INTEGRATED ANALYSIS OF MULTIPLE DATA SETS WITH BIOMEDICAL APPLICATIONS

Gen Li

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

Chapel Hill
2015

Approved by:

Andrew B. Nobel

Haipeng Shen

J.S. Marron

Fred A. Wright

Kai Zhang

ABSTRACT

GEN LI: INTEGRATED ANALYSIS OF MULTIPLE DATA SETS
WITH BIOMEDICAL APPLICATIONS
(Under the direction of Andrew B. Nobel and Haipeng Shen)

It is increasingly common to have measurements from multiple platforms on the same set of samples in modern biomedical sciences. In this dissertation, we develop novel methodologies for integrated analysis of multiple data sets. In particular, we devise a supervised principal component analysis framework that achieves dimension reduction of the primary data with guidance from an auxiliary data set. It extracts accurate and interpretable low-rank structures that are potentially driven by the auxiliary information. We further extend the method to accommodate special features of data such as functionality and high dimensionality through regularization. Numerical examples demonstrate that the proposed methodologies have clear advantages over existing methods. In addition, we develop a Bayesian hierarchical model for multi-tissue eQTL analysis. It exploits shared information in multiple tissues to increase the power of eQTL discovery and improve tissue specificity assessment. The method has been adopted by the Genotype-Tissue Expression (GTEx) consortium and successfully applied to the nine-tissue pilot data.

ACKNOWLEDGMENTS

I am very lucky to have Andrew and Haipeng as my advisors during my PhD study in the Department of Statistics and Operations Research at UNC Chapel Hill. I am greatly indebted to them for their patience, dedication, and constructive supervision. Their extraordinary empathy and valuable advice have guided me through many difficult situations and helped me make important decisions. I am also grateful for numerous enlightening and fruitful discussions with Drs. Jianhua Huang, J.S. Marron and Fred Wright. They are definitely role models in my life.

I would also like to thank my colleagues, Patrick Kimes, James Wilson, Xuan Wang, Dan Shen, Eric Lock and many others, for their insightful comments, stimulating discussions, and helpful criticism of my research. In particular, I will miss the time when Patrick and I tackled math problems together using the white board in our office.

I would not have made it without the great company of my friends. I hereby would like to thank: Juana, my beloved girlfriend, for her tremendous support and company; Minghui Liu and Dong Wang for being wonderful roommates; Dongqing Yu and Siying Li for being great neighbors; Yu Zhang and Haojin Zhai for being gym buddies; many others for sharing happiness and sorrow.

Finally, I would like to express my ultimate gratitude to my parents. Without their mental and financial support, I would not have had the opportunity to pursue my dream. They encourage me to follow my interest and stick to what I have chosen. They always cheer me up when I'm down and prevent me from being dizzy when I succeed. I owe my everything to their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Collection of multiple data sets on the same set of samples becomes increasingly common. Many scientific research fields, such as genetics, finance and economics, now involve the analysis of multiple data types. Multiple data sets provide vast opportunities and challenges to statistics.

As an example, The Cancer Genome Atlas Network (2012) (TCGA) aims at understanding how genetic variations interact to drive cancers. The TCGA consortium collected multiple data types such as gene expression data, genotype data, and DNA methylation data from over 500 subjects for each cancer selected for study. Conceptually, different data types are inherently related and may shed light on the mechanism of disease from different perspectives. Jointly analyzing the multiple genetic data types may help us get a more comprehensive understanding. Another example is the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013; The GTEx Consortium, 2015). The primary goal is to create a comprehensive public atlas of gene expression and regulation across multiple human tissues. Noticing the commonality among tissues, one may borrow strength across tissues when studying genetic regulation in one tissue. Moreover, the joint analysis of multiple tissues may expand the scope of single tissue analyses by addressing more fundamental biological questions about the nature and source of variation among tissues.

There are two primary ways of analyzing multiple data types: individual analysis and integrated analysis. Individual analysis focuses on a single data set at a time. It neither accounts for interaction between data sets, nor fully recognize shared information across multiple sets. Therefore, integrated analysis is usually preferable in most cases. However, multiple data types usually have different scales, units, and formats. They cannot be simply concatenated for joint analysis. Most existing statistical methodologies are not specially developed for analyzing multiple data types. The integrated analysis of multiple data sets is an open and promising research area.

In this dissertation, we develop two statistical methodologies for integrated analysis of multiple data sets: one is *supervised dimension reduction* and the other is *multi-tissue eQTL model*. For the first topic, we are interested in dimension reduction of a primary data set, with the presence of an auxiliary data set. The goal is to exploit the auxiliary information to improve the accuracy and interpretability of the reduced primary data. Motivated by many applications, we assume the auxiliary data set potentially drives the underlying structure of the primary data. We develop a latent variable model to account for the supervision effect in dimension reduction. For the second topic, we are interested in studying eQTL, or significant gene-SNP associations, in multiple tissues. The goal is to utilize genetic data across multiple tissues to increase eQTL detection power and improve tissue-specificity assessment. We develop a hierarchical Bayesian model to jointly analyze gene-SNP associations in multiple tissues.

In this chapter, we briefly review some relevant concepts and existing methods. In Section 1.1, we introduce several dimension reduction methods for a single multivariate data set, including Singular Value Decomposition (SVD), Principal Component Analysis (PCA), and factor analysis. In Section 1.2, we extend to methods involving multiple data sets, such as Sufficient Dimension Reduction (SDR), Canonical Correlation Analysis (CCA), and parsimonious multivariate regression. In Section 1.3, we give an overview of the expression quantitative trait loci (eQTL) analysis. In Section 1.4, we briefly introduce the empirical Bayes framework. A summary of our contributions and the outline of the dissertation is given in Section 1.5.

Notation: Throughout the dissertation, without special notification, we use bold capital letters (e.g., $\mathbf{X}$*,* $\mathbf{Y}$*) to denote matrices, bold small letters (e.g.,* $\mathbf{u}$*,* $\mathbf{v}$*) to denote column vectors, and plain letters (e.g.,* λ*,* c*) to denote scalars. For a random data matrix, we assume each row corresponds to a sample and each column corresponds to a variable.*

1.1 Dimension Reduction of Single Dataset

Singular Value Decomposition

SVD is a popular matrix decomposition approach. It factorizes a matrix into a sum of several unit-rank layers. Formally, let $\mathbf{X}$ be an $n \times p$ data matrix of rank k, where $k \leq \min(n, p)$. The SVD of $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{i=1}^{k} d_i \mathbf{u}_i \mathbf{v}_i^T$$

where $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_k)$ is an $n \times k$ matrix of orthonormal left singular vectors, $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_k)$ is a $p \times k$ matrix of orthonormal right singular vectors, and $\mathbf{D} = \text{diag}(d_1, \cdots, d_k)$ is a diagonal matrix with positive singular values $d_1 \geq \cdots \geq d_k > 0$. In particular, $d_i \mathbf{u}_i \mathbf{v}_i^T$ is the ith unit-rank layer with Frobenius norm d_i. If all singular values are distinct, the decomposition is uniquely defined. When some of the d_i's are equal, the column space spanned by the corresponding left (or right) singular vectors is unique, while the specific singular vectors are determined up to an orthogonal rotation.

SVD can be used as a dimension reduction approach. We can obtain a low rank approximation of $\mathbf{X}$ by taking the summation of the first few unit-rank layers. In particular, for any rank $r \leq k$, we have

$$\sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i^T = \underset{\mathbf{C} \in \mathbb{R}_r^{n \times p}}{\arg\min} \|\mathbf{X} - \mathbf{C}\|_{\mathbb{F}} = \underset{\mathbf{C} \in \mathbb{R}_r^{n \times p}}{\arg\min} \text{tr}\{(\mathbf{X} - \mathbf{C})(\mathbf{X} - \mathbf{C})^T\}$$

where $\mathbb{R}_r^{n \times p}$ is the set of all $n \times p$ matrices with rank r, and $\|\cdot\|_{\mathbb{F}}$ represents the Frobenius norm. In the sense, SVD provides the best low rank approximation of a matrix in terms of minimizing the Frobenius norm of the difference between the low rank matrix and the original one.

