Continuum Direction Vectors in High Dimensional Low Sample Size Data

Myung Hee Lee

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 (Statistics).

Chapel Hill
2007

Approved by

Advisor: Dr. J. S. Marron

Reader: Dr. Yufeng Liu

Reader: Dr. Andrew Nobel

Reader: Dr. Ivan Rusyn

Reader: Dr. Haipeng Shen

Reader: Dr. David Threadgill

ABSTRACT

MYUNG HEE LEE: Continuum Direction Vectors in High Dimensional Low Sample Size Data
(Under the direction of Dr. J. S. Marron)

This dissertation consists of three parts regarding High Dimensional Low Sample Size (HDLSS) data analysis. Dimension reduction techniques in high dimensional space, based on a small number of direction vectors, will be the common theme. In the first part of the dissertation, Continuum Regression, originally proposed by Stone and Brooks (1990), will be understood as a family of methods for searching among direction vectors. Continuum Regression includes three popular methods- Ordinary Least Squares, Partial Least Squares, and Principal Component Regression - as special cases. The novel use of Continuum Regression in HDLSS settings will be illustrated by an application to microarray experiments. In the second part of the dissertation, we will extend the Continuum Regression idea to the challenging case of paired HDLSS data. The extended method, Continuum Canonical Correlation, is proposed, as a family of methods for searching direction vectors over two high dimensional spaces simultaneously. The last part of the dissertation studies the HDLSS asymptotic behavior of the maximum covariance direction vectors over two data spaces, i.e., the singular vectors of the sample cross-covariance matrix. We find some conditions under which consistency and strong inconsistency of the singular vectors in HDLSS is observed.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Professor Steve Marron for his guidance and support. I am very grateful for all the effort that he has made for me, especially over the past two years when he has gone through challenging times in his life. I am privileged to be his student.

I would like to deliver thanks to other committee members, Andrew Nobel, Yufeng Liu, Haipeng Shen, David Threadgill, and Ivan Rusyn for their valuable suggestions on this dissertation. Especially, I would like to thank Ivan Rusyn for kindly providing the data which led to an interesting application.

I also thank all my friends in the department and Hanmaum church for helping me out in many ways. I am grateful to Brenda Trapp, whom I met as a host mom. Her warm support helped me adjust to the new culture. I thank my dearest five-month-old son, Nathan Jinseok Lee, for adding a joyous and adventurous dimension to my life. My special thanks go to my husband, Chihoon Lee. Without his unselfish support and countless prayers, getting to this point would have not been possible. Last but not the least, I owe my deepest gratitude to my family, especially my parents, for their love and belief in me.

CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

Data with more variables than observations are emerging in a number of fields. For example, in typical microarray experiments, expression level of numbers of genes ranging from the thousands to the tens of thousands are measured, while the number of observations (i.e., tissue samples) is typically a few tens or hundreds. Data from text recognition and signal processing also often have a much larger dimension d than sample size n. The term High Dimensional Low Sample Size (HDLSS) will be used to refer to this type of data in this dissertation. Another term in use for this type of setting are "large p, small n" .

The statistical analysis of HDLSS data has become a serious challenge to statisticians over a number of years. Classical multivariate tools, originally developed under the assumption of $d < n$, seldom provide satisfying results for HDLSS data. One of the main reasons for this failure is because there is not enough information (observations) to estimate the full underlying covariance matrix. An important step in multivariate analysis is to sphere the data by pre-multiplying the root inverse of the covariance matrix. The fact that the estimated covariance matrices from HDLSS data are inevitably singular hinders this key step in practice.

There have been attempts to modify the existing tools, and also invention of new methodologies, for HDLSS data over the last decades. In the linear regression context, ridge regression (see Hoerl and Kennard (1970); Hastie *et al.* (2001) for useful

introduction and overview), the LASSO (Tibshirani, 1996), and the elastic net (Zou and Hastie, 2005) can be viewed as modified Ordinary Least Squares (OLS) methods, which can be applied to HDLSS input data. The Support Vector Machine (SVM) (see e.g. Vapnik (1982); Vapnik (1995); Shawe-Taylor and Cristianini (2000)) is a clever and powerful discrimination method. Marron *et al.* (2007) developed Distance Weighted Discrimination (DWD), which can be seen as an improved version of the SVM which has good performance with HDLSS data.

Among many attempts to analyze HDLSS data, dimension reduction techniques, based on some directions of interest, are mainly considered in this dissertation. For example, Principal Component Analysis (PCA) finds a set of direction vectors for explaining most of the variability in the data. These direction vectors can be used in several ways. One application is directions for visualization. One can produce 1-dimensional projection plots on those directions or low dimensional projection plots on the subspaces generated by those direction vectors. Another application is to find important variables by sorting on the components of the direction vectors.

HDLSS data also motivate a new type of mathematical theory. Along with the development of new methodologies has come a new family of asymptotics, with the dimension d increasing. More detailed discussion on this topic, along with a literature review, can be found in Chapter 4.

In Chapter 2, Continuum Regression (CR) (Stone and Brooks, 1990) will be viewed as a family of direction searching methods, which includes three popular methods, Ordinary Least Squares (OLS), Partial Least Squares (PLS) and Principal Component Analysis (PCA) as special cases. Each of these will be studied in the HDLSS setting. The novel HDLSS use of this methodology is illustrated by an application to microarray experiments. The change in the relative gene weights, implied by these different directions, will be studied.

The analysis of paired HDLSS data is considered in Chapter 3. When two multi-

variate data sets are obtained in pairs, Canonical Correlation Analysis (CCA) provides a simple method for understanding the connection between the two data sets. PLS and PCA are also generalizable in this scenario. All of these methods will be understood as simultaneous direction searching methods over two high dimensional spaces. As a generalization of CR for paired HDLSS data settings, we propose Continuum Canonical Correlation (CCC), which includes the above as special cases.

In Chapter 4, some asymptotic analysis is developed. We focus on the maximum covariance direction vector for two multivariate data sets. The SVD of the sample cross-covariance matrix provides the solutions to this problem, i.e., singular vectors are the maximum covariance direction vectors. In a spiked marginal population model, we establish the consistency of the sample singular vectors in the sense of convergence in probability. However, in a spherical marginal population model, the sample singular vector is strongly inconsistent to the population singular vector. In both cases, the limiting distribution of the sample singular value is derived.

CHAPTER 2

Continuum Regression

Suppose we have a set of data which consists of n pairs (x_i, y_i) for $i = 1, \cdots, n$. Let $x_i \in \mathbb{R}^d$ represent a regressor vector (i.e., input, i.e., covariates) and $y_i \in \mathbb{R}$ denote the response (output) value from the i-th observation. In the regression problem, the goal is to explain or predict the quantitative response variable Y as a function $f(X)$ of the d-dimensional regressor vector $X = (X_1, \cdots, X_d)'$ based on the training data $\{(x_i, y_i)\}_{i=1}^n$. The linear model assumes that the regression function, or the conditional expectation of $E(Y|X) = f(X)$ has the form $f(X) = \beta_0 + \sum_{j=1}^d X_j \beta_j$. Despite the limitation of model structure, the linear model

(i) is simple and gives an interpretable description,

(ii) has been studied for a long time, so the resulting algorithms are efficient and well understood, and

(iii) can be generalized into non-linear regression via transformation of the regressor variables.

In the example that we have in mind, the response variable Y is a phenotypic measurement assumed to be quantitative. The regressor variables $(X_1, \cdots, X_d)$ are gene expressions from microarrays. The microarray technology enables us to study thousands of genes simultaneously from a sample. Because of its expensive experimental cost, the number of samples, however, reaches only up to a few tens or hundreds.

As a result, in microarray studies, the number of variables, d, oftentimes, is much larger than the number of samples, n. This type of data will be refereed to as High Dimensional Low Sample Size (HDLSS) data in this dissertation.

Probably the most simple way to fit a linear regression function is the Ordinary Least Squares (OLS) method. However, the OLS method directly applied to HDLSS data usually does not provide satisfactory results. See Section 2.2 for details. When the input variables d exceeds the number of data points n, among many alternatives to the OLS method, two commonly considered approaches are the shrinkage method and the method of linear transformation. The former includes ridge regression (see Hoerl and Kennard (1970); Hastie *et al.* (2001) for useful introduction and overview), the LASSO (Tibshirani, 1996), and the elastic net (Zou and Hastie, 2005). All these methods impose a penalty on the regression coefficient size (in L_1, L_2, or in both senses) and these constraints make the solution coefficients shrink.

We focus on linear transformation methods in this dissertation; a detailed explanation can be found in Section 2.3. Special examples include Principal Component Regression (PCR) and Partial Least Squares (PLS). These methods sequentially produce linear transformations (i.e., direction vectors) of the input variables, and use a small number of the direction vectors as regressors. Continuum Regression (CR), proposed by Stone and Brooks (1990), brings OLS, PLS and PCR under one mathematical umbrella. Including these three methods as special cases, they formulated a richer family of regression procedures by introducing a continuous parameter $\alpha \in [0, 1]$, which controls the tradeoff between the covariance (between the input and the output data) and the variance (of the input data). In particular, for $\alpha = 0, 1/2$, and 1 CR corresponds to OLS, PLS, and PCR, respectively. The selection of α and the number of regressors to be considered in regression, can be determined based on Cross Validation.

In this dissertation, however, CR will be applied to microarray data in a slightly differently sense than its original purpose (i.e., predictive modeling). CR will be viewed as a family of direction searching methods. As opposed to the original CR or shrinkage regression methods mentioned in the paragraphs above, choice of a single "best" tuning parameter is not our goal. Rather, we study the entire family of direction vectors over the whole range of the continuum parameter because each of these illustrates a different aspect of the data. This point is reflected in the term "continuum direction vectors" in the title of this dissertation. See Section 2.4 for details.

In this chapter, we will first review CR and its special cases in the original context, i.e., from the regression point of view. We begin the chapter by introducing notation in Section 2.1 and review the OLS and study its behavior in HDLSS settings in the following section. In Section 2.3, we establish a general framework for dimension reduction and will discuss PCR, PLS, and CR, with examples as special cases of this framework. Finally, in Section 2.4, an application of CR to microarray data will be discussed.

2.1 Notation and Assumptions

We denote the regressor (input) variables by $X = (X_1, \cdots, X_d)'$ and a scalar response (output) variable by Y. The data, n samples of those variables, are written in lower case letters $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, \cdots, n$. Bold upper case letters refer to data matrices; we write input and output data matrices as

$$\mathbf{X} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}_{n \times d} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}.$$

Note that the i-th input data is denoted by a column vector x_i, but is stored as a row vector x_i' in the data matrix $\mathbf{X}$ to be consistent with the typical way of writing the data matrix in the linear regression literature.

Column vectors with n components are written in bold lower case, for example, the j-th column of $\mathbf{X}$ is denoted by $\mathbf{x}_j$ for $j = 1, \cdots, d$. It contains n observations on the j-th regressor variable, X_j. This convention distinguishes the data vector on the j-th variable $\mathbf{x}_j \in \mathbb{R}^n$ from the input data from the i-th sample $x_i \in \mathbb{R}^d$.

We assume that the columns of $\mathbf{X}$ (variables) and $\mathbf{Y}$ have been centered to have sample mean zero; i.e., each element of the data matrices, x_{ij} and y_j, have been replaced by $x_{ij} - \frac{1}{n}\sum_{i'=1}^{n} x_{i'j}$ and $y_i - \frac{1}{n}\sum_{i'=1}^{n} y_{i'}$, respectively. Thus, the sample covariance matrix $\mathrm{Cov}(X)_{d\times d}$ and $\mathrm{Cov}(X, Y)_{d\times 1}$ (multiplied by $n-1$) will be denoted as

$$\mathbf{S} = \mathbf{X}'\mathbf{X} \text{ and } \mathbf{s} = \mathbf{X}'\mathbf{Y}. \tag{2.1}$$

Linear regression techniques will be applied to the centered data as the result can always be transformed back to the original data later.

2.2 Ordinary Least Squares

The linear model assumes that the regression function has the form $f(X) = \sum_{j=1}^{d} X_j\beta_j$. There are many ways to estimate the regression coefficients, β_j, and by far the Ordinary Least Squares (OLS) method is the most common and convenient way. The OLS estimates $\hat{\beta}_j$, $j = 1, \cdots, d$ are chosen to minimize the residual sum of squares over the data,

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{d} x_{ij}\beta_j)^2.$$

We can write this in a matrix notation,

$$RSS(\beta) = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

where $\beta = (\beta_1, \beta_2, \cdots, \beta_d)$. Differentiating with respect to β and setting it to be 0, we obtain the normal equations

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta = 0. \tag{2.2}$$

If the inverse of $\mathbf{X}'\mathbf{X}$ exists, then the OLS estimates are uniquely given as $\hat{\beta}^{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. The fitted values at the input training data are

$$\begin{aligned} \hat{Y} &= \mathbf{X}\hat{\beta}^{OLS} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \end{aligned}$$

and at an arbitrary point $x_0 = (x_{01}, \cdots, x_{0d})'$, we use the OLS predictors

$$\hat{Y} = x_0'\hat{\beta}^{OLS} \tag{2.3}$$

as a predicted value.

2.2.1 Geometry in $\mathbb{R}^n$

The geometry of $\mathbb{R}^n$ is helpful to understand how the OLS works with n training data, and it is illustrated in Figure 2.1. $\mathbf{Y} \in \mathbb{R}^n$ denotes the output data vector and $\mathbf{X}\beta = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_d\beta_d \in \mathbb{R}^n$ is a linear combination of column vectors of $\mathbf{X}$. Recall that $\mathbf{x}_j \in \mathbb{R}^n$ is the data vector for the j-th input variable. Let $\mathcal{M}$, represented by the cyan plane in Figure 2.1, be a closed subspace of $\mathbb{R}^n$ that is generated by the n-dimensional data vectors for d input variables, namely,

$$\begin{aligned} \mathcal{M} &= \{\mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_d\beta_d | \beta_j \in \mathbb{R}, j = 1, \cdots, d\} \\ &= \{\mathbf{X}\beta | \beta \in \mathbb{R}^d\}. \end{aligned} \tag{2.4}$$

Now the minimization problem (2.2) can be rephrased as follows; what is the closest element to $\mathbf{Y} \in \mathbb{R}^n$ in the subspace $\mathcal{M}$?

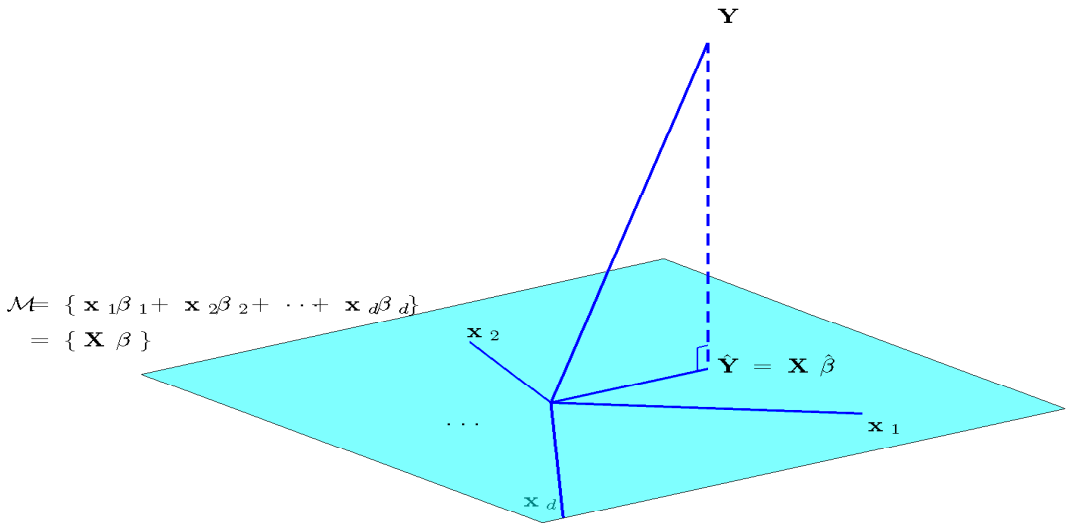

Figure 2.1: Illustrates that the OLS fit $\mathbf{X}\hat{\beta}^{OLS}$ is the projection of $\mathbf{Y}$ onto the space $\mathcal{M} = \{\mathbf{X}\beta | \beta \in \mathbb{R}^d\} \subseteq \mathbb{R}^n$.

The classical projection theorem tells that there is a unique element $\mathbf{X}\hat{\beta} \in \mathcal{M}$ closest to $\mathbf{Y}$ and it is obtained by making the residual vector $\mathbf{Y} - \mathbf{X}\hat{\beta}$ orthogonal to $\mathcal{M}$. The unique element $\mathbf{X}\hat{\beta}$ is called the projection of $\mathbf{Y}$ onto $\mathcal{M}$. The normal equation (2.2) ensures that the least squares fit $\mathbf{X}\hat{\beta}^{OLS}$ is indeed the projection of $\mathbf{Y}$ onto $\mathcal{M}$, $\mathbf{X}\hat{\beta}$, since

$$
\begin{aligned}
&\mathbf{Y} - \mathbf{X}\hat{\beta} \quad \text{is orthogonal to } \mathcal{M}. \\
&\iff \mathbf{x}_j'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0 \qquad\qquad j = 1, \cdots, d. \\
&\iff \mathbf{x}_j'\mathbf{Y} = \mathbf{x}_j'\mathbf{X}\hat{\beta} \qquad\qquad j = 1, \cdots, d. \\
&\iff \mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\beta}.
\end{aligned}
$$

2.2.2 Motivation for dimension reduction

It might happen that the dimension of the subspace, $\mathcal{M}$, is less than d, i.e. $\mathbf{X}$ is not of full column rank because

- regressors have an exact linear dependence (e.g. $\mathbf{x}_1 = 2\mathbf{x}_2$), or
- the number of variables, d, exceeds the number of observations, n.

For both cases, any element in $\mathcal{M}$, say $\mathbf{X}\beta$, can have a different representation $\mathbf{X}\beta^*$, where $\beta^* \neq \beta$. Referring to the geometry of $\mathbb{R}^n$, the projection of $\mathbf{Y}$ onto $\mathcal{M}$, $\hat{\mathbf{Y}}$, is still unique, we just have more than one expression for it in terms of $\mathbf{X}\beta$. The OLS method for these type of data is not satisfactory for predicting the Y value for a new input x_0 since the choice of the OLS solutions β will give an arbitrary prediction $x_0'\beta$.

When there is only an approximate linear dependence among regressors, $\mathbf{X}'\mathbf{X}$ could be nearly singular. The OLS solution to the normal equation (2.2) is unique, but it could be very unstable, resulting in the estimated coefficients having large variance, hence, poor prediction accuracy.

The first case, including the case when regressors have an approximate linear dependence, has been a motivation for dimension reduction in the classical statistics literature. For these data redundancies, it can be helpful to consider a reduced set of regressors. With a subset selection approach, we could select a subset of regressors according to a certain rule, for example, forward or backward selection procedures. Assume for now that the rows of the data matrix are "ordered" in a way that the "reduced" model can be written in the form

$$f(X) = \sum_{i=1}^{\omega} X_i \beta_i \tag{2.5}$$

for some $1 \leqslant \omega \leqslant d$. If ω is too large, then the model will have the same over-fitting problems as the full model, but if too small, under-fitting will make it unlikely that

the model is rich enough to explain important underlying phenomena.

The data dimension reduction schemes which we will study in this dissertation are motivated by the second case. Keeping in mind the applicability of the regression techniques to microarray data, it is likely that many genes are working together for regulating a specific phenotype. If prediction is the only concern, we could reduce dimension by dropping some genes from the model just like (2.5), and improve the prediction accuracy. This approach has the potential risk of discarding important genes from the model. It can be very important to biologists not to miss genes which play important roles. The way we reduce the data dimensionality should be different from (2.5) in this case, and we defer the detailed review of this issue to Section 2.3.

2.2.3 The choice of OLS solution for HDLSS case

As we saw in the previous section, the OLS solution for HDLSS is not uniquely determined. In some situations, however, we might want to choose an OLS estimate to make a comparison with other methods. When we do not have enough information to specify a solution, an approach to resolve this is to restrict or bias the solution in some way. Two ways of resolving this are discussed in the following sections.

Restriction of the subspace

If we restrict the OLS solution, β, to be in the subspace, $\mathcal{U} \subseteq \mathbb{R}^d$ generated by the d- vectors of training inputs, i.e.,

$$\begin{aligned} \mathcal{U} &= \{\textstyle\sum_{i=1}^{n} x_i\alpha_i | \alpha_i \in \mathbb{R}, i = 1, \cdots, n\} \\ &= \{\mathbf{X}'\alpha | \alpha = (\alpha_1, \cdots, \alpha_n)' \in \mathbb{R}^n\}, \end{aligned}$$

then we now project $\mathbf{Y}$ onto the subspace $\mathcal{M}^* = \{\mathbf{X}\beta | \beta \in \mathcal{U}\} \subseteq \mathcal{M} = \{\mathbf{X}\beta | \beta \in \mathbb{R}^d\}$. Thus, the least squares problem becomes

$$\arg\min_{\alpha \in \mathbb{R}^n} (\mathbf{Y} - \mathbf{X}\mathbf{X}'\alpha)'(\mathbf{Y} - \mathbf{X}\mathbf{X}'\alpha).$$

Assuming that $\mathbf{XX'}$ is invertible, the solution is given as $\hat{\alpha} = (\mathbf{XX'})^{-1}\mathbf{Y}$. Call $\hat{\beta}^{dual} = \mathbf{X'}\hat{\alpha} = \mathbf{X'}(\mathbf{XX'})^{-1}\mathbf{Y}$.

Ridge solution and its dual solution

The Ridge Regression solution $\hat{\beta}^{Ridge} = (\mathbf{X'X} + \lambda\mathbf{I})^{-1}\mathbf{X'Y}$ was proposed by Hoerl and Kennard (1970) as a remedy to instability of the OLS solution when $\mathbf{X'X}$ is singular or nearly singular. The $\hat{\beta}^{Ridge}$ solves a modified least squares problem,

$$\hat{\beta}^{Ridge} = argmin_{\beta}(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) + \lambda||\beta||^2$$

where $\lambda > 0$, called the ridge parameter, controls the balance between the residual sum of squares and the squared length of the parameter vector β; the ridge regression solution is *shrinked toward 0* in the sense that, the λ larger, the length of β is smaller. When $\lambda = 0$, we get back to the OLS problem. A dual version of the ridge solution, (Shawe-Taylor and Cristianini (2004) and Saunders *et al.* (1998)), namely, as a form of $\hat{\beta}^{Ridge} = \mathbf{X'}\alpha = \sum_{i=1}^{n} x_i\alpha_i$ for some $\alpha \in \mathbb{R}^n$, is obtained with $\alpha = (\mathbf{XX'} + \lambda\mathbf{I})^{-1}\mathbf{Y}$ for $\lambda \geqslant 0$. The OLS dual solution corresponds to the special case of ridge dual solution with the choice of $\lambda = 0$. From now on, we let $\hat{\beta}^{OLS} = \mathbf{X'}(\mathbf{XX'})^{-1}\mathbf{Y}$.

2.3 Dimension Reduction

In this section, a general approach (Stone and Brooks, 1990) to reduction of the dimensionality for HDLSS data, is introduced. The idea behind this approach is that

- if we can summarize data set of high dimensional covariates, $X = (X_1, \cdots, X_d)'$, into linearly independent factors, $Z = (Z_1, \cdots, Z_\omega)'$, where $Z_m = v_m'X$, a linear transformation of the covariates for some $v_m \in \mathbb{R}^d$, (e.g. v_m could be PCR or PLS direction), for $m = 1, \cdots, \omega \ll d$,
- the small number of factors, Z, in the regression analysis, instead of using the full set of X, could avoid the overfitting problem.

This framework contains Principal Component Regression (PCR), Partial Least Squares (PLS), and Continuum Regression (CR) as special cases.

2.3.1 Linear Transformation of the Input Data

Motivated by thinking of (2.5), we now consider a reduced regression model

$$f(Z) = \sum_{m=1}^{\omega} Z_m \theta_m, \tag{2.6}$$

where $Z_m = v_m' X$, is a linearly transformed variable for $v_m \in \mathbb{R}^d$ and θ_m is a new model parameter for $m = 1, \cdots, \omega$, for some $1 \leqslant \omega < d$.

Now the important questions are:

- how to choose useful vectors v_m for $1 \leqslant m \leqslant \omega$
- how many transformed variables do we need, namely, what is the value of ω?

The discussion of the second question is important, we briefly mention a suggested approach to the selection of ω in following sections, but will not focus on this issue in this dissertation.

For the first question, we impose two reasonable restrictions on the v_m;

R1 The v_m is a unit vector (i.e., $||v_m|| = 1$) for $m = 1, \cdots, \omega$. The v_m can be thought of as a direction vector in the d-dimensional space. Another reason we keep $||v_m|| = 1$ is to make θ_m identifiable.

R2 The $v_1, \cdots, v_\omega$ are **S**-orthogonal to each other, meaning that $v_m' \mathbf{S} v_l = 0$ for $1 \leqslant l < m \leqslant \omega$. It appears natural to make direction vectors "orthogonal" to each other in some sense and this constraint, **S**-orthogonality, ensures that the current variable Z_m is uncorrelated with previously constructed variables,

$Z_1 \cdots, Z_{m-1}$ over the data;

$$\begin{aligned}(n-1)\mathrm{Cov}(Z_m, Z_l) &= (\mathbf{X}v_m)'\mathbf{X}v_l \\ &= v_m'\mathbf{S}v_l \\ &= 0\end{aligned}$$

for $l = 1, \cdots, m-1$, where Cov is the sample covariance over the data.

