-\gamma} - d^{-\gamma}} = \begin{cases} \infty & \text{if } \gamma > 1, \\ 0 & \text{if } \gamma < 1. \end{cases}$$

From this we then have

$$\frac{1}{1+c} = \begin{cases} 0 & \text{if } \gamma > 1, \\ 1 & \text{if } \gamma < 1. \end{cases}$$

Conditions for consistency and strong inconsistency are described in the statement of the theorem.

$S2$: The sum of the eigenvalues for this population model is

$$\sum_{i=1}^{d} \lambda_X^i = (d^\alpha + d^\beta)^2 + d^{2\beta} + d - 1$$

Recall that the conditions necessary for the ϵ-condition to hold are either $0 \leq \alpha < \beta < \frac{1}{2}$ or $0 \leq \beta < \alpha < \frac{1}{2}$. The behavior of c once again depends on the leading term d

$$c = \lim_{d\to\infty} \frac{1}{(d^{\alpha-\frac{\gamma}{2}} + d^{\beta-\frac{\gamma}{2}})^2 + d^{2\beta-\gamma} + d^{1-\gamma} - d^{-\gamma}} = \begin{cases} \infty & \text{if } \gamma > 1, \\ 0 & \text{if } \gamma < 1. \end{cases}$$

From this we then have that

$$\frac{1}{1+c} = \begin{cases} 0 & \text{if } \gamma > 1, \\ 1 & \text{if } \gamma < 1. \end{cases}$$

Consistency and strong inconsistency of the $\hat{\rho}_i$'s is described in the statement of the theorem.

3. *Model 3*: The sum of the eigenvalues under this population model is

$$\sum_{i=1}^{d} \lambda_X^i = d^{2\alpha} + d$$

In order for the ϵ-conditions to hold we must have that $0 \leq \alpha < \frac{1}{2}$. Then c behaves

as

$$c = \lim_{d\to\infty} \frac{1}{d^{2\alpha-\gamma} + d^{1-\gamma}} = \begin{cases} \infty & \text{if } \gamma > 1, \\ 0 & \text{if } \gamma < 1. \end{cases}$$

Consistency and strong inconsistency of the $\hat{\rho}_i$'s is described in the statement of the theorem.

This completes the proof. □

Remark 5.2.11. Note that throughout the proof of Theorem 5.2.10 the convergence of the sample canonical correlations depended a considerable amount more on the regularization parameter κ then on the relationship between the population parameters α and β. Specifically the relationship between γ (recall $\kappa \sim d^{\gamma}$) and the highest order term, d played the biggest role in determining the convergence of the sample canonical correlation. In contrast the convergence of the sample canonical correlations when we were exploring the behavior of the sample cross-correlation matrix did depend in large part on the relationship between the population parameters. This suggests that the regularization parameter is relatively more important for KCCA then for CCA.

In our analysis the kernel induced feature space that we were mapping into was characterized by the standard inner product. While the inner products defined in other kernel induced feature spaces may be considerably different, many are still a function of Euclidean distance, which is itself simply the inner product of the difference between two observations,

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &= f(||\mathbf{x} - \mathbf{x}'||^2) \\ &= f(\langle \mathbf{x} - \mathbf{x}', \mathbf{x} - \mathbf{x}' \rangle) \\ &= f(\mathbf{x}^T\mathbf{x} + \mathbf{x}'^T\mathbf{x}' - 2\mathbf{x}^T\mathbf{x}'). \end{aligned}$$

While we do not provide any formal proof, based on our results using the standard inner product kernel, it seems reasonable to conclude that the selection of the regularization

parameter for a general kernel plays a critical role in KCCA. This is not to say that the regularization parameter does not play an important role in standard CCA. However, based on our results from Section 5.2.5, where the sample cross-correlation matrix was studied, when the regularization parameter converged to 0, and the population canonical correlation converged 0, this did not immediately imply that the sample canonical correlation would converge to 1 (Corollary 5.2.5). However, when studying the sample kernel cross-correlation matrix when the regularization parameter converged to 0 the sample canonical correlation always converged to 1 (Theorem 5.2.10).