Principal Component Analysis

PCA is one of the most widely used dimension reduction techniques in multivariate statistics. It seeks directions that are mutually orthogonal and sequentially maximize variations of data. Mathematically, assume $\mathbf{x}$ is a length-p random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The first principal component loading vector is the solution of the following

criterion:

$$\underset{\{\mathbf{v}_1 \in \mathbb{R}^p : \|\mathbf{v}_1\|_2 = 1\}}{\arg\max} \quad \mathrm{var}(\mathbf{v}_1^T \mathbf{x}),$$

and subsequent loading vectors ($k = 2, \cdots, p$) are obtained by solving

$$\underset{\{\mathbf{v}_k \in \mathbb{R}^p : \|\mathbf{v}_k\|_2 = 1, \mathbf{v}_k^T \mathbf{v}_j = 0, j = 1, \cdots, k-1\}}{\arg\max} \quad \mathrm{var}(\mathbf{v}_k^T \mathbf{x}).$$

It is easy to see that the PC loadings are the eigenvectors of $\mathbf{\Sigma}$ on population level.

When the true $\mathbf{\Sigma}$ is unknown, PC loadings can be estimated from a sample covariance matrix. This is closely related to the SVD method. Let $\mathbf{X}$ denote an $n \times p$ data matrix where each row is an independent identically distributed (i.i.d.) sample with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$. Without loss of generality, we assume $\mathbf{X}$ has been column centered. PCA can be computed by the SVD of $\mathbf{X}$. In particular, if we write the SVD of $\mathbf{X}$ as $\mathbf{UDV}^T$ where $\mathbf{U}$ is the left singular matrix $\mathbf{V}$ is the right singular matrix and $\mathbf{D}$ is the diagonal singular value matrix, the columns of $\mathbf{UD}$ correspond to PC scores and the columns of $\mathbf{V}$ correspond to PC loadings. Alternatively, one may also obtain PC loadings by decomposing $\mathbf{X}^T\mathbf{X}$, and obtain PC scores by decomposing $\mathbf{XX}^T$.

Recently, many variants of PCA have been investigated and adapted for different applications. For example, Shen and Huang (2008b) proposed a sparse PCA method by imposing ℓ_1 penalty on loading vectors. The method achieves dimension reduction and variable selection simultaneously. Also see Zou et al. (2006) and Yang et al. (2014). Witten et al. (2009) and Lee et al. (2010) generalized the idea by imposing penalties on both loading and score vectors to get two-way sparse PCA methods. In functional data analysis, each sample is an observation from some underlying smooth function. A variety of functional extensions of PCA have been studied (cf. Huang et al., 2008; Silverman, 1996). As an analog of the two-way sparse PCA, Huang et al. (2009) proposed a two-way functional PCA method. Zhang et al. (2013) further modified the method to be robust against outliers. Very recently, Allen (2013) developed a general framework of two-way sparse and functional PCA that unifies many existing

regularized PCA methods.

Factor Analysis

Factor analysis describes variability of correlated variables in terms of a small number of latent factors. Formally, the factor model for a length-p random vector $\mathbf{x}$ is

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{f} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\mu}$ is a mean vector, $\mathbf{f}$ is an length-r random vector ($r < p$) with entries being uncorrelated latent factors, $\mathbf{L}$ is a $p \times r$ loading matrix, and $\boldsymbol{\varepsilon}$ is a $p \times 1$ error vector. We assume $\mathbf{f}$ has mean zero and covariance $\mathbf{I}$, and is independent of $\boldsymbol{\varepsilon}$ which has mean zero and covariance $\boldsymbol{\Sigma}$. Hence the covariance of $\mathbf{x}$ is $\mathbf{L}\mathbf{L}^T + \boldsymbol{\Sigma}$. One may impose different covariance structures on $\boldsymbol{\Sigma}$ to form different factor models.

To estimate $\boldsymbol{\mu}$, $\mathbf{L}$, and $\boldsymbol{\Sigma}$ in a factor model, people usually use the *Expectation-Maximization* (EM) algorithm. It is an iterative algorithm alternating between an E step and an M step. In the E step, we calculate the conditional distribution of the latent factor $\mathbf{f}$ given the observed data $\mathbf{x}$ and parameters estimated from the previous iteration; in the M step, we maximize the conditional expectation of the joint likelihood of $\mathbf{x}$ and $\mathbf{f}$ with respect to model parameters.

Factor analysis is closely related to PCA, but they are not the same. By definition, the PC loadings of $\mathbf{x}$ are the eigenvectors of $\mathbf{L}\mathbf{L}^T + \boldsymbol{\Sigma}$, which are not necessarily the columns of $\mathbf{L}$. However, when $\boldsymbol{\Sigma}$ is isotropic (i.e., $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$) and $\mathbf{L}$ has orthogonal columns, it is easy to see the first r eigenvectors of $\mathbf{L}\mathbf{L}^T + \boldsymbol{\Sigma}$ are proportional to the columns of $\mathbf{L}$. In fact, when $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, the factor model is the probabilistic PCA model proposed by Tipping and Bishop (1999).

1.2 Dimension Reduction with Multiple Datasets

Sufficient Dimension Reduction

Assume we have a univariate response y and a multivariate predictor vector $\mathbf{x} \in \mathbb{R}^p$. The goal is to predict y based on $\mathbf{x}$. When p is large, we may want to reduce the dimensionality of $\mathbf{x}$ to facilitate visualization and interpretation. The idea of SDR is to get a reduced version

of $\mathbf{x}$, denoted by $R(\mathbf{x})$, that lies in a lower dimensional subspace $\mathbb{R}^r$ $(r < p)$ and contains all relevant information about y in $\mathbf{x}$. Formally, a reduction $R(\cdot) : \mathbb{R}^p \mapsto \mathbb{R}^r$ is sufficient if it satisfies one the following conditions:

(1) $\mathbf{x}|(y, R(\mathbf{x})) \sim \mathbf{x}|R(\mathbf{x})$,

(2) $y|\mathbf{x} \sim y|R(\mathbf{x})$,

(3) $\mathbf{x} \perp\!\!\!\perp y|R(\mathbf{x})$,

where $\sim$ means identically distributed and $\perp\!\!\!\perp$ means independent. These conditions are equivalent when $(y, \mathbf{x})$ has a joint distribution. Different conditions may be useful in different situations. For example, when the response is assumed fixed, only condition (1) makes sense since it does not require y to be random; for fixed design where $\mathbf{x}$ is assumed fixed, only condition (2) is meaningful.

The sufficient reduction $R(\cdot)$ can be of any form. In the extreme case where no reduction is available, $R(\cdot)$ is a one-to-one mapping of $\mathbf{x}$. For simplicity, people usually assume that $R(\cdot)$ is a collection of linear transformations: $R(\mathbf{x}) = (\boldsymbol{\beta}_1^T\mathbf{x}, \cdots, \boldsymbol{\beta}_r^T\mathbf{x})^T = \mathbf{B}^T\mathbf{x}$, where $\mathbf{B} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_r)$ is a $p \times r$ coefficient matrix. The column space of $\mathbf{B}$ is called a *dimension reduction subspace*, denoted by $\mathcal{S}_{\mathbf{B}}$. Since any superspace of $\mathcal{S}_{\mathbf{B}}$ also contains all relevant information about y in $\mathbf{x}$, the dimension reduction subspace is not unique. Under mild conditions, Cook (1996) shows that the intersection of two dimension reduction subspaces is still a dimension reduction subspace. Consequently, the inferential target in sufficient dimension reduction is often taken to be the intersection of all dimension reduction subspaces which is uniquely defined. It is called the *central subspace*, denoted by $\mathcal{S}_{y|\mathbf{x}}$.