With $v_1, \cdots, v_\omega$ fixed, the coefficients $\theta^\omega = (\theta_1, \cdots, \theta_\omega)'$ are estimated by the OLS method with the transformed input data,

$$\mathbf{Z}_\omega = \mathbf{X}\mathbf{V}_\omega, \text{ where } \mathbf{V}_\omega = (v_1, \cdots, v_\omega), \tag{2.7}$$

hence, the least squares solution is given as

$$\hat{\theta}^\omega = (\mathbf{Z}_\omega'\mathbf{Z}_\omega)^{-1}\mathbf{Z}_\omega'\mathbf{Y}.$$

Note that the matrix $\mathbf{Z}_\omega'\mathbf{Z}_\omega$ is a diagonal matrix by the $\mathbf{S}$-orthogonality constraint. This will lead to a computational saving when fitting multiple regression; the OLS solution of the m-th coefficient in the multiple regression fit

$$\hat{\theta}_m = \frac{\mathbf{z}_m'\mathbf{Y}}{\mathbf{z}_m'\mathbf{z}_m}$$

is just the OLS solution of the regression fitted with a single regressor Z_m. Namely, the regression fit of Y in the ω-dimensional subspace is the sum of ω univariate fits on each transformed variable, Z_m, separately.

The fitted value can be written in term of the original data,

$$\begin{aligned}\hat{\mathbf{Y}}_\omega &= \mathbf{Z}_\omega \hat{\theta}^\omega \\ &= \mathbf{X}\mathbf{V}_\omega \hat{\theta}^\omega \\ &= \mathbf{X}\hat{\beta}^\omega\end{aligned}$$

where $\hat{\beta}^\omega = \sum_{m=1}^{\omega} \hat{\theta}_m v_m$ is the corresponding linear transformation of the direction vectors $v_1, \cdots, v_\omega$. The estimated coefficient in terms of the original input variables, $\hat{\beta}^\omega$, now can be used to predict the value of Y at a general point $x_0 = (x_{01}, \cdots, x_{0d})'$ as

$$\hat{Y} = x_0' \hat{\beta}^\omega. \tag{2.8}$$

Geometry in $\mathbb{R}^n$

The least squares problem with the transformed input data matrix $\mathbf{Z}_\omega$ can be viewed as a projection problem of $\mathbf{Y} \in \mathbb{R}^n$ onto the space

$$\mathcal{M}^* = \{\mathbf{z}_1\theta_1 + \cdots + \mathbf{z}_\omega\theta_\omega | \mathbf{z}_m = \mathbf{X}v_m, \theta_m \in \mathbb{R}, m = 1, \cdots, \omega\},$$

rather than $\mathcal{M} = \{\mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_d\beta_d | \beta_j \in \mathbb{R}, j = 1, \cdots, d\}$. A set of linearly transformed data, $\{\mathbf{z}_m\}_{m=1}^M$, where $M \equiv \dim(\mathcal{M})$ is an orthogonal basis of the subspace, $\mathcal{M}$, so, clearly, the choice of $\omega = M$ will get back to the original OLS regression problem.

The idea of data reduction is to use the first ω vectors $\{\mathbf{z}_m\}_{m=1}^\omega$ for regression, but discard $M - \omega$ vectors $\{\mathbf{z}_m\}_{\omega+1}^M$. By doing so, we project $\mathbf{Y}$ on a subspace $\mathcal{M}^* \subset \mathcal{M}$.

The regression procedures which we will discuss in the following sections differ by the way they construct the direction vectors, $\{v_m\}_{m=1}^\omega$; hence, the transformed data $\{\mathbf{z}_m\}_{m=1}^\omega$.

2.3.2 OLS Revisited

For understanding the relationship between OLS and the data reduction scheme introduced in the previous section - OLS is presented as a special case of the general method. The geometry of OLS on page 8 informs that $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}^{OLS}$, the projection of $\mathbf{Y}$ onto $\mathcal{M}$, creates the maximum angle between $\mathbf{Y}$ among all the elements in $\mathcal{M}$. When studying angles, we observe that the square of the sample correlation between $v'X$ and Y emerges naturally via the following relationship;

$$\begin{aligned} \text{Corr}^2(v'X, Y) &= \frac{\text{Cov}^2(v'X, Y)}{\text{Var}(v'X)\text{Var}(Y)} \\ &= \frac{< \mathbf{X}v, \mathbf{Y} >^2}{||\mathbf{X}v||^2||\mathbf{Y}||^2} \\ &= \cos^2\theta, \end{aligned} \tag{2.9}$$

where θ is the angle between $\mathbf{X}v \in \mathcal{M}$ and $\mathbf{Y}$ in $\mathbb{R}^n$. Since it is scale invariant, $v_1 = \hat{\beta}^{OLS}/||\hat{\beta}^{OLS}||$ solves the optimization,

$$v_1 = \arg\max_{||v||=1} \text{Corr}^2(v'X, Y). \tag{2.10}$$

To search for the next direction vector, v_2, let v denote any unit vector that satisfies $v'\mathbf{S}v_1 = 0$. Then,

$$\begin{aligned} < \mathbf{X}v, \mathbf{Y} > &= v'\mathbf{X}'\mathbf{Y} \\ &= v'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{Y} \\ &= ||\hat{\beta}^{OLS}||v'\mathbf{S}v_1 \\ &= 0. \end{aligned}$$

Thus, there is no well defined maximum of (2.9) for v satisfying $v'\mathbf{S}v_1 = 0$, and no gain for adding more transformed variables $v'X$ to the model. So, the sequential construction defined in Section 2.3.1 terminates with just the first direction vector,

v_1, i.e. $\mathbf{X}v_1$ summarizes the full explanatory (of $\mathbf{Y}$) power of $\mathbf{X}$.

2.3.3 Principal Component Regression

Following the general framework for data reduction in Section 2.3.1, the $m-$th direction vector for PCR, v_m is chosen to maximize the sample variance of $v'X$, under the constraint $||v|| = 1$ and $\mathbf{S}$-orthogonality with the $m-1$ previous vectors;

$$\begin{aligned} v_m &= \arg\max_v \mathrm{Var}(v'X) \\ &= \arg\max_v \textstyle\sum_1^n (v'x_i)^2 \\ &= \arg\max_v v'\mathbf{S}v, \\ \text{subject to } ||v|| = 1, &\quad v'\mathbf{S}v_l = 0, \mathrm{l} = 1, \cdots, m-1, \end{aligned} \tag{2.11}$$

Principal Component Analysis

The method of Principal Component Analysis (PCA) is often used in multivariate analysis (Anderson, 1958) as a method of dimension reduction: given a large number of measurements on each observation, PCA tries to find a small number of linear combinations of the measurements, called principal components, which explain most of the variability across the observations.

In more detail, PCA begins with the eigenvector decomposition of the sample covariance matrix (multiplied by $n-1$),

$$\mathbf{S} = \lambda_1 v_1 v_1' + \cdots + \lambda_M v_M v_M'$$

where $M = \mathrm{rank}(\mathbf{S})$, $\lambda_1 \geqslant \cdots \geqslant \lambda_M \geqslant 0$ are the eigenvalues of $\mathbf{S}$, and $\{v_1, \cdots, v_M\}$ are their corresponding eigenvectors of length 1. (i.e., $||v_1|| = \cdots = ||v_M|| = 1$). The eigenvectors, $\{v_m\}$, are called Principal Component (PC) direction vectors of $\mathbf{X}$ and the entries the loadings. The projection of data onto the m-th PC direction vector, $\mathbf{z}_m = \mathbf{X}v_m$, is the m-th PC of $\mathbf{X}$.

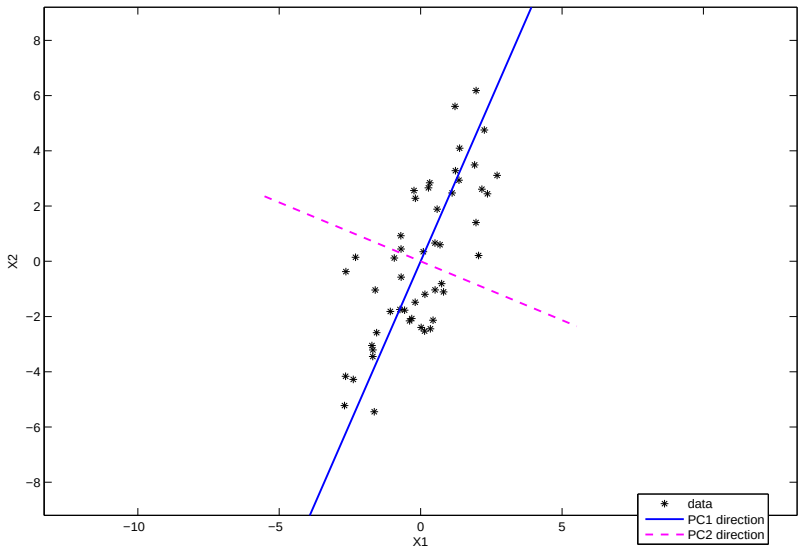

Figure 2.2: Depicts the construction of the PC direction vector. The first direction accounts for most variation of the data and the second direction is orthogonal to the first.

In particular, the m-th eigenvector, v_m, is the solution to

$$\begin{aligned} \arg\max_v v'\mathbf{S}v &= \arg\max_v \textstyle\sum_{i=1}^{n}(v'x_i)^2 \\ &= \arg\max_v \mathrm{Var}(v'X) \\ \text{subject to } ||v|| = 1 \text{ and} \quad & v'v_l = 0 \text{ for } l = 1, 2, \cdots, m-1, \end{aligned} \tag{2.12}$$

where Var is the shorthand for the variance over the sample. Put in words, the measurements on observations spread out the most along the first PC direction vector v_1 in the d-dimensional space. The second PC direction vector, v_2, is the maximal variance amongst all direction vectors perpendicular to v_1 as shown in Figure 2.2.

Note that a unit vector v orthogonal to the first eigenvector v_1 is also $\mathbf{S}$ orthogonal to v_1, and vice versa. So, for $m = 2$, if we replace the orthogonality condition, $v'v_1 = 0$, in (2.12) by the $\mathbf{S}$-orthogonality, $v'\mathbf{S}v_1 = 0$, the second eigenvector of $\mathbf{S}$, v_2, remains

the solution of the optimization. The same argument can be used subsequently to show that the m-th eigenvector of $\mathbf{S}$, v_m, satisfies

$$\begin{aligned} &\arg\max_v \mathrm{Var}(v'X) \\ &\text{subject to } ||v|| = 1 \text{ and } v'\mathbf{S}v_l = 0 \text{ for } l = 1, 2, \cdots, m-1. \end{aligned} \tag{2.13}$$

The variance of the data along the m-th PC direction is

$$\begin{aligned} \mathrm{Var}(v'_m X) &\propto \textstyle\sum_{i=1}^{n} (x'_i v_m)^2 \\ &= v'_m \mathbf{S} v_m \\ &= \lambda_m. \end{aligned}$$

The amount of variability that can be explained by the direction v_m is reflected on the magnitude of its corresponding eigenvalue. If eigenvalues beyond the m-th are small, we might expect that discarding the principal components corresponding to or beyond the m-th might not lead to loss of too much information.

Principal Component Regression

Principal Component Regression (PCR) (Massy, 1965), regresses the output variable onto the first ω PCs, which contain most of the variability in the original input variables. For the choice of ω, Cross-Validation (CV) (Stone, 1974) has been suggested as an approach to this (Frank and Friedman, 1993).

The obvious disadvantage of PCR is that there is no reason why the first few PCs, that are important to explain the variability of the input variables, will take into account the output Y. This approach could discard a low X-variance direction, which, in fact, could carry the important information about Y. To facilitate this idea, a toy example showing this deficiency is shown in the next paragraph.

Toy Example

We generate a 2−dimensional regression data set $\{(x_{i1}, x_{i2}, y_i)\}$ of size 50. The input data, $\{(x_{i1}, x_{i2})\}$, are seen in Figure 2.2. This plot gives an image of how the regression data $\{(x_{i1}, x_{i2}, y_i)\}$ are distributed in the 2-dimensional input space only.

The movie, Lee (2005b), shows the data $\{(x_{i1}, x_{i2}, y_i)\}$ in the 3-dimensional space, from a rotating viewpoint. This can be helpful to have a feeling about how the data appear in the 3-dimensional space. Notice that the data appear to spread more along the X_2 axis than the X_1 axis. As the view angle spins, the data points, sometimes, appear to lie near a line (correlation ≈ 1), but sometimes they seem to spread very randomly (correlation ≈ 0).

In the movie, Lee (2005c), the viewpoint is fixed, but the direction vector in the X space, v, denoted as the blue line, is rotating to search for the maximal input data variation vector. Figure 2.3 is a snapshot from this movie. The left panel illustrates how the dimension reduction of the input data can be attained via the linear transformation. The right plot shows the resulting projected data and linear regression. For an arbitrarily chosen direction vector, v, the corresponding linear transformed variable is given as $Z_1 = v'X$. The regression of Y on Z_1 now is a simple linear regression, and the data pairs for this regression analysis are obtained by projecting data on the 2-dimensional face determined by Z_1 and Y, which is denoted as the cyan phase. The projections - denoted as circles - are linked with the original data - denoted as crosses- for $i = 1, \cdots, 50$. For better demonstration of the regression analysis with the transformed variable, the blue plane is extracted and displayed on the right hand side.

The rotation stops at the vector v_1 (PC1 direction vector) to maximize the variation of the input data. One can see that the performance of the regression on the first PC very poor.

Figure 2.4 demonstrates how the important component concerning the variability

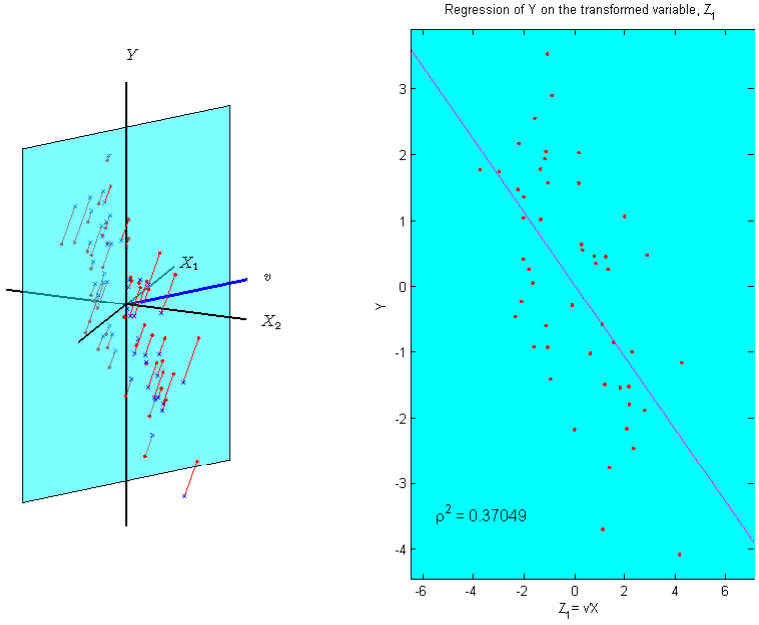

Figure 2.3: Toy Example with 2−dimensional regression data $(x_{i1}, x_{i2}, y_i)\}_1^{50}$. A snapshot from the movie, Lee (2005c). *Left*: A subspace, represented by the cyan plane, is generated by the arbitrarily chosen direction vector, v, and the output variable, Y. Each data point- blue cross- and its projection on this plane- red circle- are linked together by red lines. *Right*: The blue plane pulled from the left side. The red dots form the data for the 1-dimensional regression analysis using the transformed variable, $v'X$ as a regressor.

of the input data could be irrelevant to the output variable. On the top left panel, the two PC vectors, v_1 and v_2 are shown in the input space. Having fixed the direction vector as the first PC vector, regression on the first PC is done on the top right panel (as we saw in Figure 2.3). The fitted regression line is overlayed as the magenta line. The bottom plots are for the second PC. We hardly see any pattern between the first PC and the output on the top right. But the low variance component on the bottom right is much more relevant to the output variable, Y, than the first PC.

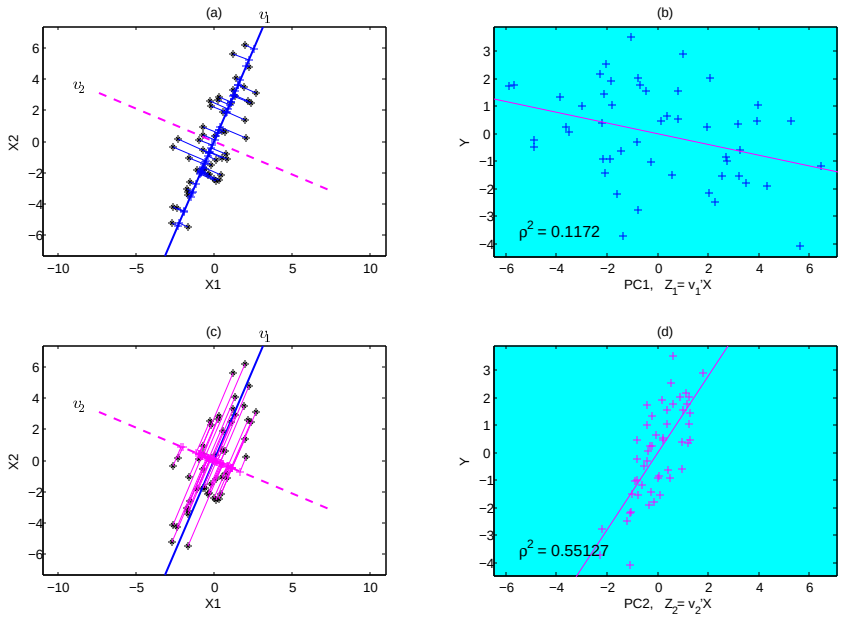

Figure 2.4: Toy Example with the same 2−dimensional regression data $(x_{i1}, x_{i2}, y_i)\}_1^{50}$ as in Figures 2.2 and 2.3. (a) Projection of input data onto the first PC direction vector in the (X_1, X_2) plane (b) Scatter plot of the first PC against output data (c) Projection of input data onto the second PC direction vector (d) Scatter plot of the second PC against output data, showing that it is the second PC that carries the important information about the output data (i.e., much stronger correlation with Y).

2.3.4 Partial Least Squares

Since Partial Least Squares (PLS) introduced by Wold (1976) in an algorithmic form, a variety of different algorithms (Naes and Martens (1985), Helland (1988)) that produce the same solutions have been proposed. It is very popular in the field of chemometrics, where HDLSS settings which often lead to multicollinerity between variables are commonplace.

PLS can be understood as an optimization problem (as was PCR in 2.13), specif-

ically the m-th PLS direction vector, v_m, solves

$$\begin{aligned} \arg\max_v \mathrm{Cov}^2(v'X, Y) &= \arg\max_v \mathrm{Corr}^2(v'X, Y)\mathrm{Var}(v'X) \\ &= \arg\max_v < \mathbf{X}v, \mathbf{Y} >^2 \\ &= \arg\max_v (v'\mathbf{s})^2 \\ \text{subject to } ||v|| = 1, v'\mathbf{S}v_l = 0, &\quad l = 1, \cdots, m-1 \text{ where } \mathbf{s} = \mathbf{X}'\mathbf{Y}. \end{aligned} \tag{2.14}$$

In contrast with the PC direction, the PLS direction is found in connection with the output variable. The first PLS direction v_1 maximizes the covariance between the output variable and the linearly transformed variable, $v'X$ over the data.

An interesting connection among the criteria for OLS in (2.10), PCR in (2.13), and PLS in (2.14) can be found: the criteria for PLS, $\mathrm{Corr}^2(v'X, Y)\mathrm{Var}(v'X)$ is the square of the geometric mean of criterion for OLS and PCR. This suggests that PLS can be regarded as a compromise between OLS and PCR. This point is illustrated in Figure 2.5. With the simulated data that we employed in the previous section, the first OLS, PCR, and PLS vectors are drawn. The PLS vector lies between the OLS and the PCR vectors. Unlike PCA, PLS makes direct use of the information about the output variable, but to a smaller extent than OLS does.

While building up the PLS vectors sequentially, we can add the corresponding factors in the regression analysis as explanatory variables. The Cross Validation criteria is one way to choose the number of PLS factors that needs to be included in the regression (Frank and Friedman, 1993).

2.3.5 Continuum Regression

Continuum Regression (CR) proposed by Stone and Brooks (1990) is a general procedure to reduce a regression model in terms of linear transformations of the original regressors as introduced in (2.6). With a criteria varying with a parameter $\alpha \in [0, 1]$, CR constructs regressors sequentially. In particular, the family of built

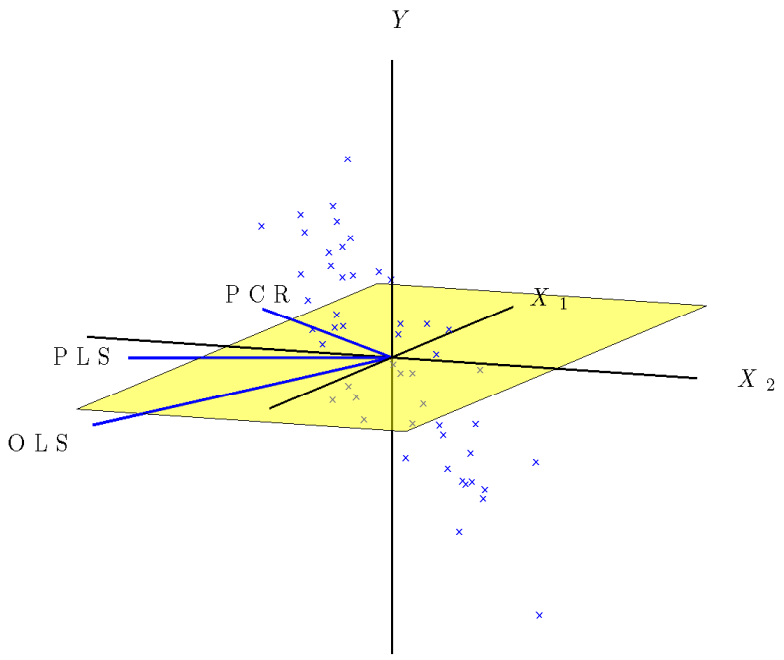

Figure 2.5: Toy Example with the same 2−dimensional regression data $\{(x_{i1}, x_{i2}, y_i)\}_1^{50}$ as in Section 2.3.3. The cyan plane stands for the X-space. The first direction vectors for OLS, PLS, and PCR are drawn. The OLS direction is nearly orthogonal to the PCR direction, and the PLS direction lies between the OLS and the PCR directions.

regressors embrace OLS, PLS, and PCR.

The fact that OLS, PLS, and PCR differ only in one aspect - the target quantity maximized at each step - was pointed out by Stone and Brooks (1990). Based on this analogy, they formulated a richer family of regression methods encompassing those three methods. Namely, for each $\alpha \in [0, 1)$, a direction vector $v_m(\alpha)$ is found

sequentially to maximize

$$\begin{aligned} T(\alpha) &= \mathrm{Cov}^2(v'X, Y)\mathrm{Var}(v'X)^{\alpha/(1-\alpha)-1} \\ &= (v'\mathbf{s})^2(v'\mathbf{S}v)^{\alpha/(1-\alpha)-1} \qquad (2.15) \\ \text{subject to} \quad & ||v|| = 1 \text{ and } v'\mathbf{S}v_l' = 0 \text{ for } l = 1, \cdots, m-1. \end{aligned}$$

Here, Cov and Var are the sample covariance and sample variance over the data. For $\alpha = 0$,

$$\begin{aligned} \arg\max_v T(0) &= \arg\max_v \mathrm{Cov}^2(v'X, Y)\mathrm{Var}(v'X)^{-1} \\ &= \arg\max_v \mathrm{Corr}^2(v'X, Y). \end{aligned}$$

For $\alpha = 1/2$,

$$\arg\max_v T(\frac{1}{2}) = \arg\max_v \mathrm{Cov}^2(v'X, Y).$$

Clearly, CR for $\alpha = 0$ and $\alpha = \frac{1}{2}$ corresponds to OLS and PLS, respectively. The PCR, however, can be understood only in the limiting sense as $\alpha \to 1$. But, we can not expect that PCR is always the limit of CR as α goes to 1. In particular, even though the function (2.15) is continuous with respect to α, it is not generally true that the maximizer, $v(\alpha) \in \mathbb{R}^d$, as a function of α, is also continuous with respect to α. Björkström and Sundberg (1996) demonstrates that CR can yield a discontinuous maximizer $v(\alpha)$, as a function of α.

Having found the direction vector, $v_m(\alpha)$ at the m-step for a fixed value of α, the linearly transformed regressor, $v_m'X$ is added to the regression analysis. Thus, the CR procedure gives rise to a set of linearly transformed regressors,

$$\{(v_1(\alpha)'X, \cdots, v_\omega(\alpha)'X) | \alpha \in [0,1], 1 \leqslant \omega \leqslant M\}, \text{ where } M = \text{ rank } (\mathbf{X}).$$

For the choice of the two tuning parameters to be set- $\alpha \in [0,1]$, a continuum parameter indicating a regression procedure and ω, the number of regressors to be included in the regression- the values of (α, ω) that gives the smallest leave-one-out

Cross-Validation error is suggested by Stone and Brooks (1990).