CHAPTER 6

Proposed Future Work

In the following section we discuss possible future work which looks at providing a framework for performing variable selection using KCCA.

6.1 Variable Selection KCCA

Variable selection can be a very useful statistical task. In particular in high dimensional data, where the number of parameters is potentially greater than the number of observations. In the context of KCCA we want to find the set of variables which is most meaningful for capturing the relationship between spaces.

We look to build upon the ideas presented in Lafferty and Wasserman (2008). The key idea in our approach is as follows. Consider a kernel which takes the form

$$K^h(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}^T \mathbf{H} \mathbf{x}'), \tag{6.1}$$

where $\mathbf{H} = \text{diag}(h_1, \ldots, h_d)$ and

$$\mathbf{K}(h) = \{K^h(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n.$$

By $\mathbf{K}(1)$ we mean $\mathbf{H} = \mathbf{I}$, i.e. $h_i = 1, i = 1, \ldots, d$. Let

$$\rho_{\mathcal{H}}(h) = \alpha_X^T \mathbf{K}_X(h_X) \mathbf{K}_Y(h_Y) \alpha_Y \tag{6.2}$$

where

$$\mathbf{K}_X(h_X) = \{K_X^h(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n,$$
$$\mathbf{K}_Y(h_Y) = \{K_Y^h(\mathbf{y}_i, \mathbf{y}_j)\}_{i,j=1}^n.$$

If $P = (h(t) : 0 \leq t \leq 1)$ is a smooth path through the set of weights with $h(0) = 1$ and $h(1) = 0$ then letting $\rho_{\mathcal{H}}$ be as in (3.19) we can then write

$$\rho_{\mathcal{H}} = \rho_{\mathcal{H}}(1) = \rho_{\mathcal{H}}(0) + \rho_{\mathcal{H}}(1) - \rho_{\mathcal{H}}(0) \tag{6.3}$$
$$= \rho_{\mathcal{H}}(0) - \int_0^1 \frac{d\rho_{\mathcal{H}}(h(s))}{ds} \tag{6.4}$$
$$= \rho_{\mathcal{H}}(0) - \int_0^1 \langle D(h(s)), \dot{h}(s)\rangle ds, \tag{6.5}$$

where

$$D(h) = \nabla\rho_{\mathcal{H}}(h) = (D_X(h), D_Y(h)) = \left(\frac{\partial\rho_{\mathcal{H}}}{\partial h_X^1}, \ldots, \frac{\partial\rho_{\mathcal{H}}}{\partial h_X^{d_X}}, \frac{\partial\rho_{\mathcal{H}}}{\partial h_Y^1}, \ldots, \frac{\partial\rho_{\mathcal{H}}}{\partial h_Y^{d_Y}}\right)^T \tag{6.6}$$

is the gradient of $\rho_{\mathcal{H}}$ and $\dot{h}(s) = \frac{dh(s)}{ds}$ is the derivative of $h(s)$ along the path.

If we assume that the number of relevant variables describing the relationship between spaces is in fact some $r_X < d_X$ and $r_Y < d_Y$ then there should be a path for which $D(h)$ is also sparse. Along such a path we replace $D(h)$ with some $\hat{D}(h)$ that makes use of the sparsity assumption. Our estimate of $\rho_{\mathcal{H}}$ is then

$$\hat{\rho}_{\mathcal{H}} = \rho_{\mathcal{H}}(0) - \int_0^1 \langle \hat{D}(h(s)), \dot{h(s)}\rangle ds. \tag{6.7}$$