There has been a lot of efforts devoted to the estimation of central subspace. In particular, there are two major lines: moment-based methods and likelihood-based methods. Li (1991) proposed a sliced inverse regression (SIR) approach that exploits the first moment of $\mathbf{x}$ given y to derive the central subspace. In the discussion of the same paper, Cook and Weisberg proposed a sliced average variance estimation (SAVE) method to achieve the same goal using the first and second moments. Both methods provide consistent estimators of $\mathcal{S}_{y|\mathbf{x}}$ under standard conditions. Later, several methods have been proposed to combine the strength of

SIR and SAVE and move beyond (see Cook and Ni, 2005; Ye and Weiss, 2003; Yin and Cook, 2003, for example). Cook and Forzani (2009) proposed a maximum likelihood estimator of the central subspace based on Gaussian assumptions. They demonstrated the method outperforms moment-based methods and is robust against deviations from normality.

Parsimonious Multivariate Regression

In multivariate regression problems, the response has multiple variables. In particular, assuming $\mathbf{y} = (y_1, \cdots, y_q)^T \in \mathbb{R}^q$ and $\mathbf{x} = (x_1, \cdots, x_p)^T \in \mathbb{R}^p$, a multivariate regression has the following form

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{B}^T\mathbf{x} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\mu}$ is a $q \times 1$ intercept vector, $\mathbf{B}$ is a $p \times q$ coefficient matrix, and $\boldsymbol{\varepsilon}$ is an error vector with mean zero and covariance $\boldsymbol{\Sigma}$. Without loss of generality, in the context we always assume that all variables are centered beforehand so we get rid of the intercept term. When multiple observations are available, the model can be written in the matrix form as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

where $\mathbf{Y}$ is an $n \times q$ response matrix, $\mathbf{X}$ is an $n \times p$ design matrix, and each row of $\mathbf{E}$ is i.i.d. with mean zero and covariance $\boldsymbol{\Sigma}$. In particular, we assume columns of $\mathbf{X}$ are linearly independent to avoid indeterminacy.

Ordinary least square (OLS) is one of the most popular approaches to estimate the regression coefficient matrix $\mathbf{B}$. It minimizes the following criterion

$$\widehat{\mathbf{B}}_{\text{OLS}} = \arg\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{XB}\|_{\mathbb{F}}^2 = \arg\min_{\mathbf{B}} \text{tr}\{(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\}$$

where $\|\cdot\|_{\mathbb{F}}$ is the Frobenius norm. The above problem has a unique closed-form solution as

$$\widehat{\mathbf{B}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

when $\mathbf{X}$ has full column rank. OLS is equivalent to maximum likelihood estimate (MLE) when the random noise has a multivariate Gaussian distribution. Under Gaussian assumption, the log likelihood of the observed data is proportional to $-\text{tr}\{(\mathbf{Y}-\mathbf{XB})\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{XB})^T\}$. When $\mathbf{\Sigma}$ is isotropic, i.e., $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, the MLE has the same object function as OLS; when $\mathbf{\Sigma}$ is any positive definite matrix, using elementary matrix calculus calculations we can derive the closed form solution of MLE to be the same as that of OLS as well. Namely, $\widehat{\mathbf{B}}_{\text{MLE}} \equiv \widehat{\mathbf{B}}_{\text{OLS}}$ under normality.

In high dimension, overfitting issue may arise in estimation. Two primary solutions have been extensively studied: one is *feature selection*, and the other is *feature extraction*. The idea of feature selection is to use a subset instead of all of the p variables to construct the model. The idea of feature extraction is to transform the data from the p-dimensional space to a low-dimensional subspace. Both are achieved by adding structural constraints on the coefficient matrix $\mathbf{B}$. In particular, feature selection can be achieved by imposing sparsity on $\mathbf{B}$. Unimportant variables are removed from the analysis by the device of zero coefficients in $\mathbf{B}$. Various sparse multivariate linear regression methods have been studied in literature (cf. Turlach et al. (2005), Yuan et al. (2007), Lee and Liu (2012), Rothman et al. (2010)).

To realize feature extraction, one general approach is to impose a rank constraint on the coefficient matrix. Consider the multivariate regression model with the constraint $\text{rank}(\mathbf{B}) = r$ where r is a prespecified number much smaller than $\min(p, q, n)$. We write the QR decomposition of $\mathbf{B}$ as $\mathbf{QR}^T$ where $\mathbf{Q}$ is a $q \times r$ matrix, $\mathbf{R}$ is a $p \times r$ matrix. As a result, the constrained regression model can be written as

$$\mathbf{Y} = \mathbf{XQR}^T + \mathbf{E}.$$

$\mathbf{XQ}$ is a linear transformation of the data from the original p-dimensional space to a r-dimensional subspace. The above model is commonly referred to as the reduced rank regression (RRR) model (cf. Izenman (1975) and Reinsel and Velu (1998)). RRR reduces the number of parameters in the model and takes advantage of the interrelation of multiple variables. Therefore, the interpretation and the prediction are both enhanced. The following

OLS criterion is commonly used to estimate $\mathbf{B}$ in RRR

$$\widehat{\mathbf{B}}_{\text{RRR}} = \underset{\text{rank}(\mathbf{B})=r}{\arg\min} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_{\mathbb{F}}^2.$$

Again, this is equivalent to MLE when the random error is Gaussian with isotropic covariance structure. The explicit solution is given in Reinsel and Velu (1998) as

$$\widehat{\mathbf{B}}_{\text{RRR}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{H}\mathbf{H}^T$$

where $\mathbf{H} = (\mathbf{h}_1, \cdots, \mathbf{h}_r)$ and $\mathbf{h}_i$ is the ith eigenvector of $\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. When the random error is Gaussian but with arbitrary positive definite covariance structure, the OLS is not equivalent to the MLE anymore. The closed form MLE solution is given in Izenman (1975). Recently, Chen et al. (2012) and Chen and Huang (2012) combined feature selection and feature extraction by imposing sparsity on the low rank coefficient matrix estimation. Numerical studies show the proposed sparse RRR methods have more appealing performances in many situations.

Envelope model (Cook et al., 2010) is another parsimonious variation of multivariate regression. The motivation comes from the observation that some variations in response might be unrelated with the predictor. In that case, by separating the material variation from the immaterial one and only focusing on the former, one expects to reduce the coefficient estimation variability. In particular, the coordinate version of the envelope model is

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{\Lambda}\mathbf{\Gamma}^T + \mathbf{E} \\ \mathbf{\Sigma} &= \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T \end{aligned}$$

where $\mathbf{\Lambda}\mathbf{\Gamma}^T$ is the low rank coefficient matrix, $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ is a $q \times q$ orthogonal matrix, $\mathbf{\Sigma}$ is the covariance structure for each i.i.d. row of $\mathbf{E}$, and $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ are two positive definite matrices. $\mathbf{Y}\mathbf{\Gamma}$ envelops the material variation in the response, and $\mathbf{Y}\mathbf{\Gamma}_0$ envelops the immaterial variation. The envelope model can be viewed as a special case of the RRR model with the specially set covariance structure. Since proposed, different variations of the envelope model

have been studied in Su and Cook (2011), Su and Cook (2012), Su and Cook (2013), Cook and Su (2013) and Cook et al. (2013).