Toy Example

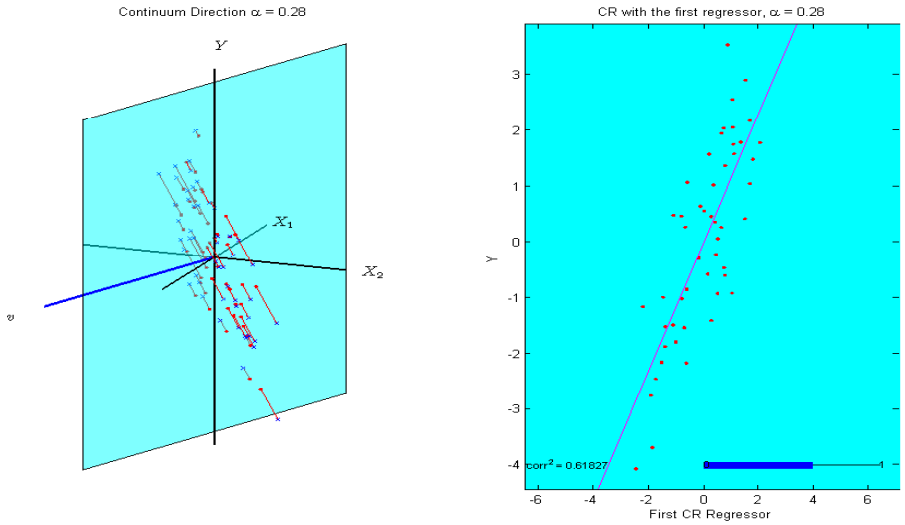

Figure 2.6: Toy Example with the same 2−dimensional regression data $(x_{i1}, x_{i2}, y_i)\}_1^{50}$ as in Section 2.3.3. *Left:* The CR direction vector for a given α is found and plotted as the blue line. The projections of input data on this direction, in pairs of output data, are plotted as red circles. *Right:* The blue plane from the left hand side shows the CR on the corresponding CR factor.

Using the same data as in Section 2.3.3, the first direction and the corresponding first factor of CR are calculated as a function of α in the movie, Lee (2005d). Figure 2.6 is a snapshot of this movie. Data pairs, (x_{i1}, x_{i2}, y_i) are shown as crosses in the 3-dimensional space. For a fixed value of α, the direction vector, v, maximizing the objective function in (2.15) is found and represented as the blue line. By projecting the input data (x_{i1}, x_{i2}) onto this direction vector, we obtain the first CR factor, which in pairs with the output data are plotted as dots. As a result, these reside on the cyan face, the span of the first CR factor and the output variable. The plot on the right hand side is basically the blue plane in the left panel. The solid line is the fitted

CR on the first factor. The blue thick bar at the bottom indicates the magnitude of the sample correlation between the output variable and the first CR factor. This movie shows the full span of the CR vectors from the OLS vector to the PCR vector.

2.4 Application to Microarray Data

This example involves a large input data set with $n = 36$ mouse samples and $d = 17000$ gene expression measurements from microarrays. The microarray measures mRNA abundance which is used to derive the level of expression for thousands of genes simultaneously. Since the activity of a gene, represented by the quantity of mRNA, reflects the molecular status of the sample, gene expression profiles can be used to classify the different subtypes of disease. There have been many attempts to adapt statistical tools for discrimination/ clustering for performing this type of diagnosis. When we have an accompanying variable that characterizes the same samples, such as survival time or other quantitative measurements, a common statistical task is to build a prediction model that uses the gene expression level, for a sample as input to predict the output value.

CR can be seen as a method to unify these two types of tasks in a way that solely unsupervised (PCR) and solely supervised (OLS) tasks occupy the two ends of the CR spectrum and that puts an interesting range of intermediate methods, (e.g. PLS), between them.

For this example, the maximizer direction vector will be computed for each value of $\alpha = 0, .1, \cdots, 1$. Entries of a direction vector can be interpreted as the contribution of genes on the particular direction vector. Sorting the entries of each direction vector in an decreasing order gives the rank of gene contributions on each direction vector. Note that genes with the largest positive (negative) entries are the ones which influence most on the given direction vector.

We will describe the data as well as the mouse experiment done by the Rusyn

lab in Section 2.4.1. In Section 2.4.2, we will use the continuum parameter, α, in a novel way, to study how gene ordering changes over the α spectrum by creating "rank tracking plots". To identify biologically relevant genes, the tracking visualization will be used in a different way in Section 2.4.3.

2.4.1 Experiment and Data

A mouse experiment was conducted in the Rusyn lab to study the genetic factors affected by exposure to alcohol that may contribute to liver disease. A panel of six inbred mouse strains (A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6J, and DBA/2J) was exposed to alcohol acutely as a bolus of 5g/kg intra-gastric dose for 6 hours. Liver and blood was taken from the alcohol-treated mice and controls to assess several measurements. Several phenotypic changes (associated with alcohol toxicity in liver) were measured from control and alcohol-treated mouse samples. In addition to that, gene expression profiling was performed from the mouse samples above.

For the application of CR, we consider the d genes in the microarray as input variables, $(X_1, \cdots, X_d)$, and the Blood Alcohol Concentration (BAC), a continuous phenotype, as the response variable, Y.

The heat map of the gene profiles (a typical way to visualize microarray data) is shown in Figure 2.7. Gene expression values in the input data matrix $\mathbf{X}$ are converted to colors ranging from green (negative) to red (positive) and displayed in rectangular pixels so that each row represents a sample and each column corresponds to a gene. Due to the resolution, not all 17,000 genes, but a subset of genes are seen. The intensities of red and green reflect the magnitude of the absolute values of gene expression levels. The rows and columns are ordered, using the clustering method in TreeView (EisenLab, 2002) so that samples and genes with similar patterns of expressions can be located close to each other. Note that within each of the six mouse strains (A/J, BALB/CJ, C3H/HeJ, DBA/2J, AKR/J, and C57BL/6J) the data are clustered with

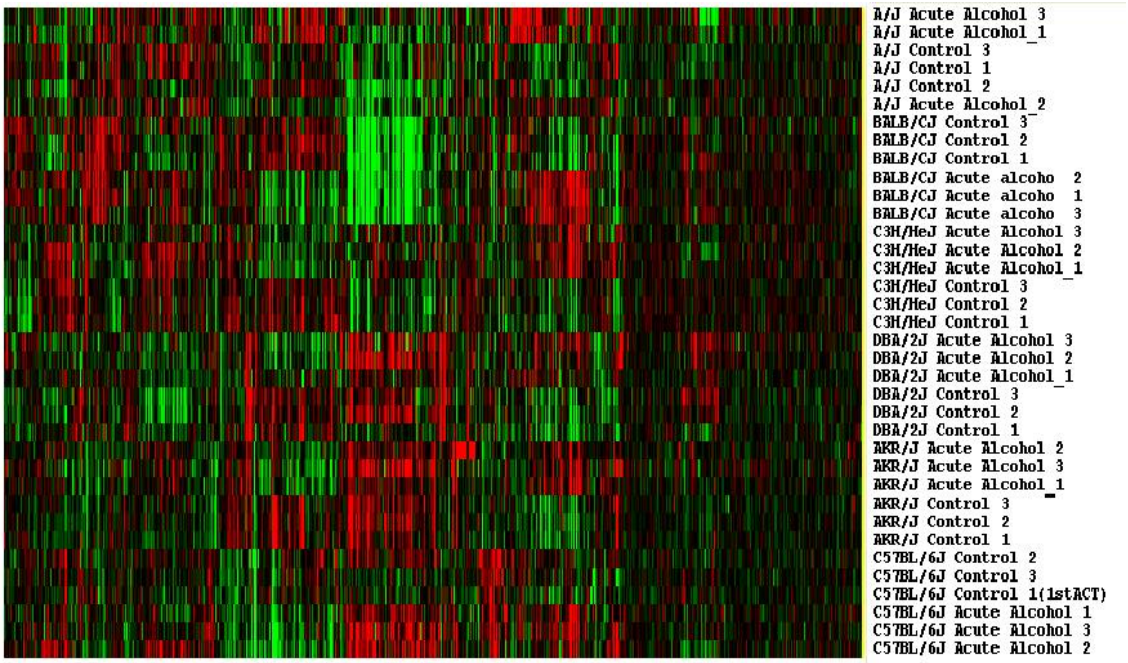

Figure 2.7: Heat map view of Gene Expression from 36 mouse samples. Columns represent genes (variables) and rows display gene profiles from samples (observation). Note that transposing this map is the usual way of displaying. We take this view because the mouse strain labels on the right are clearly labeled.

each other. This indicates that the strain factor dominates the treatment factor. Also, note that all the alcohol-treated samples and the control samples within a strain are clustered next to each other except for one strain, A/J. This indicates that this experiment is very replicable. Genes with similar expression patterns are also clustered together. In the middle, one can see a vertical stripe pattern- bright greens for BALB/CJ, reds for the bottom three strains (DBA/2J, AKR/J, and C57BL/6J) and somewhat green for the rest of two strains (A/J, C3H/HeJ). This stripe pattern seems to dominate other patterns of genes.

Figure 2.8, a PCA scatter plot, shows expression patterns more clearly via low dimensional projections. Strains and treatments are indicated by different colors and symbols. The first 4 PC directions vectors are computed. Gene expressions from 36 mouse samples are projected onto those directions and plotted on the diagonals. As

pointed out in the paragraph above, one can clearly see that the strain, BALB/CJ, is far from the rest of the strains in the projection plot on the 1st PC. In the 2nd diagonal plot, the circles (controls) and crosses (acute treatments) are nicely separated. This shows that the treatment factor also explains the variability across the samples, having adjusted the variability by the first PC. Off-diagonals are 2-dimensional projection plots on the subspaces spanned by each pair of the 4 PC directions. For example, the plot on the top and the second from the left is the projections of the expressions on the subspace generated by PC1 and PC2 direction vectors. From the 2-dimensional projection plots, one can see that

1. the same colors (strains) tend to be close to each other

2. within each color cluster, samples are grouped as symbols (treatments)

The main lessons from the PCA scatter plot are similar to the lessons from the heat map, but some patterns are more clear through appropriate low dimensional projections.

The top panel in Figure 2.9 is the bar-graph of the blood alcohol concentration (BAC) data (Y) from the samples with the microarray data and the bottom the sample-mean subtracted BAC data. The bottom panel shows the distribution of phenotypes across the mouse samples. The main variability clearly is due to treatment effect. Note that, however, variability across strains within alcohol-treated samples is noticeably bigger than variability in the control. This means that the phenotypes are affected by genetic differences. OLS, a way to associate expression data with phenotypes, clearly differs from the discrimination study in the sense that it takes into account the strain variability within treatment/contol samples.

2.4.2 CR Analysis

In this section, CR is used to find "interesting genes". Each entry of the CR direction vectors, which will be referred to as gene loading in this dissertation, can be

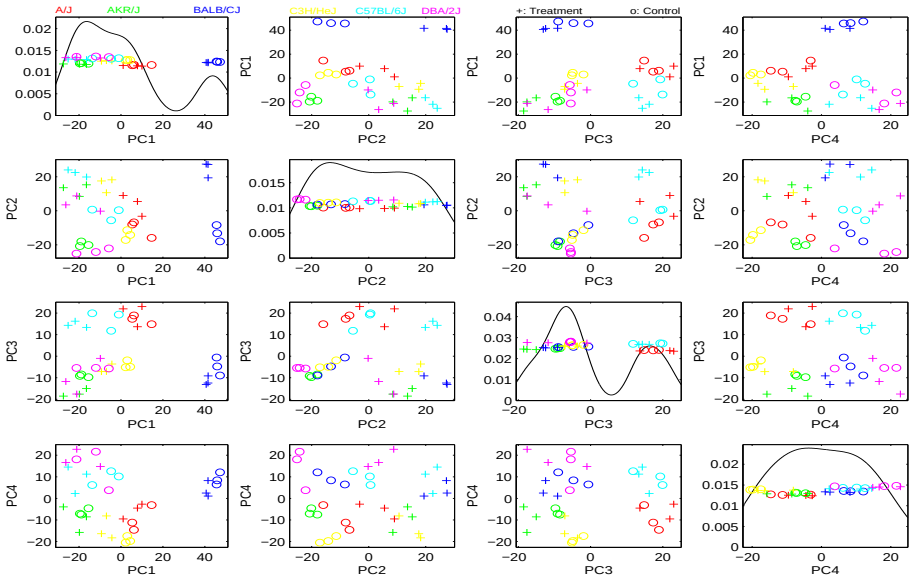

Figure 2.8: PCA Scatter plot of Gene Expression data. Colors and symbols indicate strains and treatments, respectively. Diagonals are 1-dimensional projections on the first 4 PC directions and off-diagonals are 2- dimensional projections on the subspaces generated by those directions.

seen as a contribution of the corresponding gene on the first CR vector. Suppose, for example, the first CR vector for $\alpha = 1$, i.e., the first PC vector, is $v_1(1) = (1, 0, \cdots, 0)$. An interpretation of this is that the first gene explains the most variability in the samples, whereas the rest of the genes contribute nothing to the directions of the largest variability. We will use the continuum parameter, α, in a novel way, to study loadings simultaneously over $\alpha \in [0, 1]$.

We find the first CR vector for a range of $\alpha \in [0, 1]$. Each entry of the CR vectors can be seen as a contribution of the corresponding gene on the first CR vector. The rank of the effect of genes on the CR vector can be obtained by sorting the entries of the CR vector. Since we keep the signs of the entries, the genes with largest positive (top genes) and smallest negative (bottom genes) entries have the most influence on the direction vector.

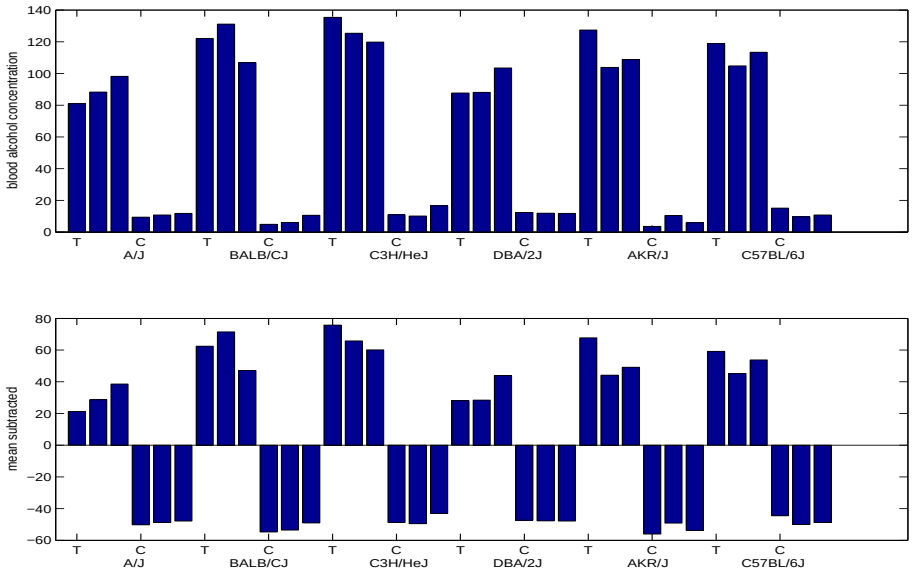

Figure 2.9: The bar graph of the output data, Blood Alcohol Concentration. Mouse samples (x-axis) are grouped as strain (A/J, ..., C57BL/6J) and within a strain, alcohol-treated(T)/ control (C) samples are grouped together. We see an obvious effect by alcohol treatment and some variability between mouse strains.

Figures 2.10 and 2.11 show the rank tracks (over the continuum parameter, α) of the 100 largest positive and negative valued genes for the OLS direction ($\alpha = 0$) over $\alpha = 0, 0.1, \cdots, 1$. Based on OLS, we select those top ranked genes (100 largest positive and 100 largest negative), i.e. genes that feel the phenotype differences across the samples. Then, we keep track of their ranks based on CR direction vectors as the continuum parameter, α marches along the range with increments .1. It is interesting to see when the top ranked genes on OLS disappear. Most genes stay influential until α reaches 0.5 (PLS). This means that a portion of genes that explains correlation between the expression and the phenotypes the most also explains the covariance. However, a large portion of genes disappear as the value of α further increases (in increments of 0.1). In fact, there is no common gene between the 100 top ranked genes for OLS and PCA, and 3 genes are common in the 100 bottom ranked genes.

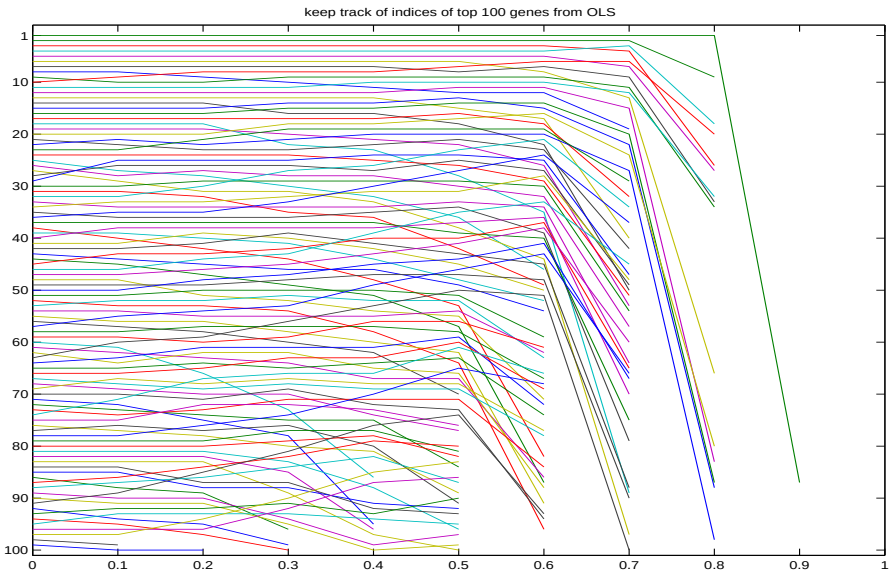

Figure 2.10: Rank tracking plot of the top 100 ranked genes from OLS ($\alpha = 0$). Most genes stay highly ranked until α goes to 0.5 (PLS). Afterwards, a portion of genes disappear (low ranked) as we get closer to PCR and none of the genes survive at the very end, $\alpha = 1$ (PCR).

PLS, a compromise between the OLS and PCA, is in fact, much closer to the OLS in this example.

2.4.3 Loading Tracking Plot

In this section, we will use the loading tracking visualization in a really different way. Instead of comparing CR direction vectors over α, we now consider different direction vectors. In addition to OLS and PLS, direction vectors considered in this section include several important discrimination methods. We view the loadings (entries) of the direction vector as gene contributions to the direction vector. The loading tracking visualization nicely conveys the changes in relative gene contribution as the direction vector changes.

The first considered vector is the second PC direction vector. This is because

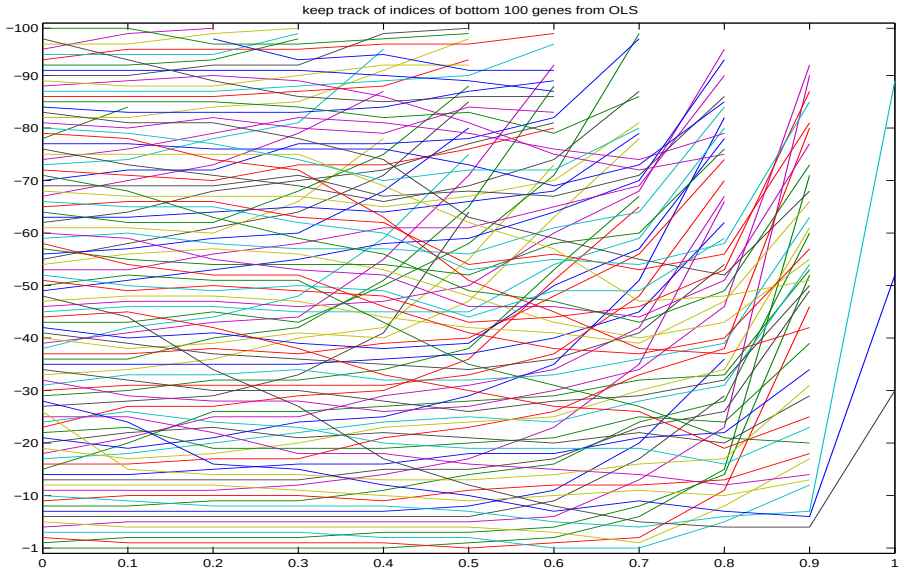

Figure 2.11: Rank tracking plot of the bottom 100 ranked genes from OLS ($\alpha = 0$). About half of genes appear to be highly ranked until α reaches 0.9 and all genes but 3 genes suddenly disappear when $\alpha = 1$. Genes responsible for the PC direction, which explains the most variability in the samples, could be very different from the genes relevant to the response variable ($\alpha = 0$).

the 1-dimensional projection on the first PC (first diagonal in Figure 2.8) shows that most of the variability comes from strain differences. On the other hand, the second most variability is mainly due to the treatment effect as seen in the second diagonal. Genes that drive the treatment effect are of more interest than genes that drive strain differences. For this reason, the second PC is considered rather than the first PC in the following analysis.

A more direct approach for identifying genes that are expressed differently between the treatment and control groups is discrimination. Discrimination methods, and also their use in the analysis of expression data, have been widely studied by many researchers. See Hastie *et al.* (2001), Duda *et al.* (2000), Tibshirani *et al.* (2003) and references therein.

Among many discrimination methods, we consider Distance Weighted Discrimi-

nation (DWD) proposed by Marron *et al.* (2007) since it performs reasonably well in HDLSS contexts. See Marron *et al.* (2007) for details. We also consider Mean Difference (MD, also called centroid by Tibshirani *et al.* (2003)), an especially simple discrimination method. MD is the normalized vector of the gene expression mean difference between the control and treatment groups.

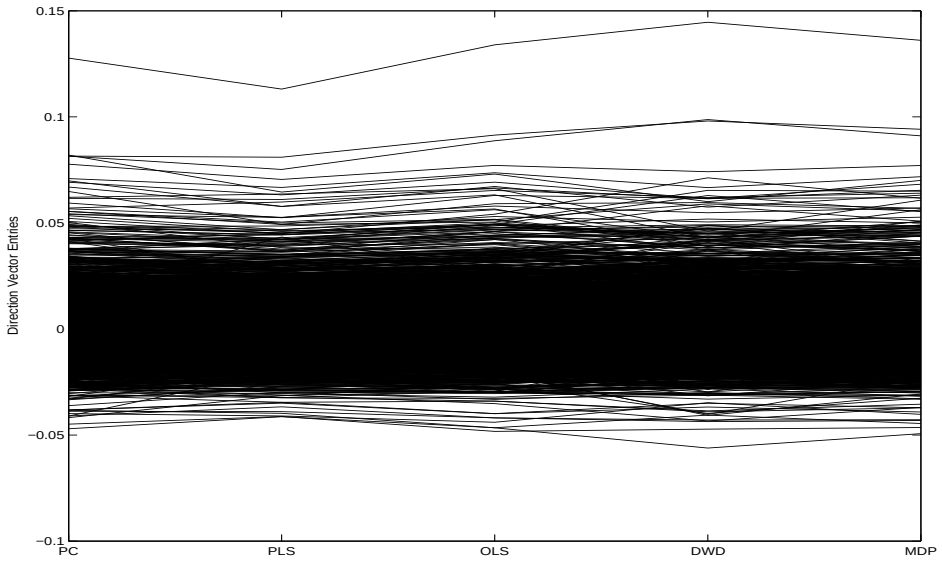

Figure 2.12: loading tracking plot for 5 direction vectors- the 2nd PC, PLS, OLS, DWD, MD. Each gene is represented by a piecewise line, connecting the entries of the respective directions. A few genes stand out as important across all the methods.

Sorting genes based on different direction vectors suggest different lists of important genes. Genes having large loadings in the second PC direction are the ones that explain the second most variability across the samples in the gene expressions (which might be related to treatment effect as explained above) . DWD and MD (right ends) are useful to identify genes that differentiate the alcohol treated sample from the controls. With OLS (the central ordinate), genes that best explain the phenotype, BAC, can be obtained.

Figure 2.12, the loading tracking plot, conveys changes of gene contribution to

these tasks. For each gene, we keep track of entries as the direction changes (from the second PC to PLS, OLS, DWD, and MD). These form a piecewise line in Figure 2.12, which shows the change of gene loadings for the five different direction vectors.

Most of the loadings are between -.05 and .05 whereas there are some genes standing out for some particular methods, or across all the analyses. The thick bundle of black curves in the middle hinders understanding as to whether they are actually parallel or crossed (up in one analysis, but down in the other, or vice versa). However, genes with low loadings (in absolute value) for all ordinates are not related to the treatment. A few genes stand out across all the methods, and these are worth looking at in detail.