To implement this idea we need two things

1. To find a sparse path for the derivative.

2. Take advantage of this sparsity as a method for variable selection.

The key observation is that if a particular covariate $\vec{\mathbf{x}}_j (\in \mathbb{R}^n), j = 1, \ldots, d_X$ and/or $\vec{\mathbf{y}}_k (\in \mathbb{R}^n), k = 1, \ldots, d_Y$ is irrelevant, then we would expect that changing the associated weights h_X^j and/or h_Y^k would cause little or no change to the canonical correlation $\rho_{\mathcal{H}}$. On the other hand, if $\vec{\mathbf{x}}_j$ and/or $\vec{\mathbf{y}}_k$ is important we would expect a small change in the weights h_X^j and/or h_Y^k to cause a large change in the canonical correlations. Thus the derivatives, $D_X^j(h) = \frac{\partial \rho_{\mathcal{H}}}{\partial h_X^j}$ and $D_Y^j(h) = \frac{\partial \rho_{\mathcal{H}}}{\partial h_Y^j}$ should discriminate between relevant and irrelevant covariates. To simplify the procedure we discretize the continuum of weights replacing $h_X(s)$ and $h_Y(s)$ with the sets

$$h_X^j \in \mathcal{B}_X = \{(1 - \beta_X^l)h_X^0, (1 - \beta_X^{2l})h_X^0, \ldots\} \text{ where } 0 \leq \beta_X \leq 1, l \in \mathbb{N}$$
$$h_Y^j \in \mathcal{B}_Y = \{(1 - \beta_Y^l)h_X^0, (1 - \beta_Y^{2l})h_Y^0, \ldots\} \text{ where } 0 \leq \beta_Y \leq 1, l \in \mathbb{N}.$$

Furthermore, we can proceed in a greedy fashion by estimating $D(h)$ sequentially with $h_X^j \in \mathcal{B}_X$ and $h_Y^j \in \mathcal{B}_Y$ by setting $\hat{D}_X^j(h) = 0$ when $h_X^j < \hat{h}_X^j$ and similarly setting $\hat{D}_Y^j(h) = 0$ when $h_Y^j < \hat{h}_Y^j$, where $\hat{h}_X^j$ and $\hat{h}_Y^j$ are the first h_X or h_Y, respectively, such that $|D_X^j(h)| < c_X^j(h)$ or $|D_Y^j(h)| < c_Y^j(h)$ for some threshold c_X^j and c_Y^j. Thus our estimate of $\rho_{\mathcal{H}}(h)$ is $\rho_{\mathcal{H}}\hat{h}$ and the hard threshold estimate of the derivatives are

$$\hat{D}_X(h) = D_X(h)I(|D_X(h)| > c_X(h)),$$
$$\hat{D}_Y(h) = D_Y(h)I(|D_Y(h)| > c_Y(h)).$$

The algorithm can be summarized as follows

1. Select constants $0 \leq \beta_X \leq 1$ and $0 \leq \beta_Y \leq 1$ and initial weights $0 \leq h_X^0 \leq 1$ and $0 \leq h_Y^0 \leq 1$.

2. Initialize the weights and activate all covariates:

   (a) $h_X^j = h_X^0, j = 1, \ldots, d_X$ and $h_Y^j = h_Y^0, j = 1, \ldots, d_Y$.

   (b) $\mathcal{A}_X = \{1, 2, \ldots, d_X\}$ and $\mathcal{A}_Y = \{1, 2, \ldots, d_Y\}$.

3. While $\mathcal{A}_X$ and $\mathcal{A}_Y$ are nonempty, do for each $j \in \mathcal{A}_X$ and $j \in \mathcal{A}_Y$:

   (a) Compute the thresholds c_X^j and c_Y^j.

   (b) If $|D_X^j| > c_X^j$, then set $h_X^j \leftarrow (1 - \beta_X^{tl})h_X^j$; otherwise remove j from $\mathcal{A}_X$. Similarly if $|D_Y^j| > c_Y^j$, then set $h_Y^j \leftarrow (1 - \beta_Y^{tl})h_Y^j$; otherwise remove j from $\mathcal{A}_Y$. Here t is the counter associated with the iteration number.