Canonical Correlation Analysis

Unlike regression models, CCA focuses on examining correlation structures between two multivariate random vectors. It treats both random vectors equally, without assuming one to be response and the other to be predictor. The idea of CCA is to find linear combinations such that the correlations between the two are sequentially maximized. Assume we have multivariate vectors $\mathbf{x} = (x_1, \cdots, x_p)^T \in \mathbb{R}^p$ and $\mathbf{y} = (y_1, \cdots, y_q)^T \in \mathbb{R}^q$. The first pair of canonical loadings $(\mathbf{u}_1, \mathbf{v}_1)$ is the solution of the following optimization problem

$$\max_{\mathbf{u}\in\mathbb{R}^p, \mathbf{v}\in\mathbb{R}^q} \text{corr}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}).$$

For identifiability purpose, we need to require both loading vectors have norm 1. The subsequent pairs of loadings are defined in a similar way under the orthogonal constraint that $\mathbf{u}_i^T\mathbf{u}_j = \mathbf{v}_i^T\mathbf{v}_j = \delta_{ij}$ where δ_{ij} is the kronecker delta.

Let $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_y}$ denote the covariance matrices of $\mathbf{x}$ and $\mathbf{y}$ respectively, and let $\mathbf{\Sigma_{xy}} = \text{cov}(\mathbf{x}, \mathbf{y})$ and $\mathbf{\Sigma_{yx}} = \text{cov}(\mathbf{y}, \mathbf{x})$. With some algebraic calculations, we know the loading vectors $\mathbf{u}_i$'s are the eigenvectors of $\mathbf{\Sigma_x}^{-1}\mathbf{\Sigma_{xy}}\mathbf{\Sigma_y}^{-1}\mathbf{\Sigma_{yx}}$ and $\mathbf{v}_i$'s are the eigenvectors of $\mathbf{\Sigma_y}^{-1}\mathbf{\Sigma_{yx}}\mathbf{\Sigma_x}^{-1}\mathbf{\Sigma_{xy}}$. Correspondingly, $\mathbf{u}_i^T\mathbf{x}$ and $\mathbf{v}_i^T\mathbf{y}$ are called the ith canonical scores. When the true covariance matrices are unknown, we can replace them with sample covariance matrices. Let $\mathbf{X}$ and $\mathbf{Y}$ be data matrices with n i.i.d. samples. We have $\widehat{\mathbf{\Sigma_x}} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, $\widehat{\mathbf{\Sigma_y}} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}$, $\widehat{\mathbf{\Sigma_{xy}}} = \frac{1}{n}\mathbf{X}^T\mathbf{Y}$ and $\widehat{\mathbf{\Sigma_{yx}}} = \frac{1}{n}\mathbf{Y}^T\mathbf{X}$. The canonical loadings are estimated accordingly through eigendecomposition of the product of sample covariance matrices.

However, when data are of high dimension, i.e., $p > n$ or $q > n$, corresponding sample covariance matrices may be invertible. In fact, in high dimension, there always exists infinite number of pairs with correlation 1. Namely, we encounter the overfitting problem. Witten et al. (2009) proposed a penalized approach for CCA that achieves sparse estimation of the loading vectors. With proper regularization, the coefficients for unimportant variables are set to 0 to avoid overfitting. In practice, the sparse CCA approach provides highly interpretable

results.

1.3 Expression Quantitative Trait Loci Analysis

Genetic variation in a population is commonly studied through the analysis of SNPs, which are variants occurring at specific sites in the genome. Differences among these variants drive primary phenotypic differences between members of the population. For humans these differences range from physical characteristics to disease susceptibility. Mediating the connection between genetic variation and resulting phenotypes are the effects of SNPs on the expression of different genes. The analysis of eQTL seeks to identify genetic variants that affect the expression of one or more genes: a gene-SNP pair for which the expression of the gene is associated with the value of the SNP is referred to as an eQTL. Enabled by high-throughput sequencing, eQTL analysis has proven to be an effective approach for the discovery of genomic variants that influence expression, and a potentially useful tool in the study of pathways and networks that underlie disease in human and other populations. For an overview of eQTL analysis and disease mapping, see Cookson et al. (2009), Mackay et al. (2009), Rockman and Kruglyak (2006), and the references therein. Kendziorski and Wang (2006) and Wright et al. (2012) survey existing statistical and computational methods for eQTL analysis, respectively.

To date, most eQTL studies have considered the effects of genetic variation on expression within a single tissue. Nonetheless, these studies have provided enhanced understanding of gene regulation and the etiology of various diseases, *cf.* Franke and Jansen (2009) and Westra et al. (2013). A natural next step in understanding genomic variation of expression is the simultaneous analysis of eQTLs in multiple tissues. Multi-tissue eQTL analysis has the potential to improve the findings of single tissue analyses by borrowing strength across tissues, and to expand the scope of single tissue analyses by addressing more fundamental biological questions about the nature and source of variation between tissues.

In a single tissue eQTL study, the goal is to identify gene-SNP pairs for which the expression of the gene is associated with the SNP genotype. An important feature of multiple tissue

studies is that a SNP may be associated with the expression of a gene in some tissues, but not in others. Thus a full multi-tissue analysis must identify complex patterns of association across multiple tissues. We will refer to an eQTL as 'common' if association is present in all available tissues, and 'tissue-specific' if association is present in at least one tissue, but not all. Until recently, understanding of multi-tissue eQTL relationships was limited by a shortage of true multi-tissue data sets, requiring the assimilation of data or results from different studies (one for each tissue) involving distinct populations, measurement platforms, and analysis protocols, *cf.* Emilsson et al. (2008) and Xia et al. (2012).

Recently, a number of human true multi-tissue eQTL data sets have been collected, for example by Dimas et al. (2009) and Nica et al. (2011), although these contain relatively few tissues. By contrast, the GTEx initiative (Lonsdale et al. (2013)) and related projects are generating eQTL data from dozens of tissues in several hundred individuals, greatly expanding our potential understanding of the variation and specificity of eQTL effects across multiple tissues. The size and complexity of these emerging multi-tissue data sets has created the need to expand existing statistical tools for eQTL analysis.

1.4 Empirical Bayes

Empirical Bayes methods are statistical inference procedures that combine Bayesian models with Frequentist estimation procedures. In Bayesian hierarchical models, parameters of interest are treated as random variables with prior distributions in which the parameters are called the hyperparameters. A typical Bayes approach would either integrate out the hyperparameters or set them to be values based on some subjective prior knowledge. For empirical Bayes methods, the hyperparameters are estimated from observed data through marginal maximum likelihood which is a typical Frequentist approach. Conceptually, empirical Bayes approaches fully utilize the information in the observed data.

We use a simple example to illustrate the idea of empirical Bayes methods. Suppose $\boldsymbol{\theta}$ is a unknown parameter vector of length p, and $\mathbf{x}$ is an observation vector of the same length,

such that the observations are normally distributed as

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I}).$$

We are interested in estimating the parameter vector $\boldsymbol{\theta}$ using the single observation vector $\mathbf{x}$. For simplicity, we assume σ^2 is known. In Frequentist inference, the least square estimate of $\boldsymbol{\theta}$ is just $\mathbf{x}$.

In Bayesian framework, the parameter vector $\boldsymbol{\theta}$ is assumed random, and one can impose a flexible prior distribution on $\boldsymbol{\theta}$ based on prior knowledge. In particular, we set the prior distribution of $\boldsymbol{\theta}$ to be $\mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I})$ where τ^2 is an unknown hyperparameter. The posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{x}$ is

$$\boldsymbol{\theta}|\mathbf{x} \sim \mathcal{N}\left(\frac{\tau^2}{\tau^2 + \sigma^2}\mathbf{x}, \frac{\tau^2\sigma^2}{\tau^2 + \sigma^2}\right).$$

Consequently, for any fixed τ^2, the Bayes estimate of $\boldsymbol{\theta}$ is $[\tau^2/(\tau^2 + \sigma^2)]\mathbf{x}$. Notice the Bayes estimate depends on the subjective choice of the hyperparameter τ^2.