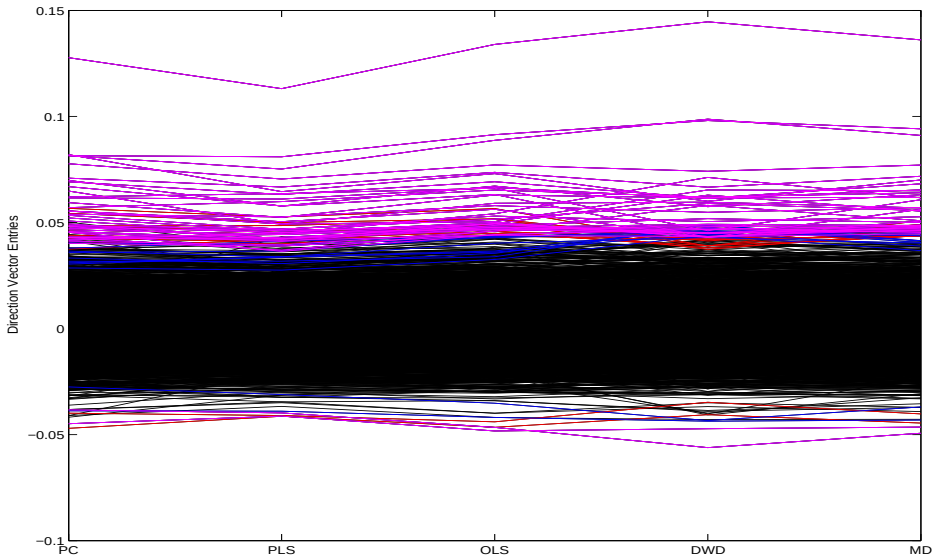

Figure 2.13: loading tracking plot, with several genes of interest highlighted. The 50 top genes based on the absolute value of OLS are colored as red, 50 based on DWD colored as blue, and purple for genes selected by both.

Let us now focus on lines that stand out either by OLS or DWD, or by both. This can be done by highlighting particular genes of interest. For example, in Figure 2.13 we highlight 50 genes based on the absolute values of the OLS directions as red,

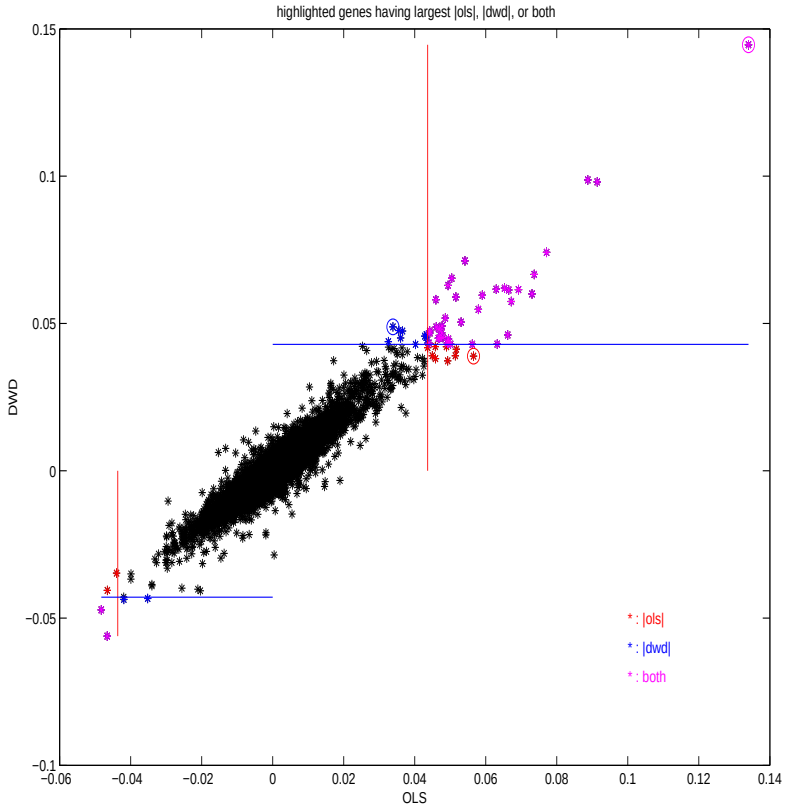

Figure 2.14: Scatter plot of OLS and DWD gene loadings. Genes are distributed along the 45°, which indicates OLS and DWD loadings strongly correlated. The 50 top genes based on the absolute values of the OLS are colored as red, 50 based on DWD colored as blue, and purple if selected by both.

50 based on DWD as blue, and purple if selected by both. While most of them lie on the top area (i.e., entries are large positive values), a few highlighted lines lie on the bottom. Out of 50 highlighted lines, 39 turn out to be purple. These 39 genes are selected both by OLS and DWD, i.e., important to explain the phenotype and also to differentiate alcohol samples from the controls. Some blue lines are partially hidden around DWD as some purple lines lie exactly on top of them. These 11 genes are in the top 50 list based on DWD, but not based on OLS. Conversely, genes that correspond to 11 red lines are important based on OLS, but not DWD.

Analyzing the gene loading changes that correspond to the black lines is challenging as the visualization is obscured by the thick bundle of lines in the middle. To focus on the comparison of OLS and DWD only, we create a scatter plot of entries for these two vectors in Figure 2.14. This scatter plot shows the loading distribution over the two direction vectors. Genes are distributed along the 45° line with some variations, i.e., some genes have larger loadings in OLS than DWD, and vise versa. However, the 45° line pattern is pretty clear, which indicates that the gene contributions to OLS and DWD are strongly correlated with each other.

Note that the highlighted lines in Figure 2.13, are now converted into highlighted crosses. The 50 genes that are most away from 0 horizontally (vertically) are colored as red (blue), respectively. The 39 genes, colored as purple, are the intersection of the two lists and located in the top-right or bottom-left corner. The 39 common genes selected by both are worth looking at in detail.

CHAPTER 3

Continuum Canonical Correlation

There are cases where two sets of multidimensional variables $X = (X_1, \cdots, X_{d_1})^T$ and $Y = (Y_1, \cdots, Y_{d_2})^T$, are observed in pairs

$$\{(\mathbf{x}_i, \mathbf{y}_i) |\ \mathbf{x}_i \in \mathbb{R}^{d_1}, \mathbf{y}_i \in \mathbb{R}^{d_2}, i = 1, \cdots, n\}$$

and the distinction between explanatory variables and response variables are not so clear. In that case, analysis dealing with two sets of variables in a symmetric manner would be more appropriate than the regression type analysis, done in this context by Stone and Brooks (1994).

Canonical Correlation Analysis (CCA), proposed by Hotelling in (Hotelling, 1936), is an example of this type of analysis. CCA is a method of finding linear relationships between two multi-variables. CCA seeks for two direction vectors, one for each variable set, such that the sample correlation between the projections of the data onto those two direction vectors are maximized.

In this chapter, we propose a generalization of CR, studied in Chapter 2, for two sets of multivariate data cases. The new method will be obtained by extending CCA into a family of CCA type analyses. Hence it will be called Continuum Canonical Correlation(CCC). CCC bears resemblance with CR in that

(i) a free parameter, $\alpha \in [0, 1]$ controls the balance between covariance and variances and

(ii) several existing methods such as Canonical Correlation Analysis (CCA), PLS (maximal covariance direction), and PCA, which we will study more in detail later in this chapter are special cases.

However, CCC differs from CR in that direction vectors in both sets of variables are of explicit interest.

In Section 3.1, we will give the description of the new method CCC. Three special cases of CCC will be reviewed in the later sections; CCA, the maximum correlation method in Section 3.3, PLS, the maximum covariance method in Section 3.4, and the PCA, the maximum variance method in Section 3.5. CCA ($\alpha = 0$) and PCA ($\alpha = 1$) occupy the ends of the spectrum of the new method and PLS ($\alpha = 1/2$) lies in the middle. These three special methods can be put in a common mathematical framework, and solved as a generalized eigenvalue problem (Borga *et al.* (1997), Shawe-Taylor and Cristianini (2004)). In Section 3.2, the generalized eigenvalue problem will be described. A numerical algorithm to find CCC direction vectors for general values of $\alpha \in [0, 1]$ is proposed in Section 3.6.

3.1 Continuum Canonical Correlation

Define the two data matrices as $\mathbf{X}_{d_1 \times n} := [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ and $\mathbf{Y}_{d_2 \times n} := [\mathbf{y}_1, \cdots, \mathbf{y}_n]$. The direction vectors, $\mathbf{u}_m \in \mathbb{R}^{d_1}$ and $\mathbf{v}_m \in \mathbb{R}^{d_2}$, are taken as the following maximizer:

$$\begin{aligned} T(\alpha) &= \max_{\mathbf{u} \in \mathbb{R}^{d_1}, \mathbf{v} \in \mathbb{R}^{d_2}} \{\mathrm{Cov}(\mathbf{u}^T X, \mathbf{v}^T Y)\}^2 \{\mathrm{Var}(\mathbf{u}^T X)\mathrm{Var}(\mathbf{v}^T Y)\}^{\alpha/(1-\alpha)-1} \\ &= \max_{\mathbf{u} \in \mathbb{R}^{d_1}, \mathbf{v} \in \mathbb{R}^{d_2}} (\mathbf{u}^T \mathbf{X}\mathbf{Y}^T \mathbf{v})^2 (\mathbf{u}^T \mathbf{X}\mathbf{X}^T \mathbf{u}\mathbf{v}^T \mathbf{Y}\mathbf{Y}^T \mathbf{v})^{\alpha/(1-\alpha)-1} \end{aligned} \tag{3.1}$$

where $0 \leqslant \alpha < 1$ subject to $\mathbf{u}_m^T \mathbf{u}_m = \mathbf{v}_m^T \mathbf{v}_m = 1$, $\mathbf{u}_m^T \mathbf{X}\mathbf{X}^T \mathbf{u}_j = 0$ and $\mathbf{v}_m^T \mathbf{Y}\mathbf{Y}^T \mathbf{v}_j = 0$ for $j = 1, \cdots, m-1$.

The parameter α controls the balance between the covariance and the variance. Continuum Canonical Correlation has the same spirit as CR in that it solves a max-

imization problem indexed by the free parameter, α, but differs in that it finds two direction vectors, one for each variable set X and Y. As for CR, CCC also embraces 3 existing methods: CCA, PLS, and PCA. These three special cases can be formulated as generalized eigen-problems, which will be introduced in the following section. In Sections 3.3 - 3.5, CCA ($\alpha = 0$), PLS ($\alpha = 1/2$), and PCA ($\alpha \to 1$) will be studied in detail.

3.2 The Generalized Eigenproblem

Following the development in Chapter 6 of Shawe-Taylor and Cristianini (2004), the generalized eigenproblem will be reviewed in this section. The generalized eigenproblem is closely related to the problem of finding the maximum point of a ratio of quadratic forms

$$r = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}}$$

where $\mathbf{A}$ and $\mathbf{B}$ are both symmetric and $\mathbf{B}$ is positive definite. This ratio is known as the *Rayleigh quotient*. Taking the derivatives with respect to $\mathbf{w}$ and setting them to zero gives the equation:

$$\frac{\partial r}{\partial \mathbf{w}} = \frac{2}{\mathbf{w}^T \mathbf{B} \mathbf{w}} (\mathbf{A}\mathbf{w} - r\mathbf{B}\mathbf{w}) = 0$$

or equivalently the generalized eigenproblem:

$$\mathbf{A}\mathbf{w} = r\mathbf{B}\mathbf{w}. \tag{3.2}$$

Since by assumption $\mathbf{B}$ is positive-definite, by pre-multiplying the equation with $\mathbf{B}^{-1}$, we can covert (3.2) to an ordinary eigenproblem

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{w} = r\mathbf{w}.$$

Let $r_1 \geqslant \cdots \geqslant r_m$ be the eigenvalues of $\mathbf{B}^{-1}\mathbf{A}$ and $\mathbf{w}_1, \cdots, \mathbf{w}_m$ be the corresponding eigenvectors.

It can be shown that (Shawe-Taylor and Cristianini (2004), Borga *et al.* (1997)) the eigenvector corresponding to the maximum (minimum) eigenvalue, $\mathbf{w}_1(\mathbf{w}_m)$, is the global maximum (minimum) point of the *Rayleigh quotient* and the remaining eigenvectors are saddle points. Therefore, searching for the maximum point of the *Rayleigh quotient* (3.2) is equivalent to finding the eigenvector of the generalized eigen-problem (3.2) corresponding to the largest eigenvalue. It is useful to view the generalized eigenproblem from a sequential viewpoint. After the eigenvector corresponding to the largest eigenvalue, $\mathbf{w}_1$, is computed, the eigenvector corresponding to the second largest eigenvalue, $\mathbf{w}_2$, solves the following maximization problem under an orthogonality condition:

$$\mathbf{w}_2 = \text{argmax}_{\mathbf{w}\,:\,\mathbf{w}^T\mathbf{B}\mathbf{w}_1=0} \frac{\mathbf{w}^T\mathbf{A}\mathbf{w}}{\mathbf{w}^T\mathbf{B}\mathbf{w}}.$$

A similar sequential representation holds for the remaining eigenvectors. Thus, the solution of the generalized eigenproblem gives a sequence of direction vectors to maximize the *Rayleigh quotient* in (3.2) under the $\mathbf{B}$-orthogonality condition.

In the following sections, we will see that CCA, PLS, and PCA can be formulated as a generalized eigenproblem, with a special choice of the matrices $\mathbf{A}$ and $\mathbf{B}$.

3.3 CCA: Direction of maximum Correlation

CCA (Hotelling, 1936) is one of the principal tools in multivariate statistics for studying the relationship between two paired sets of multivariate data. The goal of CCA is to find two direction vectors, one for each variable set, such that the correlation between the projections of variables onto these vectors are maximized.

The empirical correlation between $Z := \mathbf{u}^T X$ and $W := \mathbf{v}^T Y$ can be written as

$$\begin{aligned} \rho &\doteq \frac{\mathrm{Cov}(Z, W)}{\sqrt{\mathrm{Var}(Z)}\sqrt{\mathrm{Var}(W)}} \\ &= \frac{\mathbf{u}^T \mathbf{X}\mathbf{Y}^T \mathbf{v}}{\sqrt{\mathbf{u}^T \mathbf{X}\mathbf{X}^T \mathbf{u}}\sqrt{\mathbf{v}^T \mathbf{Y}\mathbf{Y}^T \mathbf{v}}}. \end{aligned} \tag{3.3}$$

Note that this corresponds to the CCC when $\alpha = 0$. Since the correlation is scale invariant, the vectors $\mathbf{u}$ and $\mathbf{w}$ are determined up to direction. The requirement $\mathbf{u}^T \mathbf{X}\mathbf{X}^T \mathbf{u} = \mathbf{v}^T \mathbf{Y}\mathbf{Y}^T \mathbf{v} = 1$ can resolve the ambiguity of scale issue. Then, CCA is equivalent to the following:

$$\begin{gathered} \max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{X}\mathbf{Y}^T \mathbf{v} \\ \text{subject to } \mathbf{u}^T \mathbf{X}\mathbf{X}^T \mathbf{u} = \mathbf{v}^T \mathbf{Y}\mathbf{Y}^T \mathbf{v} = 1. \end{gathered}$$

The corresponding Lagrangian is

$$\mathbf{u}^T \mathbf{X}\mathbf{Y}^T \mathbf{v} - \frac{\lambda_x}{2}(\mathbf{u}^T \mathbf{X}\mathbf{X}^T \mathbf{u} - 1) - \frac{\lambda_y}{2}(\mathbf{v}^T \mathbf{Y}\mathbf{Y}^T \mathbf{v} - 1).$$

Taking the partial derivatives with respect to $\mathbf{u}$ and $\mathbf{v}$, and setting the derivatives to zero give the equations,

$$\begin{cases} \mathbf{X}\mathbf{Y}^T \mathbf{v} = \lambda_x \mathbf{X}\mathbf{X}^T \mathbf{u} \\ \mathbf{Y}\mathbf{X}^T \mathbf{u} = \lambda_y \mathbf{Y}\mathbf{Y}^T \mathbf{v}. \end{cases} \tag{3.4}$$

Subtracting $\mathbf{u}^T$ times the first from $\mathbf{v}^T$ times the second, we obtain $\lambda_x \mathbf{u}^T \mathbf{X}\mathbf{X}^T \mathbf{u} - \lambda_y \mathbf{v}^T \mathbf{X}\mathbf{X}^T \mathbf{v} = 0$, which implies $\lambda_x = \lambda_y$. Denoting this value by λ and letting

$$\mathbf{A} = \begin{pmatrix} 0 & \mathbf{X}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{X}^T & 0 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{X}\mathbf{X}^T & 0 \\ 0 & \mathbf{Y}\mathbf{Y}^T \end{pmatrix}, \text{ and } \mathbf{w} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}, \tag{3.5}$$

then the equation (3.4) can be reduced to solving a generalized eigenvalue problem, $\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w}$. As a result, the eigenvector $\mathbf{w}$ corresponding to the largest eigenvalue

will give the canonical direction vectors, $\mathbf{u}_1$ and $\mathbf{v}_1$. The maximum of the correlation ρ over $\mathbf{u}$ and $\mathbf{v}$ is called the *canonical correlation*, and the linearly transformed variables $Z = \mathbf{u}_1^T X$ and $W = \mathbf{v}_1^T Y$ are called *canonical variates*.

3.3.1 Example

In the movie, (Lee, 2005a), the canonical directions are studied for a $d_1 = d_2 = 2$ example, using simulated data. Figure 3.1 is a snap shot from the movie. The top panel on the left shows the X data as black stars, and the projections of data, onto one arbitrarily chosen direction vector $\mathbf{u}$, as red circles. The bottom panel is for the Y data and another arbitrary $\mathbf{v}$. The panel on the right shows the scatter plot of the paired projections, onto the 2-d subspace of $\mathbb{R}^4$ generated by the direction vectors $\mathbf{u}$ and $\mathbf{v}$, which shows a weak correlation for this choice of $\mathbf{u}$ and $\mathbf{v}$.

In the first part of the movie, the direction vector $\mathbf{u}$ in the X space is fixed as shown in Figure 3.1, and the direction vector $\mathbf{v}$ in the Y space is rotated. As the direction rotates, the joint distribution of the projections in the right panel changes. Rotation stops at the direction $\mathbf{v}_1$ which gives the maximum correlation. In the second part of the movie, $\mathbf{v}_1$ in the Y space is fixed, and direction vector in the X space is rotated. Similarly, rotation stops at the vector $\mathbf{u}_1$ to maximize the correlation of the paired projections.

Figure 3.2 is the last snap shot from this movie. The panel on the right shows the scatter plot of data projections onto CCA direction vectors, which clearly shows a strong linear correlation. Note that the CCA vectors, $\mathbf{u}_1$ and $\mathbf{v}_1$, indicated by blue lines on the left, are nearly orthogonal to the PCA vectors, in each of X and Y spaces, the directions along which data points are spread the most, highlighting the important point that these can be quite different.

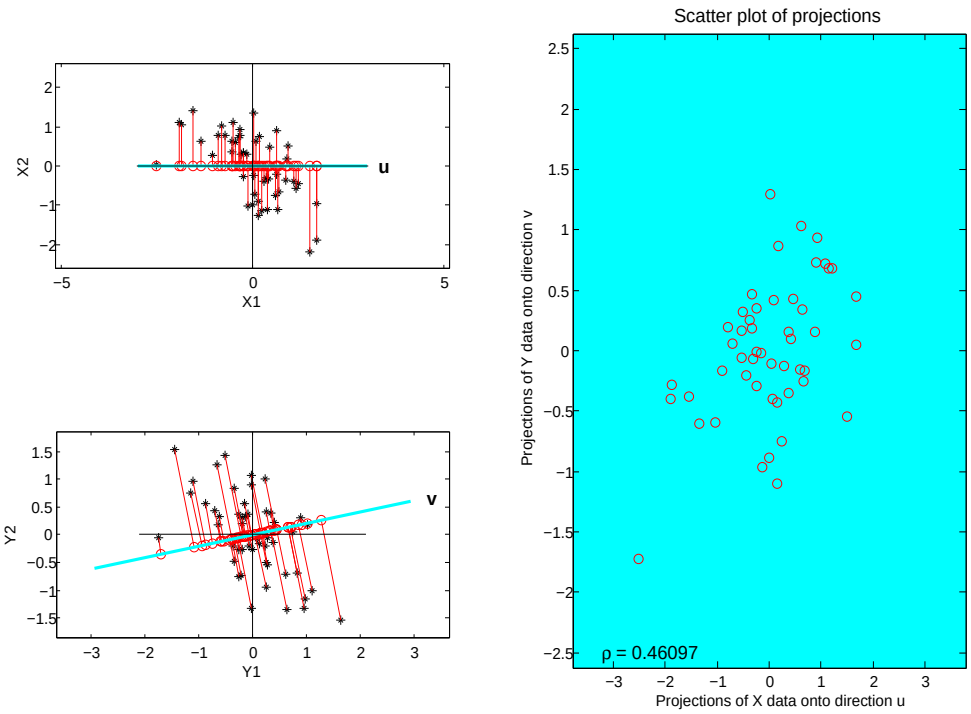

Figure 3.1: Two paired sets of 2-d data vectors and the projections of the two direction vectors are shown in the left panels. The scatter plot of data projections are seen in the right panel. This show a weak correlation between the paired projections.

3.3.2 HDLSS CCA

For HDLSS data in the sense that either $d_1 > n$ or $d_2 > n$ holds, or both hold, the matrix $\mathbf{B}$ in the equation (3.5) becomes singular. An ordinary approach to solving the eigenvalue problem by pre-multiplying the inverse $\mathbf{B}$ on both sides of the equation can not be applied. For the sake of simplicity, assume for now that only $d_1 > n$ is true. Then, the rank of the covariance matrix $\mathbf{X}\mathbf{X}^T$ is less than its dimension d_1. In particular, this means that null space $\{\mathbf{u} \mid \mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u} = 0\}$ is not empty. For any direction vectors in this null space, the denominator of the correlation in (3.3)

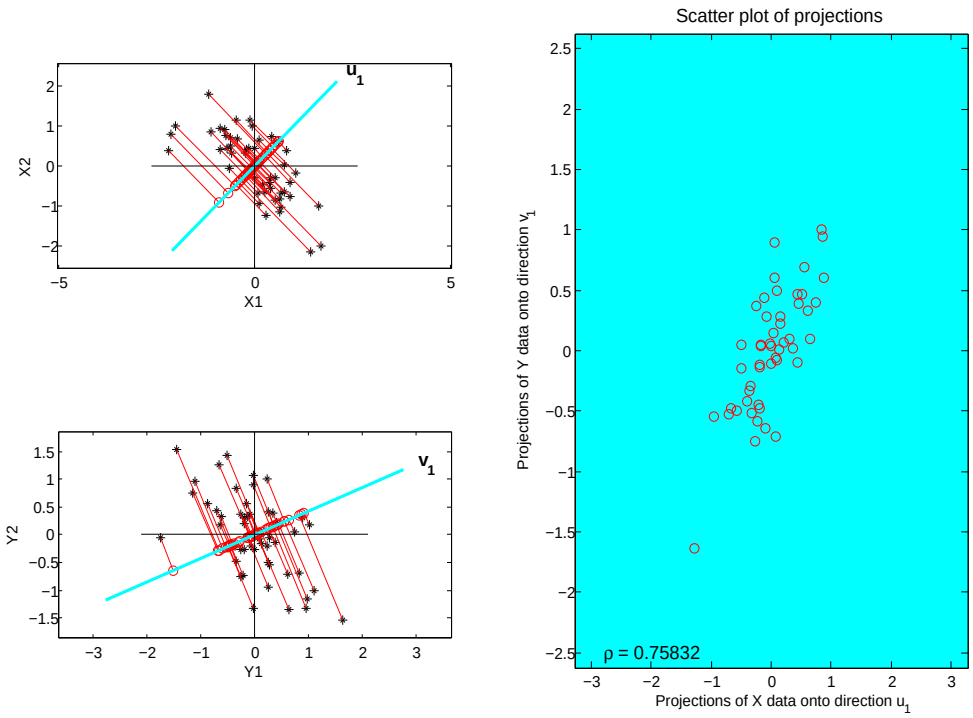

Figure 3.2: Two paired sets of 2-d data vectors and the projections of the CCA direction vectors are shown in the left panels. The scatter plot of data projections onto CCA vectors are seen in the right panel, which clearly shows a strong correlation.

is 0, thus the ratio is undefined. Therefore, for a singular covariance matrix $\mathbf{X}\mathbf{X}^T$, maximizing the correlation is not well- defined. This is true in all of the cases $d_2 > n$, or both $d_1 > n$ and $d_2 > n$ are true.

To circumvent the singularity problem, confine the direction vectors to be in the subspace generated by the data, i.e., $\mathbf{u} = \mathbf{X}\alpha$ and $\mathbf{v} = \mathbf{Y}\beta$ for some $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^n$. Substituting into the equation (3.3) we obtain the following

$$\rho = \frac{\alpha^T \mathbf{X}^T \mathbf{X}\mathbf{Y}^T \mathbf{Y}\beta}{\sqrt{\alpha^T \mathbf{X}^T \mathbf{X}\mathbf{X}^T \mathbf{X}\alpha}\sqrt{\beta^T \mathbf{Y}^T \mathbf{Y}\mathbf{Y}^T \mathbf{Y}\beta}}.$$

Assume Rank($\mathbf{X}^T\mathbf{X}$) = n and Rank($\mathbf{Y}^T\mathbf{Y}$) = n. Then, the maximization problem over α and β is now well defined. As for the ordinary CCA, it can be formulated as a generalized eigenvalue problem.