4. Output weight vectors $\hat{h}_X = (h_X^1, \ldots, h_X^{d_X})$ and $\hat{h}_Y = (h_Y^1, \ldots, h_Y^{d_Y})$ and updated canonical correlation $\hat{\rho}_{\mathcal{H}}(h)$.

Future work will focus on the implementation of this approach and a detailed study of algorithm performance and convergence.

BIBLIOGRAPHY

Ahn J., Marron J., Muller K. and Chi Y. (2007). The high dimension, low sample size geometric represetnation holds under mild conditions. *To appear in Biometrika* .

Ahn J. and Marron J.S. (2009). The maximal data piling direction for discrimination. *Biometrika* .

Alpay D. (2001). The schur algorithm, reproducing kerne spaces and system theory. *SMF/AMS Texts and Monographs* **5**.

Anderson T.W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience.

Bach F.R. and Jordan M.I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research* **3**, 1–48.

Bai Z. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica* **9**, 611–677.

Bai Z. and Yin Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix .

Baik J. and Silverstein J. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97**, 1382–1408.

Belkin M. and Niyogi P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15**, 1373–1396.

Bengio Y., Delalleau O., Roux N., Paiement J., Vincent P. and Ouimet M. (2004). Learning eigenfunctions links spectral embedding and kernel pca. *Neural Computation* **16**, 2197–2219.

Bie T.D. (2005). *Semi-supervised learning based on kernel methods and graph cut algorithms*. Phd thesis, K.U.Leuven (Leuven, Belgium), Faculty of Engineering.

Boyd S. and Vandeberghe L. (2004). *Convex Optimization*. Cambridge University Press.

Buja A., Hastie T. and Tibshirani R. (1995). Penalized discriminant analysis. *The Annals of Statistics* **23**, 73–102.

Chen J. and Ye J. (2008). Training svm with indefinite kernels. In: *ICML '08: Proceedings of the 25th international conference on Machine learning*, pp. 136–143. ACM, New York, NY, USA.

Chung F. (1997). *Spectral Graph Theory*. AMS.

Cristianini N. and Shawe-Taylor J. (2000). *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Daylight (2004). World drug index. URL `www.daylight.com`.

Duda R.O., Hart P.E. and Stork D.G. (2000). *Pattern Classification*. Interscience. Wiley, 2 edition.

Fisher R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Fukumizu K., Bach F.R., Gretton A. and Hyvarinen A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research* **8**, 361–383.

Geman S. (1980). A limit theorem for the norm of random matrices. *The Annals of Statistics* **8**, 252–261.

Haasdonk B. (2005). Feature space interpretation of svms with indefinite kernels. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **27**, 482–492.

Hall P., Marron J. and Neeman A. (2005). Geometric representation of high dimension low sample size data. *Journal of the Royal Statistical Society* **67**, 427–444.

Hardoon D.R., Szedmak S. and Shawe-Talyor J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* **16**, 2639–2664.

Hastie T., Buja A. and Tibshirani R.J. (1995). Penalized discriminant analysis. *The Annals of Statistics* **23**, 73–102.

Hotelling H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321–377.

Hoyle D. and Rattray M. (2004). Principal component analysis eigenvalue spectra from data with symmetry breaking structure. *Physical Review E* .

John S. (1971). Some optimal multivariate tests. *Biometrika* **58**, 123–127.

John S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika* **59**, 169–173.

Johnstone I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29**, 295–327.

Johnstone I. and Lu A. (2004). Sparse principal component analysis. Technical report, Stanford University.

Jung S. and Marron J. (2009). Pca consistency in high dimension, low sample size context .

Kettenring J.R. (1971). Canonical analysis of several sets of variables. *Biometrika* **58**, 433–451.

Kullback S. (1997). *Information Theory and Statistics*. Dover Publications.

Kuss M. and Graepel T. (2003). The geometry of kernel canonical correlation analysis. Technical Report 108, Max Planck Institute for Biological Cybernetics.

Lafferty J. and Wasserman L. (2008). *The Annals of Statistics* **36**, 28–63.