The empirical Bayes approach takes advantage of the flexible Bayesian model while estimating the hyperparameter from the data. In particular, from the marginal distribution of $\mathbf{x}$ we know

$$\mathbb{E}\left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{x}\|_2^2}\right) = \frac{\tau^2}{\tau^2 + \sigma^2}$$

for any $p > 2$. We can substitute this into the Bayes estimate and get the empirical Bayes estimate as $\left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{x}\|_2^2}\right)\mathbf{x}$. This estimate has the same form with the James-Stein estimate (Stein, 1956). It has been proved (Efron and Morris, 1973) that the empirical Bayes estimate strictly outperforms the Frequentist estimate when $p \geq 3$ in terms of the mean square error for any true value.

The above idea can be easily extended to more general frameworks. Beginning with the work of Newton et al. (2001) and Efron et al. (2001), empirical Bayes approaches have been applied to hierarchical models in a number of genetic applications, most notably the study

of differential expression and co-expression in gene microarrays, cf. Kendziorski et al. (2003), Newton et al. (2004), Smyth (2004), Efron (2008), and Dawson and Kendziorski (2012).

1.5 New Contributions and Outline

In this dissertation, we focus on two topics of the integrated analysis of multiple data sets, i.e., supervised dimension reduction and multi-tissue eQTL analysis. We shall illustrate that in both studies integrated analysis outperforms separate analysis by borrowing strength across data sets. Briefly, the remainder of the dissertation is organized as follows:

In Chapter 2, we develop a supervised PCA framework that extends standard PCA to incorporate auxiliary data (Li et al., 2015). The auxiliary information that potentially drives the underlying structure of the primary data of interest is referred to as *supervision*. The goal is to obtain a more interpretable and accurate low-rank approximation of the primary data with the help of supervision. It is different from the scope of SDR or parsimonious multivariate regression which seeks a reduced version of the primary data that keeps all information about the supervision. We treat the auxiliary data as covariates for the intrinsic structure of the primary data rather than response. An appealing feature of the method is that it learns the amount of supervision needed in dimension reduction adaptively, and reduces to the standard PCA method when the auxiliary data are actually irrelevant to the low rank structure. We apply the supervised PCA to a gene expression data set of breast cancer tumors where disease subtypes are treated as supervision. By incorporating the auxiliary subtype information in dimension reduction, we obtain a low-rank structure consisting of clear patterns driven by subtypes and patterns from within-subtype variations. We also consider an arrival rate data set from a call center where incorporating the day-of-week index as supervision in dimension reduction reveals interpretable arrival patterns and increases call volume forecasting accuracy.

In Chapter 3, we extend the supervised PCA framework to incorporate regularization to better accommodate high dimensional data and functional data (Li et al., 2014b). Smoothness and sparsity constraints are imposed on loading vectors to reduce variability and enhance interpretability of estimation. In addition, we also impose sparsity on supervision coefficients

to identify auxiliary variables with no supervision effect. The resulting methodology subsumes the original supervised PCA method, as well as existing regularized PCA methods, such as functional PCA and sparse PCA as special cases. Numerical studies show the proposed method outperforms competitive approaches in terms of low-rank structure recovery accuracy in a wide range of settings. In an application example concerning yeast cell cycle-related genes, the supervised regularized PCA method takes advantage of auxiliary transcription factors binding information and captures underlying cyclic patterns of gene expressions in two cell cycles. Moreover, it simultaneously identifies important transcription factors that regulate cell cycles, which is not achieved by other dimension reduction methods.

In Chapter 4, we study an empirical Bayes approach for joint eQTL analysis in multiple tissues (Li et al., 2014a). We build a Multi-Tissue eQTL model (MT-eQTL) that captures the presence or absence of an eQTL and accounts for the heterogeneity of effect size variations in multiple tissues simultaneously. The model can flexibly identify whether a gene and a SNP are significantly associated in all tissues, or a subset of tissues, or no tissues. As genetic data are often of several gigabytes, fast computation is an extremely desirable feature. The proposed method employs an empirical Bayes approach for model estimation and inferences where the computational speed is tenfold faster than standard permutation-based methods, making it preferable in practice. In collaboration with experts in the GTEx consortium, we apply the method to a 9-tissue data set from the pilot project. We show that jointly analyzing data from multiple tissues increases the statistical power of eQTL detection and improves the tissue-specificity assessment.

CHAPTER 2: SUPERVISED SINGULAR VALUE DECOMPOSITION AND ITS ASYMPTOTIC PROPERTIES

2.1 Introduction

As high dimensional data become increasingly common, dimension reduction becomes more and more important, since it is easier to visualize and analyze a low dimensional structure in high dimensional data. SVD is a fundamental tool used in multivariate analysis to decompose a high-dimensional data matrix into a sum of unit-rank layers ordered by importance. The first few layers, which often capture the majority of the variation, act as a low rank approximation or dimension reduction of the original data.

However, one drawback of SVD is that it only makes use of a single data set, and by default the resulting dimension reduction cannot incorporate any additional information that may be relevant. When multiple related data sets are available on the same set of samples, sharing information across data sets may lead to recovery of a low rank structure that is more interpretable. Several approaches have been developed for analyzing multiple data sets. For example, Lock et al. (2013) develops an integrative approach to study joint and individual variations simultaneously; Bair et al. (2006) develops a supervised principal component regression method to select predictors and do prediction. In this chapter, we propose a supervised SVD (SupSVD) model to achieve dimension reduction that incorporates auxiliary information. We assume that the auxiliary data set, which we refer to as the *supervision*, is a potential driving factor for the low rank structure of the *primary* data of interest.

The assumption is reasonable in many applications. For example, some genetic studies collect both gene expression and single-nucleotide polymorphism (SNP) data on the same group of subjects. One interesting topic is to investigate intrinsic patterns of the expression data. Biologically, expression of some genes is regulated by SNPs known as eQTL. In other words, SNPs indeed drive underlying structure in the gene expression data which one can

potentially get a better understanding of if we take advantage of the supervision (SNP) data.

We now introduce the SupSVD model using matrix notation. Let $\mathbf{X}$ denote the data matrix of primary interest which has n rows (or samples) and p columns (or variables). Let $\mathbf{Y}$ denote the supervision data matrix which has n rows (matched with $\mathbf{X}$) and q columns. We assume that the intrinsic information in $\mathbf{X}$ is low dimensional with rank r ($r \leq \min(n, p)$), and is possibly driven by $\mathbf{Y}$, in a linear fashion. In matrix form, the SupSVD model can be expressed as follows:

$$\begin{cases} \mathbf{X} = \mathbf{U}\mathbf{V}^T + \mathbf{E}, \\ \mathbf{U} = \mathbf{Y}\mathbf{B} + \mathbf{F}, \end{cases} \tag{2.1}$$

where $\mathbf{U}$ is an $n \times r$ latent score matrix, $\mathbf{V}$ is a $p \times r$ full-rank loading matrix, and $\mathbf{B}$ is a $q \times r$ coefficient matrix, with $\mathbf{F}$ and $\mathbf{E}$ being $n \times r$ and $n \times p$ error matrices, respectively.

Overall, the SupSVD model captures situations in which $\mathbf{X}$ has an intrinsic low rank structure and the structure is partially affected by $\mathbf{Y}$. The first equation in (2.1) is motivated by the additive-multiplicative low-rank approximation model for SVD, as in Dozier and Silverstein (2007) and Shabalin and Nobel (2013). It indicates that the observed data matrix $\mathbf{X}$ consists of the low rank structure $\mathbf{U}\mathbf{V}^T$ plus measurement errors $\mathbf{E}$. We use a multivariate linear regression model to capture the potential supervising effect of $\mathbf{Y}$ on the score matrix $\mathbf{U}$. In particular, the matrix $\mathbf{F}$ captures information in $\mathbf{U}$ that cannot be explained by $\mathbf{Y}$. We note that very recently Fan et al. (2014) proposed a projected PCA method that generalizes the second equation of (2.1) to a semi-parametric model.