3.4 PLS: Direction of maximum Covariance

PLS is a widely used method in chemometrics when tackling a regression problem. The PLS begins with canonical covariance analysis; maximize the covariance between projections of two paired sets of data onto two directions specified by $\mathbf{u}$ and $\mathbf{v}$. There are several versions of PLS for the multiple response case; they differ in the way that orthogonality constraint for subsequent direction vectors is imposed (Phatak and Jong, 1997). In this work we stick to a particular version of PLS introduced in Borga *et al.* (1997).

The set of vectors are obtained in pairs and we are interested in studying the covariance between the two parts of the paired data set. This is in contrast to the CCA which normalizes with respect to the variances of two linear combinations, thus studying correlation. The directions $\mathbf{u}$ and $\mathbf{v}$ of maximum covariance can be formulated as follows:

$$\begin{gathered}\max_{\mathbf{u},\mathbf{v}} \ \mathrm{Cov}(\mathbf{u}^T X, \mathbf{v}^T Y) = \max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{X}\mathbf{Y}^T \mathbf{v} \\ \text{subject to } \mathbf{u}^T\mathbf{u} = 1 \text{ and } \mathbf{v}^T\mathbf{v} = 1.\end{gathered} \tag{3.6}$$

The objective function is the same as that of CCC for $\alpha = 1/2$. Applying the Lagrange multiplier technique to the maximization problem (3.6) gives the optimization problem

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{X}\mathbf{Y}^T \mathbf{v} - \frac{\lambda_x}{2}(\mathbf{u}^T\mathbf{u} - 1) - \frac{\lambda_y}{2}(\mathbf{v}^T\mathbf{v} - 1).$$

Taking derivatives with respect to $\mathbf{u}$ and $\mathbf{v}$ and setting them to zeroes, we have the

equations

$$\begin{cases} \mathbf{X}\mathbf{Y}^T\mathbf{v} - \lambda_x\mathbf{u} = 0 \\ \mathbf{Y}\mathbf{X}^T\mathbf{u} - \lambda_y\mathbf{v} = 0. \end{cases} \tag{3.7}$$

Subtracting $\mathbf{u}^T$ times the first from $\mathbf{v}^T$ times the second, we obtain $\lambda_x\mathbf{u}^T\mathbf{u} - \lambda_y\mathbf{v}^T\mathbf{v} = 0$, which implies $\lambda_x = \lambda_y$. Denoting this value by λ and letting

$$\mathbf{A} = \begin{pmatrix} 0 & \mathbf{X}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{X}^T & 0 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{pmatrix}, \text{ and } \mathbf{w} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix},$$

the equation (3.7) can be written as a generalized eigen-problem, $\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}$.

Another way to give the solutions to the maximum covariance problem (3.6) (Shawe-Taylor and Cristianini, 2004) is the Singular Value Decomposition (SVD) of matrix, $\mathbf{X}\mathbf{Y}^T$, the sample covariance matrix between two sets of variables. Namely, if the SVD of $\mathbf{X}\mathbf{Y}^T$ is given as

$$\mathbf{X}\mathbf{Y}^T = s_1\mathbf{u}_1\mathbf{v}_1^T + \cdots + s_m\mathbf{u}_m\mathbf{v}_m^T \tag{3.8}$$

where $s_1 \geqslant \cdots \geqslant s_m > 0$ are singular values, and $\mathbf{u}_i$ and $\mathbf{v}_i$ are corresponding singular vectors and rows of $\mathbf{X}\mathbf{Y}^T$, then the first singular vector and row, $\mathbf{u}_1$ and $\mathbf{v}_1$ indeed are the solution to the problem (3.6). Equation (3.8) will be fundamental in Chapter 5.

3.5 PCA: Direction of Maximum Variance

The most distinctive feature of PCA from PLS or CCA is that PCA focuses on searching for the direction vector which can explain the most variability of data in each space separately, which thus does not take into account any relationship between two sets of vectors.

Finding direction vectors $\mathbf{u}$ and $\mathbf{v}$ such that the linear combination $\mathbf{u}^T X$ and $\mathbf{v}^T Y$

have the maximum variation can be written as the following optimization problem:

$$\max_{\mathbf{u},\mathbf{v}} \mathrm{Var}(\mathbf{u}^T X)\mathrm{Var}(\mathbf{v}^T Y) = \max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u}\ \mathbf{v}^T\mathbf{Y}\mathbf{Y}^T\mathbf{v}$$

subject to $\mathbf{u}^T\mathbf{u} = \mathbf{v}^T\mathbf{v} = 1$ or equivalently:

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathrm{Var}(\mathbf{u}^T X)}{\mathbf{u}^T\mathbf{u}} \frac{\mathrm{Var}(\mathbf{v}^T Y)}{\mathbf{v}^T\mathbf{v}} = \max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u}}{\mathbf{u}^T\mathbf{u}} \frac{\mathbf{v}^T\mathbf{Y}\mathbf{Y}^T\mathbf{v}}{\mathbf{v}^T\mathbf{v}}. \tag{3.9}$$

Looking back to the criteria of CCC in (3.1), as the parameter α tends to 1, we essentially search for vectors $\mathbf{u}$ and $\mathbf{v}$ such that the variance of the projections of the data onto those directions are maximized. Thus, PCA can be regarded as a limiting case of CCC as α tends to 1.

Letting

$$\mathbf{A} = \begin{pmatrix} \mathbf{X}\mathbf{X}^T & 0 \\ 0 & \mathbf{Y}\mathbf{Y}^T \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{pmatrix}, \text{ and } \mathbf{w} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix},$$

the ratio (3.9) is of the form

$$\frac{\mathbf{w}^T\mathbf{A}\mathbf{w}}{\mathbf{w}^T\mathbf{B}\mathbf{w}}.$$

As explained in Section 3.2, maximization of the *Rayleigh Quotient* can be formulated as a generalized eigen-problem.

Example. With the same simulated data as in Figure 3.1 and 3.2 for the CCA illustration in Section 3.3.1, we find the PCA directions and compare with the CCA direction vectors. In Figure 3.3, data vectors for the X variables are seen in the top, Y in the bottom. The CCA direction vectors and the data projections onto them are shown in the left, PCA directions and projections are presented in the right. One can easily see that the CCA and the PCA direction vectors in each set of variables are nearly orthogonal to each other. This suggests that "interesting" vectors could

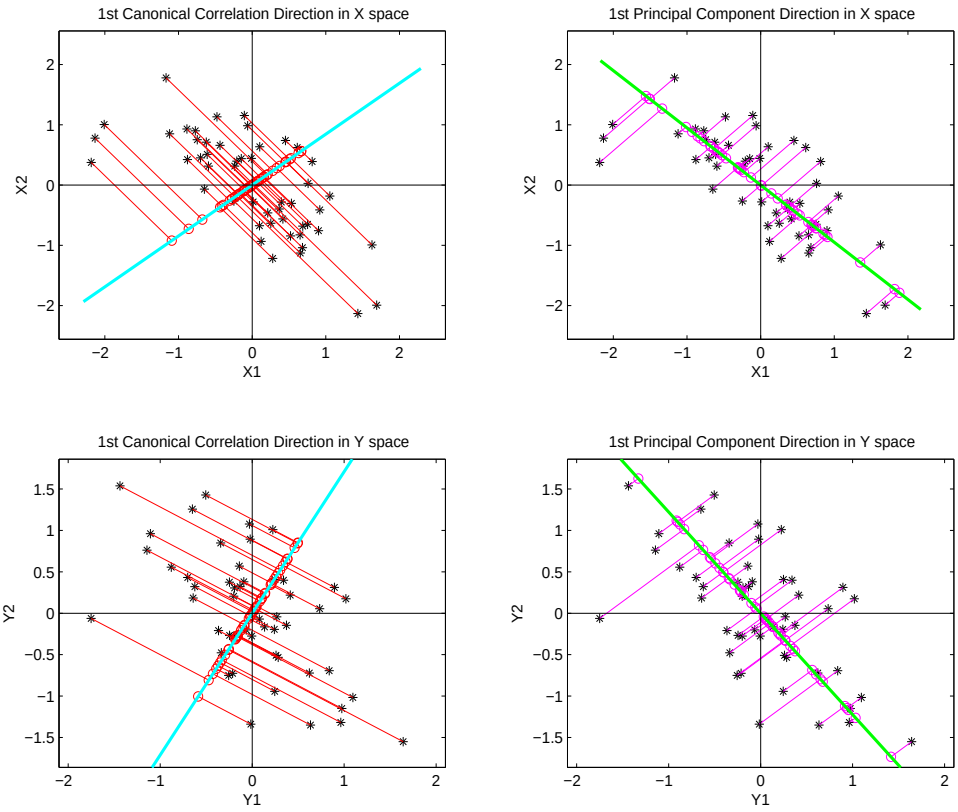

Figure 3.3: For the same two sets of 2-d data vectors as in Figure 3.1 and 3.2, shown in the left are the CCA direction vectors, and in the right PCA direction vectors. The two vectors are almost perpendicular, in both the X and Y spaces, showing CCA can be very different from PCA.

be very different depending on the task at hand.

3.6 Algorithm

In this section, a numerical algorithm for solving optimization for CCC is proposed. Note that for the three special cases of CCA, PLS, and PCA, the objective functions in (3.1) are in quadratic form, in which case the maximizer can be obtained by solving a generalized eigen-problem. However, this is no longer the case for the other values of α, so a more complicated numerical approach is needed.

Instead of maximizing the objective function (3.1) with respect to $\mathbf{u}$ and $\mathbf{v}$ simul-

taneously, we fix one of the vector (i.e., either $\mathbf{u}$ or $\mathbf{v}$), then maximize the function with respect to the other vector. We will perform this optimization iteratively until the solutions converge. Each iterative maximization step is exactly the CR problem, which we can solve numerically using the algorithm by Stone and Brooks (1990).

Suppose that we have already found $m-1$ pairs of CCC vectors, i.e., $\mathbf{u}_1, \cdots, \mathbf{u}_{m-1}, \mathbf{v}_1 \cdots,$ and $\mathbf{v}_{m-1}$ are available. The strategy for numerically solving for the $m-$th CCC vectors, $\mathbf{u}_m$ and $\mathbf{v}_m$, is the following:

(i) For the initial step, $j = 0$, set

$$\begin{aligned} \mathbf{u}_m^{(j)} &= (1-\alpha)\mathbf{u}_{CCA,m} + \alpha \mathbf{u}_{PCA,m} / ||(1-\alpha)\mathbf{u}_{CCA,m} + \alpha \mathbf{u}_{PCA,m}|| \\ \mathbf{v}_m^{(j)} &= (1-\alpha)\mathbf{v}_{CCA,m} + \alpha \mathbf{v}_{PCA,m} / ||(1-\alpha)\mathbf{v}_{CCA,m} + \alpha \mathbf{v}_{PCA,m}||, \end{aligned}$$

where $\mathbf{u}_{CCA,m}$, $\mathbf{v}_{CCA,m}$, $\mathbf{u}_{PCA,m}$ and $\mathbf{v}_{PCA,m}$ are the m-th CCA and PCA direction vectors, respectively.

(ii) For $j = 0, 1, \cdots,$ fix $\mathbf{v}_m^{(j)}$ and let $\mathbf{y} = \mathbf{Y}^T \mathbf{v}_m^{(j)}$. Use the algorithm of CR to maximize the objective function (3.1) with respect to $\mathbf{u}$, i.e.,

$$\mathbf{u}^* = \text{argmax}_{\mathbf{u} \in \mathbb{R}^{d_1}} (\mathbf{u}^T \mathbf{X} \mathbf{y})^2 (\mathbf{u}^T \mathbf{X} \mathbf{X}^T \mathbf{u})^{\alpha/(1-\alpha)-1}$$

subject to $\mathbf{u}^T \mathbf{u} = 1$ and $\mathbf{u}_m^T \mathbf{X} \mathbf{X}^T \mathbf{u}_i = 0$ for $i = 1, \cdots, m-1$. Let $\mathbf{u}_m^{(j+1)} = \mathbf{u}^*$.

(iii) For $j = 1, 2, \cdots,$ fix $\mathbf{u}_m^{(j)}$ and let $\mathbf{u} = \mathbf{X}^T \mathbf{u}_m^{(j)}$. Again, use the algorithm of CR to maximize the objective function (3.1) with respect to $\mathbf{v}$, i.e.,

$$\mathbf{v}^* = \text{argmax}_{\mathbf{v} \in \mathbb{R}^{d_1}} (\mathbf{v}^T \mathbf{Y} \mathbf{x})^2 (\mathbf{v}^T \mathbf{Y} \mathbf{Y}^T \mathbf{v})^{\alpha/(1-\alpha)-1}$$

subject to $\mathbf{v}^T \mathbf{v} = 1$ and $\mathbf{v}_m^T \mathbf{Y} \mathbf{Y}^T \mathbf{v}_i = 0$ for $i = 1, \cdots, m-1$. Let $\mathbf{v}_m^{(j+1)} = \mathbf{v}^*$.

Repeat (ii) and (iii) until the solutions converge.

Solving the d_1 or d_2-dimensional problem in each iterative step, (ii) and (iii), and continually updating solutions may appear to be a daunting task, especially when d_1 or d_2 is large. However, the numerical algorithm for solving CR, described by Stone and Brooks (1990), is actually performing an n-dimensional optimization problem for HDLSS cases ($n < d_1, d_2$), which makes the algorithm more tractable.

A useful property of the algorithm above is that each iteration does not decrease the objective function. In particular, suppose that we just finished step (ii) at the j-th iteration. Then, the target function (3.1) evaluated at $\mathbf{u} = \mathbf{u}_m^{(j+1)}$, $\mathbf{v} = \mathbf{v}_m^{(j)}$ is always greater than the value evaluated at $\mathbf{u} = \mathbf{u}_m^{(j)}$, $\mathbf{v} = \mathbf{v}_m^{(j)}$. This is true for all iteration steps corresponding to (iii). For a given pair of data sets, $\mathbf{X}$ and $\mathbf{Y}$, the target function (3.1) is bounded. Thus, the sequence of function values evaluated at iterative solutions will converge to a local maximum value.

However, there is no guarantee that the sequence of iterative solutions converges to the global maximum. Depending on the starting point, the solution can converge to a local maximum. Improvement for escaping from a local maximum can be made by setting a better starting point (i.e., a initial point closer to the global maximum). One approach toward this goal is to make use of the fact that we can obtain CCC solution for $\alpha = .5$. For given α, say .4, the solution is expected to be closer to the solution for $\alpha = .5$ than a convex combination of the two known solutions for the extreme cases of $\alpha = 0$ and 1, as the suggested above in step (i). For this reason, any given α, we will use the solutions for $\alpha = 0$, .5, or 1, as a starting point, depending on whichever α is closest to among 0, .5, or 1, respectively. Thus, the step (i) will be replaced by the following.

(i') For the initial step, $j = 0$, set

$$\mathbf{u}_m^{(j)} = \begin{cases} \mathbf{u}_{CCA,m} & \text{if } 0 \leqslant \alpha < .25, \\ \mathbf{u}_{PLS,m} & \text{if } .25 \leqslant \alpha \leqslant .75, \\ \mathbf{u}_{PCA,m} & \text{if } .25 < \alpha \leqslant 1, \end{cases}$$

and

$$\mathbf{v}_m^{(j)} = \begin{cases} \mathbf{v}_{CCA,m} & \text{if } 0 \leqslant \alpha < .25, \\ \mathbf{v}_{PLS,m} & \text{if } .25 \leqslant \alpha \leqslant .75, \\ \mathbf{v}_{PCA,m} & \text{if } .75 < \alpha \leqslant 1. \end{cases}$$

3.7 Future Work

In this section, a potential usage of CCC is suggested. Data that we have in mind for the application are 1) gene expressions and 2) metabolomics data. They are obtained in pairs and both are HDLSS data. We can compute CCC vectors for a range of α values from 0 to 1, for example, with an increment of .1. The product of the computation is a number of pairs of direction vectors, one in gene expression space and the other in the metabolomics space. As α changes, we can analyze how the genes and metabolomics loadings change as seen in Sections 2.4.2 and 2.4.3. Loading tracking plots can give us insight as to how the important genes and the important metabolites selected by different directions (tasks) are relate to each other. Genes and metabolites that are important for explaining the correlation across the two data sets (i.e., a "joint" structure of the two data sets) will be important for α close to 0. As α tend to 1, CCC direction vectors eventually focus on the variation of the two data sets separately (i.e., "marginal" structure of the two data sets). Thus, two sets of CCC direction vector entries from $\alpha = 0$ to $\alpha = 1$, one for genes and the other for metabolites, can be interpreted as a span of loadings changes from a "joint" data analysis to a "marginal" data analysis. They can also help us select important genes

and metabolites across the tasks, as done in Sections 2.4.2 and 2.4.3.

CHAPTER 4

HDLSS Asymptotics

4.1 Introduction

As there are more HDLSS data emerging in various fields such as micro-array experiments, signal processing, and image analysis, there is a strong need to develop multivariate analysis tools which are designed to work well for this data type. The HDLSS data type motivates a new approach to mathematical statistics. In particular, there has been an increasing interest in a family of asymptotics, with the dimension, d, increasing.

Among many subjects in the multivariate asymptotics literature, there is a long history of the study on the eigenvalues and eigenvectors of the sample covariance matrices, i.e., study on PCA (Anderson (1963), Muirhead (1982)). Classical asymptotics deal with the case of increasing sample size but fixed dimensionality. In that scenario, most of the studies make use of the fact that the sample covariance matrix is a good approximation of the population covariance. However, this is no longer the case with increasing dimensionality, d.

In an increasing d scenario, there are two types of circumstances commonly considered. One is to let dimensionality and sample size grow together. Bai and Yin (1993), Paul (2004), and Johnstone and Lu (2004) have studied asymptotics where d and n grow with the same rate, in the sense that the sample size to dimension ratio converges to a positive constant γ. Some researchers have addressed the case where

d grows with some power of n, for example, Portnoy (1984) and Portnoy (1988) let $n \to \infty$, with d also growing as $n^{1/2}$. We refer to these types of studies as High Dimensional High Sample Size (HDHSS)- asymptotics.

The other extreme case, called HDLSS-asymptotics, emerges rather recently. In Hall *et al.* (2005), the geometric structure of HDLSS data was explored. They let d go infinity, while keeping the sample size n relatively small. This type of asymptotic result is more relevant to the analysis of HDLSS data. In the fixed n and increasing d setting, Ahn *et al.* (2007) found conditions under which the first eigenvector of the sample covariance matrix is consistent to its theoretical counterpart. They assume that the first eigenvalue of the population covariance matrix is extremely large compared to the rest of them. They also found some interesting inconsistency conditions. Namely, if the population covariance matrix is not extremely aspherical, the samples eigenvalues tend to behave as if they are from the spherical Gaussian distribution.

There is an important contrast between the aforementioned work and this dissertation work. Most of the former has a focus on the analysis of variance of the variables. Thus, eigen-analysis of the covariance matrix, (i.e., PCA), becomes an important tool. However, in the latter case, we focus our attention on the covariance analysis of two sets of variables. Thus, in this case, the SVD of the sample cross-covariance matrix plays a central role. Note that the cross-covariance matrix is not required to be square.

In the next section, some HDHSS and HDLSS asymptotic results, mostly regarding the eigenvalues and eigenvectors of the sample covariance matrix, will be reviewed. In the following two sections, HDLSS asymptotic conditions under which there is consistency and strong inconsistency of the singular vectors of the sample cross-covariance matrix will be studied, respectively.

4.2 Asymptotics of Sample Covariance Matrices

Consider a $d \times n$ data matrix $\mathbf{X} := [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ where $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are *i.i.d.* with mean zero and covariance Σ. Define the sample covariance matrix as

$$\widehat{\Sigma} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$$

and let $\widehat{\lambda}_1 \geqslant \cdots \geqslant \widehat{\lambda}_r > 0$ be the eigenvalues of $\widehat{\Sigma}$, where $r = \text{rank}(\widehat{\Sigma})$. Note that the sample mean is not subtracted as this form is commonly considered in the study of large dimensional random matrices.

In Sections 4.2.1 and 4.2.2, some HDHSS and HDLSS asymptotic results, mainly about eigenvalues and eigenvectors of $\widehat{\Sigma}$, will be reviewed.

4.2.1 HDHSS Asymptotics

In this section, we will consider the case where both the sample size and the dimension grow in a comparable manner in the sense that $\frac{d}{n} \to \gamma \in (0, \infty)$ as $n \to \infty$. In this case, data matrices will be viewed as a double array indexed by both d and n.

Spherical Distribution

Throughout this subsection, we assume the spherical population, i.e., the population covariance Σ is assumed to be identity. Define the empirical distribution of eigenvalues, often called the *Empirical Spectral Distribution (ESD)*, as

$$F(x) = \frac{1}{d} \times \{\text{number of } \widehat{\lambda}_i s \leqslant x\}.$$

Then, the ESD converges almost surely to the Marčenko-Pastur distribution, F (Marčenko and Pastur, 1967), with the density function defined as

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

where $a = (1-\sqrt{\gamma})^2$ and $b = (1+\sqrt{\gamma})^2$. Significant expansion and refinement of the theorem has been made under various assumptions by Bai and Yin (1988) and Yin (1986). For details, see the survey paper by Bai (1999), which provides extensive reviews on the spectral analysis results.

While the result above focuses on the bulk of sample eigenvalues, the extremes around the edge of the support, F, such as the largest (or the first few largest) and the smallest eigenvalues, have drawn the attention of many researchers.

Studies on the largest sample eigenvalues include Geman (1980), Yin *et al.* (1988), Silverstein (1989), and Johnstone (2001). Geman (1980) established the almost sure limit of the largest sample eigenvalue, $\widehat{\lambda}_1$:

$$\widehat{\lambda}_1 \overset{a.s.}{\to} (1+\sqrt{\gamma})^2, \tag{4.1}$$

assuming some additional conditions on the population moments. The other extreme, the smallest eigenvalue, has been studied by Bai and Yin (1993) and Silverstein (1985). They showed that under the spherical population model (not necessarily Gaussian) with finite fourth moment,

$$\widehat{\lambda}_r \overset{a.s.}{\to} (1-\sqrt{\gamma})^2.$$

The limiting distribution of $\widehat{\lambda}_1$, which thus provides information on the variability of the largest sample eigenvalue, is established by Johnstone (2001). Under a spherical

Gaussian assumption, if centered by

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

and scaled by

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

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

Spiked Distribution

For some real data, however, the spherical population seems unrealistic to assume. Among many attempts to study non-spherical population models, the so called *spiked population model* named by Johnstone (2001) is of particular interest. Examples in speech recognition (Hastie *et al.* (1995), Johnstone (2001)), financial mathematics (Laloux *et al.*, 2000), and statistical learning (Hoyle and Rattray, 2004) indicate that a few sample eigenvalues distinguish from the rest. The spiked population model assumes that all eigenvalues are one except that a finite number of eigenvalues that are bigger than one, i.e., $\Sigma_{d\times d} = \text{diag}(\lambda_1, \cdots, \lambda_M, 1, \cdots, 1)$ where $\lambda_1 > \lambda_2 > \cdots > \lambda_M > 1$.

The almost sure limit of the first few largest sample eigenvalues were established by Paul (2004) and Baik and Silverstein (2006). The former focuses on real Gaussian samples and the latter includes complex non-Gaussian cases. The overlapping result of the two papers is the following: as $n \to \infty$ with $\frac{d}{n} \to \gamma \in (0, 1)$

- if $\lambda_i \leqslant 1 + \sqrt{\gamma}$, then

$$\widehat{\lambda}_i \overset{a.s.}{\to} (1 + \sqrt{\gamma})^2 \tag{4.2}$$

- if $\lambda_i > 1 + \sqrt{\gamma}$, then

$$\widehat{\lambda}_i \overset{a.s.}{\to} \lambda_i \left(1 + \frac{\gamma}{\lambda_i - 1}\right)$$

Note that the limit in (4.2) is the same as the limit for the spherical in (4.1). In other words, when the population eigenvalue is not much different from one, then the corresponding sample eigenvalues behave as if the true population were spherical. This is a crucial observation made in both works, which is termed as *phase transition phenomenon*. Similar phenomenon is also observed in the HDLSS context (Ahn *et al.*, 2007).

The phase transition is also observed in the eigenvector analysis. Paul (2004) studied the limiting behavior of the angle between the true eigenvector and the sample eigenvector. Let $\mathbf{v}_1, \cdots \mathbf{v}_d$ be the eigenvectors of Σ, where the corresponding eigenvalues are sorted in a decreasing order. Define the sample eigenvectors, $\widehat{\mathbf{v}}_1, \cdots \widehat{\mathbf{v}}_d$ from $\widehat{\Sigma}$, similarly. Then,

- if $\lambda_i > 1 + \sqrt{\gamma}$ and of multiplicity one,

$$< \widehat{\mathbf{v}}_i, \mathbf{v}_i > \overset{a.s.}{\to} \sqrt{(1 - \frac{\gamma}{(\lambda_i - 1)^2})/(1 + \frac{\gamma}{\lambda_i - 1})} \quad \text{as } n \to \infty$$

- if $\lambda_i \leqslant 1 + \sqrt{\gamma}$,

$$< \widehat{\mathbf{v}}_i, \mathbf{v}_i > \overset{a.s.}{\to} 0 \quad \text{as } n \to \infty. \tag{4.3}$$

The result in (4.3) implies that if the population eigenvalue is not much bigger than one, then the corresponding sample eigenvector is *strongly inconsistent* to the population eigenvector in the sense that the two vectors become perpendicular.