Lee M. (2007). *Continuum Direction Vectors in High Dimension Low Sample Size Data*. Ph.D. thesis, The University of North Carolina at Chapel Hill.

Li Y. and Shawe-Taylor J. (2006). Using kcca for japanese-english cross-language information retrieval and document classification. *Journal of Intelligent Information Systems* **27**, 117–133.

Luss R. and d'Aspremont A. (2008). Support vector machine classification with indefinite kernels. *CoRR* **abs/0804.0188**.

Marcenko V. and Pasture L. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR - Sbornik* **1**, 457–486.

Marron J., Tood M. and Ahn J. (2008). Distance weighted discrimination. *Journal of the American Statistical Association* .

Mohar B. (1991). The laplacian spectrum of graphs. In: *Graph Theory, Combinatorics, and Applications*, pp. 871–898. Wiley.

Mohar B. and Juvan N.T.M. (1997). Some applications of laplace eigenvalues of graphs. In: *Graph Symmetry, Algebraic Methods and Applications, volume 497 of NATO ASI Series C*, pp. 227–275. Kluwer.

Muirhead R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Ng A., Jordan M. and Weiss Y. (2002). On spectral clustering: Analysis and an algorithm. In: *Advances in Neural Information Processing Systems 14*, edited by T. Dietterich, S. Becker and Z. Ghahramani, pp. 849–856. MIT Press.

Oloff S., Zhang S., Sukumar N., Breneman C. and Tropsha A. (2006). Chemometric analysis of ligand receptor complementarity: Identifying complementary ligands based on receptor information (colibri). *J. Chem. Inf. Model* **46**, 844–851.

Ong C.S., Canu S. and Smola A.J. (2004). Learning with non-positive kernels. In: *In Proc. of the 21st International Conference on Machine Learning (ICML)*, pp. 639–646.

Paul D. (2005). Asymptotics of the leading sample eigenvalues for a spiked covariance model. Technical report.

Portnoy S. (1984). Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large; i. consistency. *The Annals of Statistics* **12**, 1298–1309.

Portnoy S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics* .

Saul L.K. and Roweis S.T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research* **4**, 119–155.

Schölkopf B. and Smola A.J. (2002). *Learning with Kernels*. MIT press.

Selassie C. (2003). History of quantitative structure-activity relationships. *Burger's Medicinal Chemistry and Drug Discovery* **1: Drug Discovery**, 1–48.

Shi J. and Malik J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 888–905.

Silverstein J. (1985). The smallest eigenvalue of a large dimensional wishart matrix. *The Annals of Probability* **13**, 1364–1368.

Silverstein J. (1989). On the weak limit of the largest eigenvalue of a large dimensional sample covariance matrix .

Telatar E. (1999). Capacity of multi-antenna gaussian channels. *European Transactions on Telecommunications* **10**, 585–595.

Tracy C. and Widom H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics* **177**, 727–754.

v. Luxburg U. (2007). A tutorial on spectral clustering. *Statistics and Computing* **17**, 395–416.

v. Luxburg U., Belkin M. and Bousquet O. (2008). Consistency of spectral clustering. *Annals of Statistics* **36**, 555–586.

Vert J.P. and Kanehisa M. (2002). Graph-driven features extraction from microarray data using diffusion kernels and kernel cca. In: *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, edited by S. Becker, S. Thrun and K. Obermayer, pp. 1425–1432. MIT Press, Cambridge, MA.

Vinod H. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics* **4**, 147–166.

Wagner D. and Wagner F. (1993). Between min cut and graph bisection. In: *Proceedings of the 18th International Symposium on Mathematical Foundations of Computers Science (MFCS)*, pp. 744–750. Springer.

Wang R., Fang X., Lu Y. and Wang S. (2004). The pdbbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem* **47**, 2977–2980.

Yin Y., Bai Z. and Krishnaiah P. (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory Related Fields* **78**, 509–521.

Zelnik-Manor L. and Perona P. (2004). Self-tuning spectral clustering. In: *NIPS 2004*.