The SupSVD model is related to the latent variable model in Bair et al. (2006) and the surrogate variable model in Leek and Storey (2007). The latent variable model utilizes the same low-rank additive model of $\mathbf{X}$ as in the first formula of (2.1). However, it differs from the SupSVD model in the second formula: it assumes the latent variable $\mathbf{U}$ drives $\mathbf{Y}$, rather than the opposite. In other words, $\mathbf{Y}$ is regressed on $\mathbf{U}$. In the surrogate variable model, it is assumed that the auxiliary data affect each variable of the primary data directly instead of through its low-rank structure. The primary data $\mathbf{X}$ is regressed on $\mathbf{Y}$ with a structured

error term, consisting of a few latent factors and noise. The name of the model comes from the fact that the latent factors are modeled with surrogate variables. Both models are related to but different from the SupSVD model we propose here.

Compared with the SVD, the SupSVD model incorporates the auxiliary information in $\mathbf{Y}$. The potential advantages of SupSVD over SVD are two-fold. First, using additional information may help reveal interesting patterns that might otherwise be undiscovered. Second, the low rank structure recovered by the SupSVD model might have superior interpretability. Evidence can be found in the simulated examples in the appendix, Section 2.7.6. Overall we find that SupSVD performs favorably when the supervision information is indeed a driving factor of low rank data. When auxiliary data are irrelevant, for example in Case 2 of Section 2.5.1.1, SupSVD automatically adapts to the situation and performs as well as SVD.

There is a rich literature on dimension reduction of a data matrix $\mathbf{X}$ in the presence of auxiliary information $\mathbf{Y}$, for example sufficient dimension reduction Cook and Ni (2005), supervised principal components Bair et al. (2006), and principal fitted components Cook (2007); Cook and Forzani (2008). Moreover, reduced rank regression (RRR) Izenman (1975); Reinsel and Velu (1998) can also be viewed as a dimension reduction approach for $\mathbf{X}$ if we regress $\mathbf{X}$ on $\mathbf{Y}$. The focus of most existing methods is to find a dimension reduced version of $\mathbf{X}$ that keeps all the information about $\mathbf{Y}$. This is different from the scope of the current paper. Here our primary goal is to identify low rank structure of $\mathbf{X}$, whether or not the structure is related to the auxiliary information $\mathbf{Y}$. The auxiliary information $\mathbf{Y}$ offers guidance for the dimension reduction of $\mathbf{X}$. To the best of our knowledge, our work is the first to address this topic.

The rest of the chapter is organized as follows. In Section 2.2, we give more details of the SupSVD model, and explain its connections with existing models. In Section 2.3, we propose a modified version of the EM algorithm for parameter estimation. The asymptotic properties of the estimates are discussed in Section 2.4. In Section 2.5, we compare different methods using extensive simulations and apply SupSVD to a real data example. We conclude in Section 2.6, with a brief discussion of potential extensions of our framework to functional data analysis. Proofs, technical details, and additional numerical examples can be found in

the appendix, Section 2.7.

2.2 The SupSVD Model

In this section, we describe the SupSVD method in detail. Section 2.2.1 gives an equivalent formulation of the model, and discusses identifiability conditions. Section 2.2.2 establishes connections of the proposed model with some existing methods.

2.2.1 An Equivalent Form of The Model

In Model (2.1), if we substitute the latent matrix $\mathbf{U}$ in the first equation with the second equation, we get an equivalent form for the SupSVD model as:

$$\mathbf{X} = \mathbf{Y}\mathbf{B}\mathbf{V}^T + \mathbf{F}\mathbf{V}^T + \mathbf{E}. \tag{2.2}$$

Without loss of generality, we assume that both $\mathbf{X}$ and $\mathbf{Y}$ are column-centered; hence, the model does not have intercepts. The random matrices $\mathbf{E}$ and $\mathbf{F}$ are assumed independent. Each entry of the error matrix $\mathbf{E}$ is independently identically distributed (i.i.d.) with mean zero and variance $\sigma_{\mathbf{e}}^2$. This follows the signal-plus-noise model for matrix reconstruction, cf. Shabalin and Nobel (2013), as well as the r-component spiked covariance model for PCA, cf. Johnstone (2001); Paul (2007). Each row of $\mathbf{F}$ is i.i.d. with mean zero and covariance matrix $\mathbf{\Sigma_f}$, which is an unknown $r \times r$ positive definite matrix.

Furthermore, Model (2.2) can be viewed as a special setup of a multivariate linear regression model

$$\mathbf{X} = \mathbf{Y}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where the coefficient matrix $\boldsymbol{\beta}$ is $\mathbf{B}\mathbf{V}^T$ of rank $\min(r, q)$, and the random noise matrix $\boldsymbol{\varepsilon}$ is $\mathbf{F}\mathbf{V}^T + \mathbf{E}$. The rows of the noise matrix $\boldsymbol{\varepsilon}$ are i.i.d. with covariance $\mathbf{\Sigma}$ equal to $\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p$ where $\mathbf{I}_p$ is the $p \times p$ identity matrix.

The primary goal of the SupSVD model is to identify low rank structure in the observed

data $\mathbf{X}$ using $\mathbf{Y}$. Namely, we want to estimate $\mathbf{YBV}^T + \mathbf{FV}^T$, where $\mathbf{YBV}^T$ is the deterministic part and $\mathbf{FV}^T$ is the random part. The deterministic signal is driven by $\mathbf{Y}$ and the random signal captures important structures from unknown sources. The two parts are related through the common loading matrix $\mathbf{V}$, and together they form the underlying low rank representation for $\mathbf{X}$. In practice, we substitute all model parameters by estimates obtained from the observed data, and replace the random matrix $\mathbf{F}$ by its best unbiased prediction.

The SupSVD model (2.2) is identifiable in terms of the coefficient matrix $\boldsymbol{\beta} = \mathbf{BV}^T$ and the covariance matrix $\boldsymbol{\Sigma} = \mathbf{FV}^T + \mathbf{E}$, but unidentifiable in terms of the specific parameters $\mathbf{B}$, $\mathbf{V}$, $\boldsymbol{\Sigma}_\mathbf{f}$, and $\sigma^2_\mathbf{e}$. To see this, let $\mathbf{B}^\star = \mathbf{BQ}$, $\mathbf{V}^\star = \mathbf{VQ}$, and $\boldsymbol{\Sigma}^\star_\mathbf{f} = \mathbf{Q}^T\boldsymbol{\Sigma}_\mathbf{f}\mathbf{Q}$ for any $r \times r$ orthogonal matrix $\mathbf{Q}$. It is easily seen that $\mathbf{BV}^T = \mathbf{B}^\star\mathbf{V}^{\star T}$ and $\mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T = \mathbf{V}^\star\boldsymbol{\Sigma}^\star_\mathbf{f}\mathbf{V}^{\star T}$. Namely, the two sets of parameters lead to the same Model (2.2). In particular, we define two sets of parameters to be *equivalent* when they give identical likelihood functions (see (2.6) below).

For regression purpose knowing $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ is enough, but for dimension reduction purpose we need to obtain all specific parameters since each parameter has an important interpretation. For example, the columns of $\mathbf{V}$ can be interpreted as projection directions; the matrix $\boldsymbol{\Sigma}_\mathbf{f}$ gives the covariance structure of latent scores; each column of $\mathbf{B}$ indicates how the supervision matrix $\mathbf{Y}$ is related with the corresponding score vector. Therefore we impose the following constraints to identify the model.

(1) The $p \times r$ matrix $\mathbf{V}$ has orthonormal columns, i.e., $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$;

(2) The $r \times r$ matrix $\boldsymbol{\Sigma}_\mathbf{f}$ is diagonal with r distinct positive eigenvalues;

(3) The columns of $\mathbf{V}$ are sorted in the descending order in terms of column norms of $\mathbf{XV}$, and the first entry of each column is positive.