4.2.2 HDLSS Asymptotics

While the results in the previous section treat the case where d and n grow together, we let d go to infinity with n fixed in this section. Geometrical representation of HDLSS data and the HDLSS asymptotics involving the sample covariance matrices are reviewed in the following two subsections.

Geometric Representation

Hall *et al.* (2005) studied the geometrical representation of the HDLSS data. Assume $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are n independent random samples from the $d-$dimensional multivariate Gaussian distribution with mean 0 and an identity covariance matrix. Then, the distance between any two samples tends to be deterministic in the following sense:

$$\| \mathbf{x}_i - \mathbf{x}_j \| = (2d)^{1/2} + Op(1) \quad \text{as } d \to \infty.$$

In other words, n-data points independently drawn from Gaussian essentially lie at the vertices of a regular n-simplex in $\mathbb{R}^d$. Thus, the increasing randomness in the data appear only in terms of a random rotation. Modulo rotation the behavior is essentially deterministic. Also, Hall *et al.* (2005) studied the geometric representation of HDLSS data from two different distributions which arise in the context of discrimination. In particular, this was used to analyze the limiting behavior of various discriminant methods, such as Support Vector Machine (SVM) and Distance Weighted Discrimination (DWD).

Sample Covariance Matrices

Consider a sequence of $d \times n$ data matrices $\mathbf{X} := [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ from $\mathcal{N}_d(0, \Sigma_d)$ for $d = 1, 2, \cdots$. Define the $n \times n$ dual sample covariance matrix as

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

Ahn *et al.* (2007) studied the conditions under which the dual sample covariance converges to an identity matrix as d grows. Thus, in the limit, the eigenvalues of the sample covariance matrix behave as if they are from a spherical Gaussian. They

formulated the assumption based on the sphericity parameter,

$$\epsilon = \frac{1}{d}\frac{(\sum_{i=1}^{d}\lambda_i)^2}{\sum_{i=1}^{d}\lambda_i^2},$$

where $\{\lambda_i\}$ are eigenvalues of Σ_d. In particular, this parameter satisfies the inequality $1/d \leqslant \epsilon \leqslant 1$. If the underlying population is spherical, i.e., $\Sigma_d = \mathbf{I_d}$, then the parameter takes the extreme value 1. In an extreme singular case where the covariance matrix is rank 1, then ϵ achieves the other extreme value $1/d$. Assume that the population distribution is not too close to the singular case ($1/d \ll \epsilon$) in the sense that $\frac{1}{d\epsilon} \to 0$ as $d \to \infty$. Then, the scaled dual sample covariance matrix,

$$\widehat{\Sigma}_D/c_d \xrightarrow{a.s.} \mathbf{I_d} \quad \text{as } d \to \infty, \tag{4.4}$$

where $c_d = \sum_{i=1}^{d}\lambda_i/n$. This convergence of matrix happens entry-wisely, i.e., diagonals of the dual sample covariance matrix tends to 1 where as off-diagonals tends to 0.

This result will be used in the course of the proof of the inconsistency of the singular vectors of the sample cross-covariance matrix in Section 4.3.3.

4.3 HDLSS Asymptotics of Sample Cross-Covariance Matrices

Consider a data set consisting of n paired multivariate vectors,

$$\{(\mathbf{x}_i, \mathbf{y}_i) \,|\, \mathbf{x}_i \in \mathbb{R}^{d1}, \mathbf{y}_i \in \mathbb{R}^{d_2}, i = 1, \cdots, n\},$$

where the joint distribution of $(\mathbf{x}, \mathbf{y})$ is a $(d_1 + d_2)$- dimensional multivariate Gaussian with mean $\mathbf{0}$ and covariance matrix

$$\Sigma = \left(\begin{array}{c|c} \Sigma_X & \Sigma_{XY} \\ \hline \Sigma_{YX} & \Sigma_Y \end{array} \right).$$

Define the data matrices as $\mathbf{X}_{d_1 \times n} := [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ and $\mathbf{Y}_{d_2 \times n} := [\mathbf{y}_1, \cdots, \mathbf{y}_n]$ and the sample mean matrices as $\overline{\mathbf{X}}_{d_1 \times n}$ and $\overline{\mathbf{Y}}_{d_2 \times n}$ by taking the sample mean for each row (each variable) over the n samples and replicating n copies of the mean vectors.

Most of the work discussed in Section 4.2 studied asymptotic behavior of the sample covariance matrices themselves, or their eigenvalues or eigenvectors. In the following sections, we will focus our attention on the sample paired cross covariance matrices between X and Y, i.e.,

$$\hat{\Sigma}_{XY} \quad = \quad \frac{1}{n} \widetilde{\mathbf{X}} \widetilde{\mathbf{Y}}^T, \tag{4.5}$$

where $\widetilde{\mathbf{X}} = \mathbf{X} - \overline{\mathbf{X}}$ and $\widetilde{\mathbf{Y}} = \mathbf{Y} - \overline{\mathbf{Y}}$.

To distinguish (4.5) from the usual sample covariance matrices, such as $\hat{\Sigma}_X$ or $\hat{\Sigma}_Y$, we will call this the sample cross-covariance matrix in this dissertation. Topics considered here include asymptotic properties of the sample singular values and vectors of $\hat{\Sigma}_{XY}$ when the sample size, n, is fixed, but the dimensions- both d_1 and d_2, or sometimes d_1 only- tend to infinity. We view the data matrices $\mathbf{X}$ and $\mathbf{Y}$ as random matrices, indexed by d_1 and d_2. For the sake of notational simplicity, we will suppress these indices. In Section 4.3.1, SVD or maximum covariance analysis of the cross-covariance matrix is reviewed. In the following two Sections, 4.3.2 and 4.3.3, we will study conditions under which the sample maximum covariance direction vectors are consistent and strongly inconsistent, respectively.

4.3.1 SVD of the Sample Cross-Covariance Matrices

The Singular Value Decomposition (SVD) of the cross-covariance matrix is

$$\begin{aligned}\Sigma_{XY} &= \lambda_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \lambda_m \mathbf{u}_m \mathbf{v}_m^T \\ &= \mathbf{U} \Lambda \mathbf{V}^T\end{aligned}$$

where $m = \text{rank}(\Sigma_{XY})$, $\mathbf{U}_{d_1 \times m} = (\mathbf{u}_1, \cdots, \mathbf{u}_m)$, $\mathbf{V}_{d_2 \times m} = (\mathbf{v}_1, \cdots, \mathbf{v}_m)$ satisfying $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$, and $\Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_m\}$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m > 0$. The vectors, $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_i\}$, are called *singular column vectors* and *row vectors*, respectively, and the $\{\lambda_i\}$ are called *singular values*.

The SVD of the cross-covariance matrix provides the "maximum covariance direction vectors" simultaneously over both the X and Y spaces, and the "maximum covariance" of the underlying population. In particular, the singular vectors $\mathbf{u}_1$ and $\mathbf{v}_1$ solve

$$\max_{\mathbf{u},\mathbf{v}} \quad \text{Cov}(\mathbf{u}^T X, \mathbf{v}^T Y) = \max_{\mathbf{u},\mathbf{v}} \quad \mathbf{u}^T \Sigma_{XY}\ \mathbf{v}$$
$$\text{subject to} \ \| \mathbf{u} \| = \| \mathbf{v} \| = 1$$

and the value of the maximum covariance is given by the corresponding singular value, i.e., $\text{Cov}(\mathbf{u}_1^T X, \mathbf{v}_1^T Y) = \lambda_1$. The next pair of singular vectors, $\mathbf{u}_2$ and $\mathbf{v}_2$, achieves the maximum covariance amongst vectors orthogonal to the previous singular vectors. Henceforth, we will use the term *singular vectors* for maximum covariance direction vectors, and *singular values* for maximum covariance.

However, the underlying population covariance matrices are usually unknown. One can estimate population cross-covariance matrices by the sample cross-covariance matrices as in (4.5).

Letting $r = \text{rank}(\widehat{\Sigma}_{XY})$, the SVD of the sample cross-covariance matrix,

$$\widehat{\Sigma}_{XY} = \widehat{\lambda}_1 \widehat{\mathbf{u}}_1 \widehat{\mathbf{v}}_1^T + \cdots + \widehat{\lambda}_r \widehat{\mathbf{u}}_r \widehat{\mathbf{v}}_r^T,$$

provides the sample singular vectors and the sample singular values.

In the following two sections, asymptotic conditions under which the first sample singular vectors are consistent and strongly inconsistent will be studied. In both cases, the limiting distribution of the first sample singular value is derived.

4.3.2 Spiked Marginal Population Model

In this section, we formulate examples where the first singular vectors of the sample cross covariance matrix are consistent. We consider an extreme case of spiked cross covariance population model.

Asymptotic studies regarding PCA (maximum variance analysis) are usually done by putting assumptions on the population covariance matrices and studying the behavior of the sample covariance matrices in the limit. See Section 4.2 for discussion of these types of study. However, maximum covariance analysis differs from maximum variance analysis in the sense that the object of the study, the sample cross-covariance matrices, are not required to be square. Assumptions made on the population cross-covariance matrix, Σ_{XY}, readily imply some restrictions on the diagonals of the big covariance matrix, Σ, because of the Cauchy-Schwartz inequality, $\text{Cov}(X_i, Y_j)^2 \leqslant \text{Var}(X_i)\text{Var}(Y_j)$. In other words, for maximum covariance study, focus should not be confined to Σ_{XY}, rather Σ_{XY} should be understood as a sub-matrix of the big square matrix Σ.

However, constructing the big covariance matrix, Σ, directly with the non-zero sub-matrix, Σ_{XY}, is not trivial due to non-negative definiteness restrictions on Σ. We note that formulating examples using so-called factor matrix, $\mathbf{F} = \Sigma^{1/2}$, avoids the need for special structures. Assumptions on Σ for the following theorem will be made

through the factor matrix, $\mathbf{F}$.

A non-intuitive, and very interesting, result is obtained when the dimension $d_1 = d_2$ goes to infinity under a certain population model. The main theorem states that consistency of the first sample singular vectors is observed even under some challenging situations where the maximum correlation tends to 0. However, the first sample singular value never converges to the population counterpart even when the maximum correlation is 1.

In this section, we consider a sequence of paired $d \times n$ data matrices $\mathbf{X}$ and $\mathbf{Y}$, where $\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$ is from $\mathcal{N}_{2d}(\mathbf{0}, \Sigma_d)$, based on $\mathbf{F}_d \equiv \Sigma_d^{1/2}$, where

$$\mathbf{F}_d = \left[\begin{array}{c|c} \mathbf{F}_X & \mathbf{F}_{XY} \\ \hline \mathbf{F}_{YX} & \mathbf{F}_Y \end{array}\right],$$

where the $d \times d$ partition matrices are $\mathbf{F}_X = \mathbf{F}_Y = \text{diag}(d^\alpha, 1, \ldots, 1)$ and $\mathbf{F}_{XY} = \mathbf{F}_{YX} = \text{diag}(d^{\alpha\beta}, 0, \ldots, 0)$ for $\alpha > 0$ and $\beta > 0$.

As the dimension grows (as more variables are added to the system), the system gets noisier. In order to explore conditions for this type of asymptotic consistency, the signal needs to get extremely stronger as the system gets noisier. For this reason, the signal that is exponentially growing with increasing dimensionality (both in the marginal factor matrices, $\mathbf{F}_X$ and $\mathbf{F}_Y$, and the cross - factor matrix, $\mathbf{F}_{XY}$) is considered. This also explains why the two parameters α and β are required to be positive (otherwise, the signal shrinks to 0 exponentially fast as d grows.)

One can think of other parametrization, for example, $d^{\alpha+\beta}$, in the cross-factor matrix. With this parametrization, both α and $\alpha + \beta$ need to be positive in order to have increasing signals as above. As a result, the considered region for the pair of (α, β) is not the first quadrant, but a different convex hull. For the sake of simplicity, we choose the former parametrization. All of the results that follow in this section

can be transformed with the latter parametrization as well (basically by replacing $\alpha\beta$ by $\alpha + \beta$.)

The parametrization is made on the factor matrix. The signal intensity in the covariance matrix can be explained only by a combination of α and β. With the factor matrix $\mathbf{F}$ above, the corresponding covariance matrix becomes

$$\Sigma = \left[\begin{array}{c|c} \Sigma_X & \Sigma_{XY} \\ \hline \Sigma_{YX} & \Sigma_Y \end{array}\right] = \left[\begin{array}{cccc|cccc} d^{2\alpha} + d^{2\alpha\beta} & 0 & \cdots & 0 & 2d^{\alpha(1+\beta)} & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \hline 2d^{\alpha(1+\beta)} & 0 & \cdots & 0 & d^{2\alpha} + d^{2\alpha\beta} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{array}\right]. \tag{4.6}$$

For PCA of the marginal data of X or Y, then $\gamma = \max(\alpha, \alpha\beta)$ plays a driving role in the results. This is because $d^{2\gamma}$ becomes a dominant signal in the limit for the marginal covariance, Σ_X or Σ_Y. See Section 4.2.2 for details. In particular, if $\gamma > 1/2$, then the first eigenvalues in Σ_X and Σ_Y dominate so strongly in the limit that the first sample eigenvector converges to the population eigenvector. However, in a mild spiked model for $\gamma \leqslant 1/2$ cases, the dual sample covariance matrices tend to a scaled identity matrix, which makes the sample eigenvalues and eigenvectors behave as those from an identity covariance.

In the maximum covariance analysis across the two data sets, a different spike parameter, τ, defined as

$$\tau = \min(\alpha, \alpha\beta) \tag{4.7}$$

becomes important. This parameter τ reflects the signal intensity of Σ_{XY}. For ex-

ample, if τ is above a certain threshold, say, $1/2$, then the power of d in the signal for Σ_{XY} is at least 1.

Another important interpretation about β is that it can be also viewed as a correlation intensity parameter. Observe that the corresponding correlation matrices are

$$\begin{aligned}\tilde{\Sigma} &= \left[\begin{array}{c|c} \tilde{\Sigma}_X & \tilde{\Sigma}_{XY} \\ \hline \tilde{\Sigma}_{YX} & \tilde{\Sigma}_Y \end{array}\right] \\ &= \left[\begin{array}{c|c} \mathbf{I_d} & \mathrm{diag}(\frac{2d^{\alpha(1+\beta)}}{d^{2\alpha}+d^{2\alpha\beta}}, 0, \ldots, 0) \\ \hline \mathrm{diag}(\frac{2d^{\alpha(1+\beta)}}{d^{2\alpha}+d^{2\alpha\beta}}, 0, \ldots, 0) & \mathbf{I_d} \end{array}\right].\end{aligned}$$

For a fixed integer d and $\alpha > 0$, define the maximum correlation between X and Y as

$$f(\beta) \equiv \mathrm{Corr}(X_1, Y_1) = \frac{2d^{\alpha(1+\beta)}}{d^{2\alpha}+d^{2\alpha\beta}}.$$

If $\beta = 1$, X_1 and Y_1 have a perfect correlation, i.e., X_1 and Y_1 are identical. For other values of β, however, the correlation becomes strictly less than 1. In fact, the farther β is from 1, the smaller the correlation. In particular, it can be shown that f is strictly increasing if $0 < \beta < 1$ and strictly decreasing if $\beta > 1$.

Now, fix β and let d go to infinity. Under the population model, the maximum correlation tends to 0 unless $\beta = 1$. Clearly, if $\beta = 1$, $\mathrm{Corr}(X_1, Y_1) = 1$ for any values of d. For other positive values of β, $\mathrm{Corr}(X_1, Y_1) \to 0$ as $d \to \infty$. In fact, the limiting distribution of the first sample singular value of $\widehat{\Sigma}_{XY}$ is different depending on whether $\beta = 1$ or not.

The following two main theorems essentially give

(i) Consistency of the first sample singular vectors for $\tau > 1/2$ as $d \to \infty$. Note that this holds even in the case where $\beta \neq 1$, i.e., the maximum correlation tends 0.

(ii) Inconsistency of the first sample singular value. This is true even in the perfect correlation case where $\beta = 1$. However, the limiting distributions are derived, and they differ depending on the values of β.

With this covariance structure, note that the SVD of Σ_{XY} is

$$\Sigma_{XY} = 2d^{\alpha(1+\beta)}(1, 0, \cdots, 0)^T(1, 0, \cdots, 0).$$

Hence, the first population singular value and singular vectors are $\lambda_1 = 2d^{\alpha(1+\beta)}$ and $\mathbf{u}_1 = \mathbf{v}_1 = (1, 0, \cdots, 0)^T$.

Theorem 4.3.1. *For $\alpha > 0$ and $\beta > 0$, define τ as in (4.7). If $\tau > 1/2$, then the first sample singular value $\hat{\lambda}_1$ has the following asymptotic properties:*

(i) If $\beta = 1$, $\hat{\lambda}_1$ is approximately distributed as $\lambda_1 \frac{\chi^2_{n-1}}{n}$, where χ^2_{n-1} represents the Chi-square distribution with $n-1$ degrees of freedom, in the sense that

$$\frac{\widehat{\lambda}_1}{\lambda_1} \Longrightarrow \chi^2_{n-1}/n \quad \text{as } d \to \infty.$$

(ii) For other positive values of β, as $d \to \infty$,

$$2d^{-\alpha(1-\beta)}\frac{\widehat{\lambda}_1}{\lambda_1} \Longrightarrow \chi^2_{n-1}/n \quad \text{if } 0 < \beta < 1,$$

$$2d^{-\alpha(\beta-1)}\frac{\widehat{\lambda}_1}{\lambda_1} \Longrightarrow \chi^2_{n-1}/n \quad \text{if } \beta > 1.$$

Remark. Note that $\chi^2_{n-1}/n = 1 + Op(n^{-1/2})$ as the sample size $n \to \infty$. Hence, if $\beta = 1$, with a large number of samples, n, the asymptotic ratio of the sample singular value to the true singular value gets close to 1 as the dimension, d, grows.

When one gets the sample singular vector, $\widehat{\mathbf{u}}_1$ from the data, a natural measure

of the closeness of two unit vectors $\mathbf{u}_1$ and $\widehat{\mathbf{u}}_1$ is the angle between the two vectors:

$$\text{ang}(\mathbf{u}_1, \widehat{\mathbf{u}}_1) = \arccos(< \mathbf{u}_1, \widehat{\mathbf{u}}_1 >) \in [0, \pi/2],$$

where $< \mathbf{u}_1, \widehat{\mathbf{u}}_1 >$ denotes $\mathbf{u}_1^T \widehat{\mathbf{u}}_1$, the inner product of the two vectors.

In the following theorem, we measure the inner product of the true and the sample singular vectors and examine the limiting behavior of them.

Theorem 4.3.2. *For $\alpha > 0$ and $\beta > 0$, define τ as in (4.7). If $\tau > 1/2$, then the first sample singular vectors converge to the first population singular vectors as d tends to infinity, in the sense that*

$$< \mathbf{u}_1, \widehat{\mathbf{u}}_1 > \xrightarrow{p} 1 \quad \text{as } d \to \infty$$

and

$$< \mathbf{v}_1, \widehat{\mathbf{v}}_1 > \xrightarrow{p} 1 \quad \text{as } d \to \infty$$

The following lemmas and corollaries will be used to prove the two Theorems 4.3.1 and 4.3.2. For notational simplicity, in the discussion that follows, we will denote the entries of the $d \times d$ sample cross-covariance matrix, $(\widehat{\Sigma}_{XY})_{ij}$, by σ_{ij} for $i, j = 1, \cdots, d$.

Lemma 4.3.3. *For $\alpha > 0$ and $\beta > 0$, define τ as in (4.7). If $\tau > 1/2$, then entries of the sample cross-covariance matrix, except for σ_{11}, become negligible in the sense that*

$$\frac{1}{\lambda_1^2} \sum_{(i,j) \neq (1,1)} \sigma_{ij}^2 \xrightarrow{a.s.} 0 \quad \text{as } d \to \infty.$$

Corollary 4.3.4. *For $\alpha > 0$ and $\beta > 0$, define τ as in (4.7). If $\tau > 1/2$, then the sample singular values, except for the first one, converge to 0, in the following sense:*

$$\frac{1}{\lambda_1^2} \sum_{i=2}^{m} \widehat{\lambda}_i^2 \xrightarrow{a.s.} 0 \quad \text{as } d \to \infty$$

Proof of Corollary 4.3.4. We will use some properties of the SVD and the Frobenius norm of the matrix. Note that the squares the Frobenius norm of $\widehat{\Sigma}_{XY}$ is defined as the sum of squares of the entries: $\| \widehat{\Sigma}_{XY} \|_F^2 = \sum_{i,j} \sigma_{ij}^2$. One can show that the squares of the Frobenius norm of the matrix can be defined as sum of the squares of singular values, i.e., $\| \widehat{\Sigma}_{XY} \|_F^2 = \sum_{i=1}^m \lambda_i^2$. Using the equivalent definition of the Frobenius norm above,

$$\|\widehat{\Sigma}_{XY} - \widehat{\lambda}_1 \widehat{\mathbf{u}}_1 \widehat{\mathbf{v}}_1^T\|_F^2 = \| \sum_{i=2}^m \widehat{\lambda}_i \widehat{\mathbf{u}}_i \widehat{\mathbf{v}}_i^T \|_F^2 = \sum_{i=2}^m \widehat{\lambda}_i^2. \tag{4.8}$$

Since $\widehat{\lambda}_1 \widehat{\mathbf{u}}_1 \widehat{\mathbf{v}}_1^T$ is the best rank 1 approximation to $\widehat{\Sigma}_{XY}$ in the Frobenius norm sense,

$$\|\widehat{\Sigma}_{XY} - \widehat{\lambda}_1 \widehat{\mathbf{u}}_1 \widehat{\mathbf{v}}_1^T\|_F^2 \leqslant \|\widehat{\Sigma}_{XY} - \sigma_{11}(1, 0 \cdots, 0)^T (1, 0 \cdots, 0)\|_F^2 = \sum_{(i,j) \neq (1,1)} \sigma_{ij}^2. \tag{4.9}$$

Combining (4.8) and (4.9),

$$\frac{1}{\lambda_1^2} \sum_{i=2} \widehat{\lambda}_i^2 \leqslant \frac{1}{\lambda_1^2} \sum_{(i,j) \neq (1,1)} \sigma_{ij}^2$$

and the right hand side converges to 0 a.s. as d grows by Lemma 4.3.3. □

Lemma 4.3.5. *For $\alpha > 0$ and $\beta > 0$, define τ as in (4.7). If $\tau > 1/2$, then σ_{11} dominates the other entries in the following sense:*

$$\frac{1}{\sigma_{11}^2} \sum_{(i,j) \neq (1,1)} \sigma_{ij}^2 \xrightarrow{p} 0 \quad \text{as } d \to \infty.$$

Proof of the two Lammas above will be given at the end of this section.

Corollary 4.3.6. *For $\alpha > 0$ and $\beta > 0$, define τ as in (4.7). If $\tau > 1/2$, then in the limit, the first sample singular value behaves like σ_{11}, the (1,1) entry of the sample*

cross-covariance matrix, in the following sense:

$$\left| \frac{\widehat{\lambda}_1}{\sigma_{11}} - 1 \right| \xrightarrow{p} 0 \quad \text{as } d \to \infty$$

Proof of Corollary 4.3.6. By the unitary invariant property of the Frobenius norm,

$$\sigma_{11}^2 + \sum_{(i,j)\neq(1,1)} \sigma_{ij}^2 = \|\widehat{\Sigma}_{XY}\|_F^2 = \widehat{\lambda}_1^2 + \sum_{i=2}^{m} \widehat{\lambda}_i^2$$

Dividing the above by σ_{11}^2, we have

$$1 + \frac{\sum_{(i,j)\neq(1,1)} \sigma_{ij}^2}{\sigma_{11}^2} = \frac{\widehat{\lambda}_1^2}{\sigma_{11}^2} + \frac{\sum_{i=2}^{m} \widehat{\lambda}_i^2}{\sigma_{11}^2}.$$

Thus,

$$\begin{aligned} \left| \frac{\widehat{\lambda}_1^2}{\sigma_{11}^2} - 1 \right| &= \frac{1}{\sigma_{11}^2} \left| \sum_{i=2}^{m} \widehat{\lambda}_i^2 - \sum_{(i,j)\neq(1,1)} \sigma_{ij}^2 \right| \\ &\leqslant \frac{1}{\sigma_{11}^2} \sum_{i=2}^{m} \widehat{\lambda}_i^2 + \frac{1}{\sigma_{11}^2} \sum_{(i,j)\neq(1,1)} \sigma_{ij}^2 \\ &\leqslant \frac{2}{\sigma_{11}^2} \sum_{(i,j)\neq(1,1)} \sigma_{ij}^2. \end{aligned}$$

The second inequality has been proven in the proof of Corollary 4.3.4 and the last term converges to 0 in probability by Lemma 4.3.5. Therefore,

$$\frac{\widehat{\lambda}_1}{\sigma_{11}} \xrightarrow{p} 1 \quad \text{as } d \to \infty.$$

□

Now we are ready to prove the main theorems. Intuition comes from comparing the highest order in d of the entries of the sample cross-covariance matrix, $\widehat{\Sigma}_{XY}$. For $\alpha > 0$ and $\beta > 0$, if $\tau = \min(\alpha, \alpha\beta) > 1/2$, σ_{11} dominates the rest of the entries as

d tends to infinity. So, when properly scaled, only σ_{11} survives whereas the rest are negligible, which somehow resembles the population cross-covariance matrix. Now let us closely look at what happens when d tends to infinity.