The first condition is commonly used in SVD analysis. Each loading vector corresponds with a projection direction. The orthonormality of loading vectors naturally leads to an orthogonal basis with unit lengths. The second condition implies that the latent variables in $\mathbf{U}$ are uncorrelated. We assume all diagonal entries to be positive and distinct to avoid indeterminacy of the loading vectors. In practice, this condition generally holds. The third condition rules

out column and sign switches. In addition, we also assume that the supervision data matrix $\mathbf{Y}$ has linearly independent columns; in practice, one can discard linearly dependent columns in $\mathbf{Y}$. Under these conditions, the SupSVD model is identifiable. Hereafter, without special notice, we assume that the model satisfies all the aforementioned identifiability conditions. We comment that the identifiability conditions help us identify the unique representative in an equivalence class.

Proposition 2.2.1. *In Model (2.2), for any parameter set* $(\mathbf{B}, \mathbf{V}, \mathbf{\Sigma_f}, \sigma_\mathbf{e}^2)$ *such that the largest* r *eigenvalues of* $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_\mathbf{e}^2\mathbf{I}$ *are distinct and greater than the remaining eigenvalues, there exists an unique parameter set that is equivalent with* $(\mathbf{B}, \mathbf{V}, \mathbf{\Sigma_f}, \sigma_\mathbf{e}^2)$ *and satisfies the identifiability conditions.*

For cases in which two or more of the first r eigenvalues of $\mathbf{\Sigma}$ are equal, the above conditions are not sufficient for identifiability, and one may have to impose constraints on $\mathbf{B}$ as well. However, in real data examples, equal-eigenvalue cases rarely occur. Therefore, we can reasonably restrict our scope to models that satisfy the identifiability conditions.

2.2.2 Connections with Existing Models

The SupSVD model (2.2) has close connections with several existing models. On the one hand, when $\mathbf{B} = \mathbf{0}$, i.e., when the score matrix $\mathbf{U}$ equals to the random matrix $\mathbf{F}$, Model (2.2) reduces to

$$\mathbf{X} = \mathbf{F}\mathbf{V}^T + \mathbf{E}. \tag{2.3}$$

In Model (2.3), each row of $\mathbf{X}$ is i.i.d. with mean zero and covariance matrix $\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_\mathbf{e}^2\mathbf{I}_p$, which is exactly the r-component spiked covariance model for PCA, cf. Johnstone (2001); Paul (2007); Shen et al. (2013). In the model, the r columns of $\mathbf{V}$ are the first r principal component (PC) loadings, and the columns of $\mathbf{XV}$ are the corresponding PCs. Note that the PCA model is *unsupervised*, as the matrix $\mathbf{Y}$ does not appear in the model.

On the other hand, when the latent score matrix $\mathbf{U}$ is fully driven by $\mathbf{Y}$, i.e., $\mathbf{\Sigma_f} = \mathbf{0}$, the

SupSVD model reduces to

$$\mathbf{X} = \mathbf{Y}\mathbf{B}\mathbf{V}^T + \mathbf{E}, \tag{2.4}$$

where for identifiability purposes we let $\mathbf{B}$ have orthogonal columns. We note that Model (2.4) is the reduced rank regression (RRR) model (Izenman, 1975; Reinsel and Velu, 1998) with isotropic covariance structure (we will refer to isotropic RRR as RRR). The matrix $\mathbf{C} = \mathbf{B}\mathbf{V}^T$ is the rank r coefficient matrix whose least square estimator is explicitly given in Reinsel and Velu (1998). In this case, the true underlying structure of $\mathbf{X}$ is $\mathbf{Y}\mathbf{B}\mathbf{V}^T$, whose column space is a subspace of the column space of $\mathbf{Y}$. In other words, the underlying structure is fully driven by the supervision information. We therefore refer to the RRR model as *fully supervised*.

The SupSVD model (2.2) is also connected with the envelope model that was recently proposed by Cook et al. Cook et al. (2010) and further developed in (Cook et al., 2013; Cook and Su, 2013; Cook and Zhang, 2015; Su and Cook, 2011). The envelope model is a parsimonious model for multivariate regression that is based on the assumption that variation in the response can be divided into two parts: a material part that is related to the predictor, and an immaterial part that is unrelated to the predictor. The envelope model achieves substantial efficiency gain in parameter estimation by focusing on the material part of the response. The coordinate version of the envelope model can be written as

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{x} + \boldsymbol{\varepsilon} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T. \end{aligned} \tag{2.5}$$

Here $\mathbf{y}$ is a p-dimensional response, $\mathbf{x}$ is a q-dimensional predictor, $\boldsymbol{\Gamma}$ is $p \times r$ semi-orthogonal matrix and $\boldsymbol{\eta}$ is an $r \times q$ matrix. The product of $\boldsymbol{\Gamma}$ and $\boldsymbol{\eta}$ acts as a coefficient, while $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ are the intercept and the random error. The random error $\boldsymbol{\varepsilon}$ has covariance matrix $\boldsymbol{\Sigma}$ defined in the second equation, in which $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ is orthogonal, and $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are positive definite.

If we regard the response as the primary data to be approximated, and the predictor as the supervision data, it can be shown that Model (2.5) coincides with Model (2.2). The

covariance of (2.2) is slightly more specific than that of (2.5). However, we note that the two models arise in the analysis of different problems, and that they have different applications and interpretations. The SupSVD model attempts to extract a low rank representation of a primary data matrix, and is intended for dimension reduction problems in which auxiliary data is present. The goal of the envelope model is to reduce the variation of coefficient estimation in regression problems. Here we impose identifiability conditions on the model and estimate each parameter, as the parameters are directly interpretable in the context of dimension reduction. In Cook et al. (2010) the authors focus on identifying estimable subspaces that are spanned by the parameters of their model; the parameters themselves are of less importance. In addition, fitting of the SupSVD and envelope models is carried out in fundamentally different ways. We describe a computationally efficient EM type algorithm to fit the model (2.1) for which the likelihood of the observed data usually converges to a local maximum after a few iterations. In order to fit the envelope model, the authors of Cook et al. (2010) directly maximize the likelihood function, which involves optimization over a Grassmann manifold. We compared the computational speeds of both methods using various simulations, and in general the EM algorithm is faster.

SupSVD can be viewed as a general model for supervised dimension reduction. It encompasses unsupervised PCA and fully supervised RRR as two extremes. When the auxiliary information is irrelevant to low rank structure of the primary data, the SupSVD model reduces to the PCA model; when the underlying structure is totally driven by the auxiliary data, the SupSVD model reduces to the RRR model. It also connects with the envelope model from a multivariate regression point of view.

2.3 Model Estimation

In this section, we describe the parameter estimation algorithm, incorporating the identifiability constraints discussed in Section 2.2.1. We assume multivariate normality for the random matrices $\mathbf{E}$ and $\mathbf{F}$ hereafter. To begin, we assume that the rank of the underlying structure of $\mathbf{X}$ is known to be r. Data-driven selection of the rank r is discussed at the end

of this section.

Under the normality assumption for $\mathbf{E}$ and $\mathbf{F}$, we can obtain the distribution of the observed data $\mathbf{X}$ according to (2.2) as

$$\mathrm{vec}(\mathbf{X}^T) \sim \mathcal{N}_{np}\left(\mathrm{vec}(\mathbf{V}\mathbf{B}^T\mathbf{Y}^T),\ \mathbf{I}_n \otimes (\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p)\right),$$

where $\mathrm{vec}(\cdot)$ is the column-stacking operator and $\otimes$ is the Kronecker product. Thus the log likelihood of $\mathbf{X}$ can be expressed explicitly as

$$\begin{aligned}\mathcal{L}(\mathbf{X}) = &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det\left(\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p\right) \\ &-\frac{1}{2}\mathrm{tr}\left((\mathbf{X}-\mathbf{Y}\mathbf{B}\mathbf{V}^T)(\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p)^{-1}(\mathbf{X}-\mathbf{Y}\mathbf{B}\mathbf{V}^T)^T\right),\end{aligned} \tag{2.6}$$

where the parameters satisfy the identifiability conditions discussed above.