Proof of Theorem 4.3.1. The data matrices $\mathbf{X}$ and $\mathbf{Y}$ can be expressed

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \mathbf{F} \begin{bmatrix} \mathbf{Z}_X \\ \mathbf{Z}_Y \end{bmatrix} = \begin{bmatrix} \mathbf{F}_X \mathbf{Z}_X + \mathbf{F}_{XY} \mathbf{Z}_Y \\ \mathbf{F}_{XY} \mathbf{Z}_X + \mathbf{F}_Y \mathbf{Z}_Y \end{bmatrix}$$

where the entries of $\mathbf{Z}_X$ and $\mathbf{Z}_Y$ are *iid* from $\mathcal{N}(0,1)$. Then, we can write the mean-centered data matrix as

$$\begin{aligned} \mathbf{X} - \overline{\mathbf{X}} &= \mathbf{F}_X(\mathbf{Z}_X - \overline{\mathbf{Z}}_X) + \mathbf{F}_{XY}(\mathbf{Z}_Y - \overline{\mathbf{Z}}_Y) \\ &\equiv \mathbf{F}_X \widetilde{\mathbf{Z}}_X + \mathbf{F}_{XY} \widetilde{\mathbf{Z}}_Y. \end{aligned}$$

Similarly, $\mathbf{Y} - \overline{\mathbf{Y}} = \mathbf{F}_{XY} \widetilde{\mathbf{Z}}_X + \mathbf{F}_Y \widetilde{\mathbf{Z}}_Y$. Then,

$$\begin{aligned} n\hat{\Sigma}_{XY} &= (\mathbf{X} - \overline{\mathbf{X}})(\mathbf{Y} - \overline{\mathbf{Y}})^T \\ &= (\mathbf{F}_X \widetilde{\mathbf{Z}}_X + \mathbf{F}_{XY} \widetilde{\mathbf{Z}}_Y)(\mathbf{F}_{XY} \widetilde{\mathbf{Z}}_X + \mathbf{F}_Y \widetilde{\mathbf{Z}}_Y)^T \\ &= (\mathbf{F}_X \widetilde{\mathbf{Z}}_X \widetilde{\mathbf{Z}}_X^T \mathbf{F}_{XY} + \mathbf{F}_X \widetilde{\mathbf{Z}}_X \widetilde{\mathbf{Z}}_Y^T \mathbf{F}_Y + \mathbf{F}_{XY} \widetilde{\mathbf{Z}}_Y \widetilde{\mathbf{Z}}_X^T \mathbf{F}_{XY} + \mathbf{F}_{XY} \widetilde{\mathbf{Z}}_Y \widetilde{\mathbf{Z}}_Y^T \mathbf{F}_Y) \\ &= \begin{bmatrix} d^\alpha \widetilde{\mathbf{z}}_{1X}^T \\ \widetilde{\mathbf{z}}_{2X}^T \\ \vdots \\ \widetilde{\mathbf{z}}_{dX}^T \end{bmatrix} [d^{\alpha\beta} \widetilde{\mathbf{z}}_{1X}, \mathbf{0}, \ldots, \mathbf{0}] + \begin{bmatrix} d^\alpha \widetilde{\mathbf{z}}_{1X}^T \\ \widetilde{\mathbf{z}}_{2X}^T \\ \vdots \\ \widetilde{\mathbf{z}}_{dX}^T \end{bmatrix} [d^\alpha \mathbf{z}_{1Y}, \mathbf{z}_{2Y}, \ldots, \mathbf{z}_{dY}] \\ &\quad + \begin{bmatrix} d^{\alpha\beta} \widetilde{\mathbf{z}}_{1Y}^T \\ \mathbf{0}^T \\ \vdots \\ \mathbf{0}^T \end{bmatrix} [d^{\alpha\beta} \widetilde{\mathbf{z}}_{1X}, \mathbf{0}, \ldots, \mathbf{0}] + \begin{bmatrix} d^{\alpha\beta} \widetilde{\mathbf{z}}_{1Y}^T \\ \mathbf{0}^T \\ \vdots \\ \mathbf{0}^T \end{bmatrix} [d^\alpha \widetilde{\mathbf{z}}_{1Y}, \widetilde{\mathbf{z}}_{2Y}, \ldots, \mathbf{z}_{dY}]. \end{aligned}$$

The entries of the sample cross-covariance matrix are

$$\begin{aligned}
\sigma_{11} &= \frac{1}{n}(d^{\alpha(1+\beta)}\widetilde{\mathbf{z}}_{1X}^T\widetilde{\mathbf{z}}_{1X} + d^{2\alpha}\widetilde{\mathbf{z}}_{1X}^T\widetilde{\mathbf{z}}_{1Y} + d^{2\alpha\beta}\widetilde{\mathbf{z}}_{1Y}^T\widetilde{\mathbf{z}}_{1X} + d^{\alpha(1+\beta)}\widetilde{\mathbf{z}}_{1Y}^T\widetilde{\mathbf{z}}_{1Y}) \quad (4.10)\\
&= \frac{1}{n}(d^{\alpha}\widetilde{\mathbf{z}}_{1X} + d^{\alpha\beta}\widetilde{\mathbf{z}}_{1Y})^T(d^{\alpha\beta}\widetilde{\mathbf{z}}_{1X} + d^{\alpha}\widetilde{\mathbf{z}}_{1Y}),\\
\sigma_{1j} &= \frac{1}{n}(d^{\alpha}\widetilde{\mathbf{z}}_{1X}^T\widetilde{\mathbf{z}}_{jY} + d^{\alpha\beta}\widetilde{\mathbf{z}}_{1Y}^T\widetilde{\mathbf{z}}_{jY})\\
&= \frac{1}{n}d^{\alpha\beta}(d^{\alpha(1-\beta)}\widetilde{\mathbf{z}}_{1X} + \widetilde{\mathbf{z}}_{1Y})^T\widetilde{z}_{jY} \quad \text{for} \quad j = 2, \ldots, d,\\
\sigma_{i1} &= \frac{1}{n}(d^{\alpha\beta}\widetilde{\mathbf{z}}_{iX}^T\widetilde{\mathbf{z}}_{1X} + d^{\alpha}\widetilde{\mathbf{z}}_{iX}^T\widetilde{\mathbf{z}}_{1Y})\\
&= \frac{1}{n}d^{\alpha\beta}(\widetilde{\mathbf{z}}_{1X} + d^{\alpha(1-\beta)}\widetilde{\mathbf{z}}_{1Y})^T\widetilde{z}_{iX} \quad \text{for} \quad i = 2, \ldots, d,\\
\sigma_{ij} &= \frac{1}{n}\widetilde{\mathbf{z}}_{iX}^T\widetilde{\mathbf{z}}_{jY} \quad \text{otherwise.}
\end{aligned}$$

The limiting distribution of $\widehat{\lambda}_1/\lambda_1$ will be derived in the following manner. First, we will show that $\widehat{\lambda}_1/\lambda_1$ and σ_{11}/λ_1 are getting close to each other as $d \to \infty$. Next, from the exact expression of the entries as in (4.10), the limiting distribution of σ_{11}/λ_1 will be obtained. Now for the first step, the distance between $(\frac{\widehat{\lambda}_1}{\lambda_1})^2$ and $(\frac{\sigma_{11}}{\lambda_1})^2$,

$$\begin{aligned}
\frac{1}{\lambda_1^2}|\widehat{\lambda}_1^2 - \sigma_{11}^2| &= \frac{1}{\lambda_1^2}|\sum_{i=2}^{m}\widehat{\lambda}_i^2 - \sum_{(i,j)\neq(1,1)}\sigma_{ij}^2|\\
&\leqslant \frac{2}{\lambda_1^2}\sum_{(i,j)\neq(1,1)}\sigma_{ij}^2\\
&\xrightarrow{a.s.} 0 \quad \text{as } d \to \infty \qquad (4.11)
\end{aligned}$$

by Lemma 4.3.3. Next, the limiting distribution of σ_{11}/λ_1 depends on β.

- Case *(i)* : $\beta = 1$.

  From (4.10), when $\beta = 1$, note that $\frac{\sigma_{11}}{\lambda_1}$ does not depend on d. In particular,

$$\begin{aligned}
\frac{\sigma_{11}}{\lambda_1} &= \frac{1}{2d^{2\alpha}}\cdot\frac{d^{2\alpha}}{n}(\widetilde{\mathbf{z}}_{1X} + \widetilde{\mathbf{z}}_{1Y})^T(\widetilde{\mathbf{z}}_{1X} + \widetilde{\mathbf{z}}_{1Y})\\
&\overset{d}{=} \frac{1}{2n}\sum_{i=1}^{n}2(z_i - \overline{z})^2
\end{aligned}$$

$$\stackrel{d}{=} \chi^2_{n-1}/n, \tag{4.12}$$

where $\{z_i\}$ are $i.i.d\,\mathcal{N}(0,1)$ and $\overline{z} = \sum_{i=1}^{n} z_i$. Combining this with (4.11),

$$\frac{\widehat{\lambda}_1}{\lambda_1} \stackrel{a.s.}{\to} \frac{\sigma_{11}}{\lambda_1} \stackrel{d}{=} \chi^2_{n-1} \quad \text{as } d \to \infty,$$

completes the proof of Case (i).

- Case *(ii)* : $0 < \beta < 1$.

$$\begin{aligned} \frac{\sigma_{11}}{\lambda_1} &= \frac{1}{2d^{\alpha(1+\beta)}} \frac{1}{n} (d^{\alpha}\widetilde{\mathbf{z}}_{1X} + d^{\alpha\beta}\widetilde{\mathbf{z}}_{1Y})^T (d^{\alpha}\widetilde{\mathbf{z}}_{1X} + d^{\alpha\beta}\widetilde{\mathbf{z}}_{1Y}) \\ &\stackrel{d}{=} \frac{1}{2nd^{\alpha(1+\beta)}} \sum_{i=1}^{n} (d^{2\alpha} + d^{2\alpha\beta})(z_i - \overline{z})^2 \\ &\stackrel{d}{=} \frac{d^{2\alpha} + d^{2\alpha\beta}}{2d^{\alpha(1+\beta)}} \chi^2_{n-1}/n, \end{aligned} \tag{4.13}$$

where $\{z_i\}$ and $\overline{z}$ are defined as above in Case (i). Multiplying both sides by $2/d^{\alpha(1-\beta)}$,

$$\frac{2}{d^{\alpha(1-\beta)}} \frac{\sigma_{11}}{\lambda_1} \stackrel{d}{=} \frac{d^{2\alpha} + d^{2\alpha\beta}}{d^{2\alpha}} \chi^2_{n-1}/n \Longrightarrow \chi^2_{n-1}/n \quad \text{as } d \to \infty.$$

Combining this with (4.11), we get the limiting distribution of the ratio,

$$2d^{-\alpha(1-\beta)} \frac{\widehat{\lambda}_1}{\lambda_1} \Longrightarrow \chi^2_{n-1}/n \quad \text{as } d \to \infty.$$

- Case *(iii)* : $\beta > 1$.

  The representation of $\frac{\sigma_{11}}{\lambda_1}$ is the same as (4.13) in the Case *(ii)*. Now, multiplying both sides of (4.13) by $2/d^{\alpha(\beta-1)}$,

$$\frac{2}{d^{\alpha(\beta-1)}} \frac{\sigma_{11}}{\lambda_1} \stackrel{d}{=} \frac{d^{2\alpha} + d^{2\alpha\beta}}{d^{2\alpha\beta}} \chi^2_{n-1}/n \Longrightarrow \chi^2_{n-1}/n \quad \text{as } d \to \infty.$$

Combining this with (4.11), we get the limiting distribution of the ratio,

$$2d^{-\alpha(\beta-1)}\frac{\widehat{\lambda}_1}{\lambda_1} \Longrightarrow \chi^2_{n-1}/n \quad \text{as } d \to \infty.$$

□

Now the proof of the consistency of the singular vectors follows.

Proof of Theorem 4.3.2. We will first show that $\widehat{\lambda}_1\widehat{\mathbf{u}}_1\widehat{\mathbf{v}}_1^T$ is getting close to $\sigma_{11}(1,0,\cdots,0)^T(1,0,\cdots,0)$ as d grows. Then, since the scalar $\widehat{\lambda}_1$ is like σ_{11} in the limit as shown in Corollary 4.3.6, the two vectors, $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{v}}_1$ are expected to be close to $(1,0,\cdots,0)^T$ in the limit.

$$\begin{aligned}
\frac{1}{\sigma_{11}}\|\widehat{\lambda}_1\widehat{\mathbf{u}}_1\widehat{\mathbf{v}}_1^T &- \sigma_{11}(1,0,\cdots,0)^T(1,0,\cdots,0)\|_F \\
&\leqslant \frac{1}{\sigma_{11}}\|\widehat{\Sigma}_{XY} - \widehat{\lambda}_1\widehat{\mathbf{u}}_1\widehat{\mathbf{v}}_1^T\|_F + \frac{1}{\sigma_{11}}\|\widehat{\Sigma}_{XY} - \sigma_{11}(1,0,\cdots,0)^T(1,0,\cdots,0)\|_F \\
&\leqslant \frac{2}{\sigma_{11}}\|\widehat{\Sigma}_{XY} - \sigma_{11}(1,0,\cdots,0)^T(1,0,\cdots,0)\|_F \\
&= \frac{2}{\sigma_{11}}\sqrt{\sum_{(i,j)\neq(1,1)}\sigma_{ij}^2}. \qquad (4.14)
\end{aligned}$$

The second inequality holds because $\widehat{\lambda}_1\widehat{\mathbf{u}}_1\widehat{\mathbf{v}}_1^T$ is the best rank 1 approximation to $\widehat{\Sigma}_{XY}$ in the Frobenious norm sense. The last quantity goes to 0 in probability by Lemma 4.3.5. Since the Frobenius norm is unitarily invariant,

$$\begin{aligned}
\frac{1}{\sigma_{11}}\|\widehat{\lambda}_1\widehat{\mathbf{u}}_1\widehat{\mathbf{v}}_1^T &- \sigma_{11}(1,0,\cdots,0)^T(1,0,\cdots,0)\|_F \\
&= \frac{1}{\sigma_{11}}\|\widehat{\lambda}_1 - \sigma_{11}\widehat{\mathbf{u}}_1^T(1,0,\cdots,0)^T(1,0,\cdots,0)\widehat{\mathbf{v}}_1\|_F \\
&= \left|\frac{\widehat{\lambda}_1}{\sigma_{11}} - < \widehat{\mathbf{u}}_1, \mathbf{u}_1 >< \widehat{\mathbf{v}}_1, \mathbf{v}_1 >\right|. \qquad (4.15)
\end{aligned}$$

With (4.14) and (4.15) together, we obtain

$$\left| \frac{\widehat{\lambda}_1}{\sigma_{11}} - < \widehat{\mathbf{u}}_1, \mathbf{u}_1 >< \widehat{\mathbf{v}}_1, \mathbf{v}_1 > \right| \xrightarrow{p} 0 \quad \text{as } d \to \infty. \tag{4.16}$$

On the other hand,

$$\begin{aligned} |< \widehat{\mathbf{u}}_1, \mathbf{u}_1 >< \widehat{\mathbf{v}}_1, \mathbf{v}_1 > -1| &\leqslant \left| \frac{\widehat{\lambda}_1}{\sigma_{11}} - < \widehat{\mathbf{u}}_1, \mathbf{u}_1 >< \widehat{\mathbf{v}}_1, \mathbf{v}_1 > \right| + \left| \frac{\widehat{\lambda}_1}{\sigma_{11}} - 1 \right| \\ &\xrightarrow{p} 0 \quad \text{as } d \to \infty \end{aligned}$$

by (4.16) and Lemma 4.3.5. Thus,

$$< \widehat{\mathbf{u}}_1, \mathbf{u}_1 >< \widehat{\mathbf{v}}_1, \mathbf{v}_1 > \xrightarrow{p} 1 \quad \text{as } d \to \infty.$$

Since the inner product of two unit vectors cannot exceed 1, this readily means that either

$$< \widehat{\mathbf{u}}_1, \mathbf{u}_1 > \xrightarrow{p} 1 \quad \text{and} \quad < \widehat{\mathbf{v}}_1, \mathbf{v}_1 > \xrightarrow{p} 1 \quad \text{as } d \to \infty$$

or

$$< \widehat{\mathbf{u}}_1, \mathbf{u}_1 > \xrightarrow{p} -1 \quad \text{and} \quad < \widehat{\mathbf{v}}_1, \mathbf{v}_1 > \xrightarrow{p} -1 \quad \text{as } d \to \infty.$$

By negating the signs of the sample singular vectors, we can exclude the latter case without loss of generality. In conclusion, both the first sample singular row and column vectors are consistent to the population singular vectors. □

Remark on generalization of Theorem 4.3.2 . Suppose that Σ_{XY} is a diagonal matrix with the first few diagonal entries growing with some power of d, say, $d^{\tau_1}, d^{\tau_2}, \cdots,$ and d^{τ_M}, where $\tau_1 > \tau_2 > \cdots > \tau_M > 0$. Also, assume that the marginal covariance matrices Σ_X and Σ_Y have similar large diagonal entries, that are large enough that Σ is nonnegative definite. For $M = 1$, the proof of Theorem 4.3.2 essentially makes use of the fact that, if $\tau_1 > 1$, then $\sigma_{11}(1, 0, \cdots, 0)^T(1, 0, \cdots, 0)$ becomes a good

rank one approximation to $\widehat{\Sigma}_{XY}$ as d grows. A similar phenomenon is expected for $M > 1$. Namely, if $\tau_M > 1$, the first M diagonal entries of $\widehat{\Sigma}_{XY}$, $\{\sigma_{11}, \cdots, \sigma_{MM}\}$, dominate the rest of the entries in a sequential manner so that $\Sigma_{j=1}^{m} \sigma_{jj} \mathbf{e}_j \mathbf{e}_j^T$, where $\mathbf{e}_j$ is the standard j-th coordinate vector, becomes a good rank m approximation to $\widehat{\Sigma}_{XY}$ for $m = 1, \cdots, M$. However, substantial refinement of the proof will be needed to rigorously establish this.

Example 1. Here, some illustrative examples are given to show the validity of the HDLSS asymptotics of the sample singular values provided in Theorem 4.3.1. Set $n = 20$, $\alpha = 1$, $\beta = 1$ and take $d = 200$. Generate a data set $(\mathbf{X}, \mathbf{Y})$ from $\mathcal{N}_{2d}(\mathbf{0}, \Sigma_d)$, where Σ_d has the form in (4.6). Estimate the sample cross-covariance matrix, and singular value of it via SVD. Repeat this procedure $M = 500$ times to get a reasonable distribution of the singular values. In Figure 4.1, the Q-Q plot of the ratio of the sample singular value to the true singular value, $\frac{\widehat{\lambda}_1}{\lambda_1}$, against the χ^2_{n-1}/n is shown as a red curve. The green line, the 45° line, shows the theoretical quantiles from the χ^2_{n-1}/n distribution. To understand the natural variation in the red curve, we use a Q-Q Envelope plot. For this, 100 Q-Q plots of random samples from χ^2_{n-1}/n are displayed as blue curves. The red curve is inside the bundle of blue curves, which shows the validity of the asymptotics.

Proof of Lemma 4.3.3. From (4.10), the entries of the sample cross-covariance matrix can be written in terms of the row vectors of the mean-centered data matrix, $\widetilde{\mathbf{Z}}_X := \mathbf{Z}_X - \overline{\mathbf{Z}}_X$ and $\widetilde{\mathbf{Z}}_Y := \mathbf{Z}_Y - \overline{\mathbf{Z}}_Y$.

- Case *(i)*: $\beta = 1$.

$$\begin{aligned}\frac{1}{\lambda_1^2} \sum_{(i,j)\neq(1,1)} \sigma_{ij}^2 &= \frac{1}{d^{4\alpha}} \Big[\sum_{j=2}^{d} \Big\{ \frac{d^\alpha}{n} (\widetilde{z}_{1X} + \widetilde{z}_{1Y})^T \widetilde{z}_{jY} \Big\}^2 + \sum_{i=2}^{d} \Big\{ \frac{d^\alpha}{n} (\widetilde{z}_{1X} + \widetilde{z}_{1Y})^T \widetilde{z}_{iX} \Big\}^2 \\ &\quad + \sum_{i,j=2}^{d} \big(\frac{1}{n} \widetilde{z}_{iX}^T \widetilde{z}_{jY} \big)^2 \Big]\end{aligned}$$

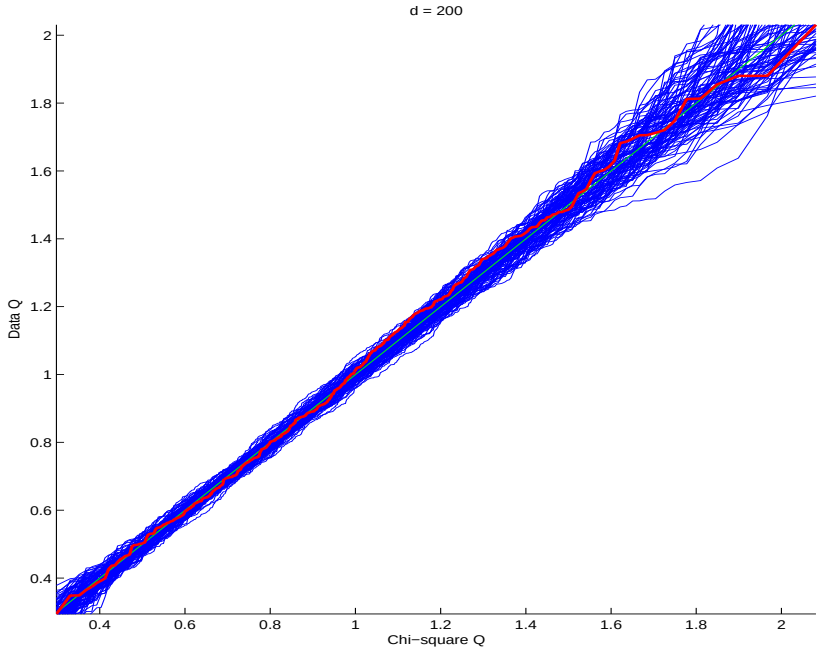

Figure 4.1: Quantile-Quantile Envelope plot for testing the distribution of $\{\frac{\widehat{\lambda}_1}{\lambda_1}\}$ against the χ^2_{n-1}/n distribution. The red curve is inside the blue bundle, which shows the validity of the asymptotics.

$$
\begin{aligned}
\leqslant \quad & \frac{1}{n^2 d^{2\alpha}} \parallel \widetilde{z}_{1X} + \widetilde{z}_{1Y} \parallel^2 \cdot \sum_{j=2}^{d} \parallel \widetilde{z}_{jY} \parallel^2 \\
& + \frac{1}{n^2 d^{2\alpha}} \parallel \widetilde{z}_{1X} + \widetilde{z}_{1Y} \parallel^2 \cdot \sum_{i=2}^{d} \parallel \widetilde{z}_{iX} \parallel^2 \\
& + \frac{1}{n^2 d^{4\alpha}} \sum_{i,j=2}^{d} \parallel \widetilde{z}_{iX} \parallel^2 \cdot \parallel \widetilde{z}_{jY} \parallel^2
\end{aligned}
$$

by the Cauchy-Schwartz inequality. Now, let's look at the first term of the right hand side of the inequality.

$$
\frac{1}{n^2 d^{2\alpha}} \parallel \widetilde{z}_{1X} + \widetilde{z}_{1Y} \parallel^2 \cdot \sum_{j=2}^{d} \parallel \widetilde{z}_{jY} \parallel^2 = \frac{1}{n^2 d^{2\alpha-1}} \parallel \widetilde{z}_{1X} + \widetilde{z}_{1Y} \parallel^2 \cdot \frac{1}{d} \sum_{j=2}^{d} \parallel \widetilde{z}_{jY} \parallel^2 \tag{4.17}
$$

Note that $\parallel \widetilde{z}_{jY} \parallel^2 = \widetilde{z}_{jY}^T \widetilde{z}_{jY}$ is just the sum of the squares of the sample mean-

centered n Gaussian draws. Thus, $\| \widetilde{z}_{jY} \|^2$ are *i.i.d.* χ^2_{n-1} for $j = 2, \cdots, d$. By the Law of Large Numbers (LLN),

$$\frac{1}{d}\sum_{j=2}^{d} \| \widetilde{z}_{jY} \|^2 \overset{a.s.}{\rightarrow} \mathbf{E} \| \widetilde{z}_{2Y} \|^2 = n-1 \quad \text{as } d \rightarrow \infty.$$

Therefore, if $\alpha > 1/2$, then (4.17) converges to 0 almost surely as $d \rightarrow \infty$ since it is the product of (1) a non-random sequence converging to 0 and (2) a random sequence converging to a constant almost surely. Similarly, the second term converges to 0 a.s. as d grows. For the third term, apply the LLN to the sequence of *i.i.d.* random variables $\{\| \widetilde{z}_{iX} \|^2 \cdot \| \widetilde{z}_{jY} \|^2\}_{i,j=2}^{d}$, with finite first moment $\mathbf{E} \| \widetilde{z}_{iX} \|^2 \cdot \| \widetilde{z}_{jY} \|^2 = (n-1)^2$. If $\alpha > 1/2$, then

$$\frac{1}{n^2 d^{4\alpha-2}} \cdot \frac{1}{d^2} \sum_{i,j=2}^{d} \| \widetilde{z}_{iX} \|^2 \cdot \| \widetilde{z}_{jY} \|^2 \overset{a.s.}{\rightarrow} 0 \quad \text{as } d \rightarrow \infty.$$

Since all of the three terms in (4.17) converge to 0 a.s., this completes the proof.

- Case *(ii)*: $0 < \beta < 1$.

  Similarly as above in Case *(i)*, by applying the Cauchy-Schwartz inequality, observe that

$$\begin{aligned}
\frac{1}{\lambda_1^2} \sum_{(i,j)\neq(1,1)} \sigma_{ij}^2 &= \frac{1}{4d^{2\alpha(1+\beta)}} \Big[\sum_{j=2}^{d} \Big\{ \frac{d^{\alpha\beta}}{n} (d^{\alpha(1-\beta)} \widetilde{z}_{1X} + \widetilde{z}_{1Y})^T \widetilde{z}_{jY} \Big\}^2 \\
&\quad + \sum_{i=2}^{d} \Big\{ \frac{d^{\alpha\beta}}{n} (\widetilde{z}_{1X} + d^{\alpha(1-\beta)} \widetilde{z}_{1Y})^T \widetilde{z}_{iX} \Big\}^2 + \sum_{i,j=2}^{d} \Big(\frac{1}{n} \widetilde{z}_{iX}^T \widetilde{z}_{jY} \Big)^2 \Big] \\
&\leqslant \frac{1}{4n^2 d^{2\alpha}} \| d^{\alpha(1-\beta)} \widetilde{z}_{1X} + \widetilde{z}_{1Y} \|^2 \sum_{j=2}^{d} \| \widetilde{z}_{jY} \|^2 \\
&\quad + \frac{1}{4n^2 d^{2\alpha}} \| \widetilde{z}_{1X} + d^{\alpha(1-\beta)} \widetilde{z}_{1Y} \|^2 \sum_{i=2}^{d} \| \widetilde{z}_{iX} \|^2 \\
&\quad + \frac{1}{4n^2 d^{2\alpha(1+\beta)}} \sum_{i,j=2}^{d} \| \widetilde{z}_{iX} \|^2 \| \widetilde{z}_{jY} \|^2 . \qquad (4.18)
\end{aligned}$$

Apply the following inequality to the first term: $\| x+y \|^2 \leqslant 2\{\| x \|^2 + \| y \|^2\}$ for two vectors, x and y. If $\alpha\beta > 1/2$, then

$$\begin{aligned}
\frac{1}{4n^2 d^{2\alpha}} \quad \cdot \quad & \| d^{\alpha(1-\beta)}\widetilde{z}_{1X} + \widetilde{z}_{1Y} \|^2 \sum_{j=2}^{d} \| \widetilde{z}_{jY} \|^2 \\
\leqslant \quad & \frac{2}{4n^2 d^{2\alpha-1}} \{d^{2\alpha(1-\beta)} \| \widetilde{z}_{1X} \|^2 + \| \widetilde{z}_{1Y} \|^2\} \cdot \frac{1}{d} \sum_{j=2}^{d} \| \widetilde{z}_{jY} \|^2 \\
\stackrel{a.s.}{\to} \quad & 0 \quad \text{as } d \to \infty.
\end{aligned}$$

Similarly, the second term can be shown to converge to 0 a.s. The third term,

$$\begin{aligned}
\frac{1}{4n^2 d^{2\alpha(1+\beta)}} \sum_{i,j=2}^{d} \| \widetilde{z}_{iX} \|^2 \| \widetilde{z}_{jY} \|^2 \quad &= \quad \frac{1}{4n^2 d^{2\alpha(1+\beta)-2}} \cdot \frac{1}{d^2} \sum_{i,j=2}^{d} \| \widetilde{z}_{iX} \|^2 \| \widetilde{z}_{jY} \|^2 \\
&\stackrel{a.s.}{\to} \quad 0 \quad \text{as } d \to \infty,
\end{aligned}$$

because the power of d,

$$2 - 2\alpha(1+\beta) < 2 - 4\alpha\beta < 0.$$

This completes the proof.

- Case *(iii)*: $\beta > 1$.

  The same as the Case (ii) except that the first two terms in (4.18) are replaced by

  $$\frac{1}{4n^2 d^{2\alpha\beta}} \| \widetilde{z}_{1X} + d^{\alpha(\beta-1)}\widetilde{z}_{1Y} \|^2 \sum_{j=2}^{d} \| \widetilde{z}_{jY} \|^2$$

  and

  $$\frac{1}{4n^2 d^{2\alpha\beta}} \| \widetilde{z}_{1X} + d^{\alpha(\beta-1)}\widetilde{z}_{1Y} \|^2 \sum_{i=2}^{d} \| \widetilde{z}_{iX} \|^2,$$

  respectively. With the same argument as above for the Case (ii), if $\alpha > 1/2$, these two terms converge to 0 a.s. as $d \to \infty$. This completes the proof.

Combining all the three cases above, if $\tau = \min(\alpha, \alpha\beta) > 1/2$, then

$$\frac{1}{\lambda_1^2} \sum_{(i,j)\neq(1,1)} \sigma_{ij}^2 \xrightarrow{P} 0 \quad \text{as } d \to \infty.$$

□

Proof of Lemma 4.3.5. Use Lemma 4.3.3 and the equation (4.12) for $\beta = 1$ case. Then, as $d \to \infty$,

$$\begin{aligned} \frac{1}{\sigma_{11}^2} \sum_{(i,j)\neq(1,1)} \sigma_{ij}^2 &= \frac{\lambda_1^2}{\sigma_{11}^2} \cdot \frac{\sum_{(i,j)\neq(1,1)} \sigma_{ij}^2}{\lambda_1^2} \\ &= O_p(1)o_p(1) \\ &= o_p(1). \end{aligned}$$

For other positive values of β, by Lemma 4.3.3 and (4.13), as $d \to \infty$,

$$\begin{aligned} \frac{1}{\sigma_{11}^2} \sum_{(i,j)\neq(1,1)} \sigma_{ij}^2 &= \frac{\lambda_1^2}{\sigma_{11}^2} \cdot \frac{\sum_{(i,j)\neq(1,1)} \sigma_{ij}^2}{\lambda_1^2} \\ &= \{\frac{2d^{2\alpha(1+\beta)}}{d^{2\alpha} + d^{2\alpha\beta}}\}^2 O_p(1) \cdot o_p(1) \\ &= o(1)O_p(1)o_p(1) \\ &= o_p(1). \end{aligned}$$

□

4.3.3 Spherical Marginal Population Model

In the previous section, we saw that the first sample singular vectors are consistent to the true singular vectors as $d \to \infty$ under the spiked marginal model. Shortly in this section, we will see the inconsistency of the sample singular vectors under the spherical marginal population model. In fact, the sample singular vectors are *strongly inconsistent* in the sense that it is orthogonal to the population singular vector with

the growing dimension d.

We consider a simple setting where $d_2 = 1$ and there is perfect correlation between Y and a linear combination of the X variables. The considered covariance matrices are of the following form:

$$\Sigma = \left(\begin{array}{c|c} \Sigma_X & \Sigma_{XY} \\ \hline \Sigma_{YX} & \Sigma_Y \end{array}\right) = \begin{pmatrix} \mathbf{I} & \mathbf{u}_1 \\ \mathbf{u}_1^T & 1 \end{pmatrix}, \tag{4.19}$$

where $\mathbf{u}_1^T = (\frac{1}{\sqrt{m}}, \cdots, \frac{1}{\sqrt{m}}, 0, \cdots, 0)$ having $\frac{1}{\sqrt{m}}$ in the first m-coordinates, and zeros elsewhere, where $1 \leqslant m \leqslant d$.

Theorem 4.3.7. *For $d_2 = 1$ and a fixed n, consider a sequence of paired data matrices $(\mathbf{X}, \mathbf{Y})$ from a $(d+1)$-dimensional multivariate Gaussian, with mean 0 and covariance matrix of the form* (4.19)*. Then,*

(i)

$$\sqrt{d} \ < \mathbf{u}_1, \widehat{\mathbf{u}}_1 > \Longrightarrow \sqrt{\chi^2_{n-1}} \quad as \quad d \to \infty$$

and

(ii)

$$d^{-1/2}\widehat{\lambda}_1 \Longrightarrow \sqrt{\chi^2_{n-1}/n} \quad as \quad d \to \infty.$$

An immediate consequence of this theorem is that

$$\mathbf{u}_1^T \widehat{\mathbf{u}}_1 \xrightarrow{P} 0 \quad \text{as } d \to \infty.$$

The sample maximum covariance vector $\widehat{\mathbf{u}}_1$ is strongly inconsistent in the sense that it is orthogonal to the theoretical covariance vector $\mathbf{u}_1$ in this setting. Note that the inner product of the two direction vectors converges to 1 in the consistent case.

Proof. Note that the singular column vector, $\widehat{\mathbf{u}}_1$, is just the normalized $d \times 1$ sample

cross-covariance matrix, $\widehat{\Sigma}_{XY}$, i.e.,

$$\widehat{\mathbf{u}}_1 = \frac{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T}{\| \widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T \|} = \frac{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T}{\sqrt{\widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T}}.$$

Write the mean-centered data matrix as $\widetilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T) = \mathbf{X}(\mathbf{I} - \frac{1}{n}\mathbf{J})$ where $\mathbf{1}$ and $\mathbf{J}$ denote an $n \times 1$ vector of ones and an $n \times n$ matrix of ones, respectively. Similarly, $\widetilde{\mathbf{Y}} = \mathbf{Y}(\mathbf{I} - \frac{1}{n}\mathbf{J})$. Using the fact that $(\mathbf{I} - \frac{1}{n}\mathbf{J})$ is symmetric and idempotent, one can write $\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T = \mathbf{X}(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y}^T = \mathbf{X}\widetilde{\mathbf{Y}}^T$ and $\widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T = \mathbf{Y}(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{X}^T\mathbf{X}(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y}^T = \widetilde{\mathbf{Y}}\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{Y}}^T$. Then,

$$\begin{aligned}
\sqrt{d}\ \mathbf{u}_1^T\widehat{\mathbf{u}}_1 &= \sqrt{d}(\frac{1}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}}, 0, \cdots, 0) \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_d^T \end{pmatrix} \widetilde{\mathbf{Y}}^T / \sqrt{\widetilde{\mathbf{Y}}\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{Y}}^T} \\
&= \frac{1}{\sqrt{m}}(\mathbf{x}_1 + \cdots + \mathbf{x}_m)^T\widetilde{\mathbf{y}} / \sqrt{\widetilde{\mathbf{y}}^T(\mathbf{X}^T\mathbf{X}/d)\widetilde{\mathbf{y}}}, \qquad (4.20)
\end{aligned}$$

where $\mathbf{x}_1^T, \ldots, \mathbf{x}_d^T$ and $\widetilde{\mathbf{y}}^T$ denote the rows of $\mathbf{X}$ and $\widetilde{\mathbf{Y}}$, respectively. Shortly, we will see that

- the two random vectors $\mathbf{y}$ and $\frac{1}{\sqrt{m}}(\mathbf{x}_1 + \cdots + \mathbf{x}_m)$ in the numerator are identical, hence, the numerator is the same as $\mathbf{y}^T\widetilde{\mathbf{y}} = \widetilde{\mathbf{y}}^T\widetilde{\mathbf{y}}$ and
- the term in the denominator, $\widetilde{\mathbf{y}}^T(\mathbf{X}^T\mathbf{X}/d)\widetilde{\mathbf{y}}$ converges to $\widetilde{\mathbf{y}}^T\widetilde{\mathbf{y}}$ in probability as $d \to \infty$.

Putting these observations altogether, we have

$$d\ (\mathbf{u}_1^T\widehat{\mathbf{u}}_1)^2 = \frac{(\widetilde{\mathbf{y}}^T\widetilde{\mathbf{y}})^2}{\widetilde{\mathbf{y}}^T\widetilde{\mathbf{y}} + o_P(1)} \Longrightarrow \frac{(\widetilde{\mathbf{y}}^T\widetilde{\mathbf{y}})^2}{\widetilde{\mathbf{y}}^T\widetilde{\mathbf{y}}} = \widetilde{\mathbf{y}}^T\widetilde{\mathbf{y}} \stackrel{d}{=} \chi^2(n-1) \text{ as } d \to \infty.$$

Finally, the Continuity mapping theorem completes part (i) of Theorem 4.3.7.

Now, let's closely look at the numerator and the denominator of (4.20). Note that

$$\mathrm{Var}(\frac{1}{\sqrt{m}}X_1 + \cdots + \frac{1}{\sqrt{m}}X_m) = \frac{1}{m}\mathrm{Var}(X_1 + \cdots + X_m) = 1$$

and

$$\mathrm{Cov}(\frac{1}{\sqrt{m}}X_1 + \cdots + \frac{1}{\sqrt{m}}X_m, Y) = \frac{1}{\sqrt{m}}\{\mathrm{Cov}(X_1, Y) + \cdots + \mathrm{Cov}(X_m, Y)\} = 1.$$

Therefore, $\mathrm{Corr}(\frac{1}{\sqrt{m}}X_1 + \cdots + \frac{1}{\sqrt{m}}X_m, Y) = 1$. In particular, this means that the linear combination $\frac{1}{\sqrt{m}}X_1 + \cdots + \frac{1}{\sqrt{m}}X_m$ is identical to Y since the joint distribution of them is the multivariate Gaussian. Hence, the data vector of length n, $\frac{1}{\sqrt{m}}(\mathbf{x}_1 + \cdots + \mathbf{x}_m)$, is the same as $\mathbf{y}$.

In the denominator of (4.20), $\mathbf{X}^T\mathbf{X}/d$, is the $n \times n$ dual sample covariance matrix, multiplied by n/d. Using the HDLSS asymptotic result by Ahn *et al.* (2007) in section 4.2.2, $(\mathbf{X}^T\mathbf{X}/d)_{ij} \to \delta_{ij}$ in probability as $d \to \infty$, so the denominator in (4.20)

$$\begin{aligned}
\widetilde{\mathbf{y}}^T(\mathbf{X}^T\mathbf{X}/d)\widetilde{\mathbf{y}} &= \sum_{ij}^{n} \widetilde{y}_i(\mathbf{X}^T\mathbf{X}/d)_{ij}\widetilde{y}_j \\
&= \sum_{ij}^{n} \widetilde{y}_i\big(\delta_{ij} + o_P(1)\big)\widetilde{y}_j \\
&= \sum_{i}^{n} \widetilde{y}_i^2 + o_P(1) \quad \text{as } d \to \infty.
\end{aligned}$$

Similarly, for the sample singular value,

$$d^{-1}\widehat{\lambda}_1^2 = d^{-1} \parallel \frac{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T}{n} \parallel^2 = \frac{\widetilde{\mathbf{y}}^T(\mathbf{X}^T\mathbf{X}/d)\widetilde{\mathbf{y}}}{n^2} = \frac{\widetilde{\mathbf{y}}^T\widetilde{\mathbf{y}} + o_P(1)}{n^2} \Longrightarrow \frac{\chi^2(n-1)}{n^2} \quad \text{as } d \to \infty,$$

which completes part (ii) of Theorem 4.3.7. □

Example 2. We use a simple example to illustrate the d-asymptotic behavior of the

sample maximum covariance vector $\widehat{\mathbf{u}}_1$, and the sample maximum covariance $\widehat{\lambda}_1$, stated in Theorem 4.3.7. Set $d = 500$ and the covariance between X and Y as $\mathbf{u}_1 = (1, 0, \cdots, 0)^T$. We generate a paired data set $\{(\mathbf{x}_i, y_i)\}$ of size $n = 20$ from $N_{d+1}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} \mathbf{I} & \mathbf{u}_1 \\ \mathbf{u}_1^T & 1 \end{pmatrix}$. In this model, only the first variable X_1 is related to Y, and the rest of them, $X_2, \cdots, X_d$, are independent of Y. For each simulation, the sample maximum covariance vector $\widehat{\mathbf{u}}_1$ is estimated, and the quantity $d\ (\mathbf{u}_1 \widehat{\mathbf{u}}_1)^2$ seen in Theorem 4.3.7 is recorded. Call this simply the "scaled inner product" in this example. The same procedure is done for $M = 250$ simulations. The Q-Q plot of the scaled inner product from 250 simulations against the χ^2_{n-1} distribution with $n = 20$ is shown in Figure 4.2. The red curve shows the quantiles of the scaled inner product, and the green line is the straight 45° line which indicates the theoretical quantiles from the χ^2_{n-1} distribution. The blues curves are the quantiles from 100 simulations from the χ^2_{n-1} distribution, showing the variation of quantiles existing in χ^2_{n-1} random samples. Note that the red curve is consistent with the envelope of blue curves, which shows the validity of the approximation of the scaled inner products to a χ^2 distribution for a large dimension d.

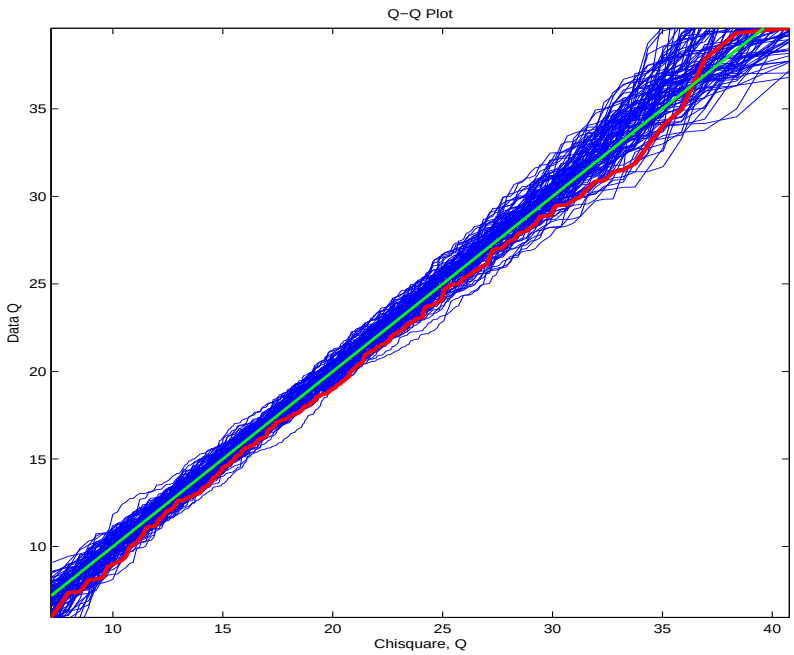

Figure 4.2: Quantile-Quantile plot for testing distributional form of $\{d\,(\mathbf{u}_1\widehat{\mathbf{u}}_1)^2\}$ against the χ^2 distribution. The red curve shows the quantiles of $\{d\,(\mathbf{u}_1\widehat{\mathbf{u}}_1)^2\}$, and the green line indicates the theoretical quantiles from the χ^2_{n-1} distribution. The blue curves show the variation of quantiles existing in a χ^2_{n-1} random sample. The red curve is inside the bundle of blue curves, which shows the validity of the asymptotics in Theorem 4.3.7.

BIBLIOGRAPHY

Ahn J., Marron J.S., Muller K.E. and Chi Y.Y. (2007). The high dimension, low sample size geometric representation holds under mild conditions. *To appear in Biometrika* .

Anderson T.W. (1958). *Introduction to Multivariate Statistical Analysis*. New York: John Wiley and Sons, Inc.

Anderson T.W. (1963). Asymptotic theory of principal component analysis. *Annals of Mathematical Statistics* **34**, 122–148.

Bai Z. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica* **9**, 611–677.

Bai Z. and Yin Y. (1988). Convergence to the semicircle law. *The Annals of Probability* **16**, 863–875.

Bai Z. and Yin Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Statistics* **21**, 1275–1294.

Baik J. and Silverstein J.W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97**, 1382–1408.

Björkström A. and Sundberg R. (1996). Continuum regression is not always continuous. *Journal of the Royal Statistical Society, Series B: Methodological* **58**, 703–710.

Borga M., Landelius T. and Knutsson H. (1997). A unified approach to PCA, PLS, MLR and CCA. Report LiTH-ISY-R-1992, ISY, SE-581 83 Linköping, Sweden.

Duda R., Hart P. and Stork D. (2000). *Pattern classification*. Wiley-Interscience.

EisenLab (2002). http://rana.lbl.gov/eisensoftware.htm.

Frank I.E. and Friedman J.H. (1993). A statistical view of some chemometrics regression tools (Disc: P136-148). *Technometrics* **35**, 109–135.

Geman S. (1980). A limit theorem for the norm of random matrices. *The Annals of Statistics* **8**, 252–261.

Hall P., Marron J.S. and Neeman A. (2005). Geometric representation of high dimension low sample size data. *Journal of the Royal Statistical Society, Series B: Methodological* **67**, 427–444.

Hastie T., Buja A. and Tibshirani R. (1995). Penalized discriminant analysis. *The Annals of Statistics* **23**, 73–102.

Hastie T., Tibshirani R. and Friedman J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer.

Helland I. (1988). On the structure of partial least squares regression. *Communications in Statistics- Simulations and Computation* **17**, 581–607.

Hoerl A.E. and Kennard R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

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 69**, 026124.

Johnstone I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29**, 295–327.

Johnstone I.M. and Lu A.Y. (2004). Sparse principal components analysis. *Technical report, Stanford University, Department of Statistics* .

Laloux L., Cizeau P., Potters M. and Bouchaud J. (2000). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance* **3**, 391–397.

Lee M.H. (2005a). http://www.unc.edu/∼mhlee/toyexample/cancor4dtoyex2.avi.

Lee M.H. (2005b). http://www.unc.edu/∼mhlee/toyexample/p13texex11-aug-2005.

Lee M.H. (2005c). http://www.unc.edu/∼mhlee/toyexample/p14texex11-aug-2005.

Lee M.H. (2005d). http://www.unc.edu/∼mhlee/toyexample/p21texex11-aug-2005.

Marron J.S., Todd M. and Ahn J. (2007). Distance weighted discrimination. *To appear in Journal of the American Statistical Association* .

Marčenko V. and Pastur L. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR - Sbornik* **1**, 457–486.

Massy W.F. (1965). Principal component regression in exploratory statistical research. *Journal of the American Statistical Association* **60**, 234–246.

Muirhead R.J. (1982). *Aspects of multivariate statistical theory*. John Wiley and Sons, Inc.

Naes T. and Martens H. (1985). Comparison of prediction methods for multicollinear data. *Communications in Statistics- Simulations and Computation* **14**, 545–576.

Paul D. (2004). Asymptotics of the leading sample eigenvalues for a spiked covariance model. *Technical Report* .

Phatak A. and Jong S.D. (1997). The geometry of partial least squares. *Journal of Chemometrics* **11**, 311–338.

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* **16**, 356–366.

Saunders G., Gammerman A. and Vovk V. (1998). Ridge regression learning algorithm in dual variables. *In Proceedings of the 15th International Conference on Machine Learning* pp. 515–521.

Shawe-Taylor J. and Cristianini N. (2000). *An Introduction to Support Vector Machines*. Cambridge.

Shawe-Taylor J. and Cristianini N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge.

Silverstein J.W. (1985). The smallest eigenvalue of a large dimensional wishart matrix. *The Annals of Probability* **13**, 1364–1368.

Silverstein J.W. (1989). On the weak limit of the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis* **30**, 321–377.

Stone M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion) (Corr: 76V38 p102). *Journal of the Royal Statistical Society, Series B: Methodological* **36**, 111–147.

Stone M. and Brooks R.J. (1990). Continuum regression: Cross-validated Sequentially Constructed prediction embracing ordinary least squares, partial least squares and principal components regression (Corr: V54 p906-07). *Journal of the Royal Statistical Society, Series B: Methodological* **52**, 237–269.

Stone M. and Brooks R.J. (1994). Joint continuum regression for multiple predictands. *Journal of the American Statistical Association* **89**, 1374–1377.

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Methodological* **58**, 267–288.

Tibshirani R., Hastie T., Narasimhan B. and Chu G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science* **18**, 104–117.

Tracy C.A. and Widom H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics* **177**, 727–754.

Vapnik V. (1982). *Estimation of dependences based on empirical data*. Springer Series in Statistics. Springer-Verlag, New York. Translated from the Russian by Samuel Kotz.

Vapnik V.N. (1995). *The nature of statistical learning theory*. Springer-Verlag, New York.

Wold H. (1976). Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. In: *Perspectives in Probability and Statistics, In Honor of M. S. Bartlett*, pp. 117–144. Academic Press.

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.

Yin Y.Q. (1986). Limiting spectral distribution for a class of random matrices. *Journal of Multivariate Analysis* **20**, 50–68.

Zou H. and Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Methodological* **67**, 301–320.