One way to estimate the parameters is to directly maximize the likelihood function (2.6) under the identifiability conditions. However, a direct constrained maximization is challenging for two reasons: 1) $\mathbf{V}$ appears in both the mean and the variance of the normal distribution; and 2) the constrained parameter space is not convex. As a remedy, we propose a modified EM algorithm, namely an *expectation-maximization-standardization (EMS)* algorithm, to efficiently estimate the model parameters. The additional standardization step guarantees that the parameter estimates satisfy the identifiability conditions.

The latent matrix $\mathbf{U}$ in Model (2.1) naturally suggests the possibility of using the EM algorithm for parameter estimation. The joint log likelihood of $\mathbf{X}$ and $\mathbf{U}$, i.e., $\mathcal{L}(\mathbf{X},\mathbf{U})$, can be separated into two parts: the conditional log likelihood of $\mathbf{X}$ given $\mathbf{U}$, and the marginal log likelihood of $\mathbf{U}$. In detail,

$$\mathcal{L}(\mathbf{X},\mathbf{U}) \quad = \quad \mathcal{L}(\mathbf{X}|\mathbf{U}) + \mathcal{L}(\mathbf{U}), \tag{2.7}$$

where

$$\begin{aligned} \text{vec}(\mathbf{X}^T) \mid \mathbf{U} &\sim \mathcal{N}_{np}\left(\text{vec}(\mathbf{V}\mathbf{U}^T), \sigma_{\mathbf{e}}^2 \mathbf{I}_{np}\right), \text{ and} \quad (2.8) \\ \text{vec}(\mathbf{U}^T) &\sim \mathcal{N}_{nr}\left(\text{vec}(\mathbf{B}^T\mathbf{Y}^T), \mathbf{I}_n \otimes \Sigma_{\mathbf{f}}\right). \quad (2.9) \end{aligned}$$

The benefit of this separation is that the parameters $(\mathbf{B}, \mathbf{\Sigma_f})$ are isolated from $(\mathbf{V}, \sigma_{\mathbf{e}}^2)$, and each parameter only contributes to one part of the likelihood. Using (2.7) the joint log likelihood has the following form:

$$\begin{aligned} \mathcal{L}(\mathbf{X}, \mathbf{U}) \propto & -np \log \sigma_{\mathbf{e}}^2 - \sigma_{\mathbf{e}}^{-2} \text{tr}\left((\mathbf{X} - \mathbf{U}\mathbf{V}^T)(\mathbf{X} - \mathbf{U}\mathbf{V}^T)^T\right) \\ & -n \log \det \mathbf{\Sigma_f} - \text{tr}\left((\mathbf{U} - \mathbf{Y}\mathbf{B})\mathbf{\Sigma_f}^{-1}(\mathbf{U} - \mathbf{Y}\mathbf{B})^T\right). \end{aligned}$$

Below we describe the steps of the EMS algorithm, which is presented as **Algorithm 1** at the end of this section. We use $\theta^{(i)} = (\mathbf{B}^{(i)}, \mathbf{V}^{(i)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i)}, \sigma_{\mathbf{e}}^{2(i)})$ to denote the parameter estimates obtained in the ith iteration, which satisfy the identifiability conditions.

E Step: We calculate the conditional expectation of $\mathcal{L}(\mathbf{X}, \mathbf{U})$ with respect to $\mathbf{U}$ given $\mathbf{X}$ and $\theta^{(i)}$, i.e., $\mathbb{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X}, \mathbf{U})|\mathbf{X}, \theta^{(i)})$. The conditional distribution of $\mathbf{U}$ given $\mathbf{X}$ and the previous parameter estimation $\theta^{(i)}$ is

$$\text{vec}(\mathbf{U}^T) \mid \mathbf{X} \sim \mathcal{N}\left(\text{vec}\left(\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}{}^T\right), \ \mathbf{I}_n \otimes \Omega_{\mathbf{U}|\mathbf{X}}^{(i)} \right), \quad (2.10)$$

where

$$\begin{aligned} \Theta_{\mathbf{U}|\mathbf{X}}^{(i)} &= \mathbb{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X}) = \left(\mathbf{Y}\mathbf{B}^{(i)}\left(\sigma_{\mathbf{e}}^{2(i)}\mathbf{\Sigma}_{\mathbf{f}}^{(i)^{-1}}\right) + \mathbf{X}\mathbf{V}^{(i)}\right)\left(\mathbf{I}_r + \sigma_{\mathbf{e}}^{2(i)}\mathbf{\Sigma}_{\mathbf{f}}^{(i)^{-1}}\right)^{-1}, \\ \Omega_{\mathbf{U}|\mathbf{X}}^{(i)} &= \left(\mathbf{\Sigma}_{\mathbf{f}}^{(i)^{-1}} + \sigma_{\mathbf{e}}^{-2(i)}\mathbf{I}_r\right)^{-1}. \end{aligned}$$

Note that the conditional expectation of $\mathbf{U}$ given $\mathbf{X}$ is a weighted average of $\mathbf{Y}\mathbf{B}^{(i)}$ and $\mathbf{X}\mathbf{V}^{(i)}$, where the weights are determined by $\sigma_{\mathbf{e}}^{2(i)}$ and $\mathbf{\Sigma}_{\mathbf{f}}^{(i)}$.

M Step: We maximize $\mathbb{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X}, \mathbf{U})|\mathbf{X}, \theta^{(i)})$ with respect to all the parameters under the identifiability constraints in Section 2.2.1. The optimization is challenging since the constraint

is not convex. As the joint distribution of $\mathbf{X}$ and $\mathbf{U}$ is identifiable even without the side conditions, we propose a modified EM algorithm that bypasses the constrained optimization problem. More specifically, we first obtain the unconstrained optimizers of $\mathbb{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X},\mathbf{U})|\mathbf{X},\theta^{(i)})$, and then find the unique set of parameters that is equivalent to the optimizers in terms of the SupSVD model, and that satisfies the identifiability conditions.

The unconstrained optimization problem can be solved analytically. Setting partial derivatives of $\mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X},\mathbf{U})|\mathbf{X},\theta^{(i)})$ with respective to each parameter to zero, we obtain

$$\begin{aligned}
\widehat{\mathbf{B}} &= (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbb{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X},\theta^{(i)}), && (2.11)\\
\widehat{\mathbf{V}} &= \mathbf{X}^T\mathbb{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X},\theta^{(i)})\left[\mathbb{E}_{\mathbf{U}}(\mathbf{U}^T\mathbf{U}\ |\mathbf{X},\theta^{(i)})\right]^{-1}, && (2.12)\\
\widehat{\mathbf{\Sigma_f}} &= \frac{1}{n}\mathbb{E}_{\mathbf{U}}\left[(\mathbf{U}-\mathbf{Y}\widehat{\mathbf{B}})^T(\mathbf{U}-\mathbf{Y}\widehat{\mathbf{B}})\ |\mathbf{X},\theta^{(i)}\right], && (2.13)\\
\widehat{\sigma^2_{\mathbf{e}}} &= \frac{1}{np}\mathbb{E}_{\mathbf{U}}\left[\mathrm{tr}((\mathbf{X}-\mathbf{U}\widehat{\mathbf{V}}^T)(\mathbf{X}-\mathbf{U}\widehat{\mathbf{V}}^T)^T)\ |\mathbf{X},\theta^{(i)}\right], && (2.14)
\end{aligned}$$

where the corresponding conditional expectations can be obtained from (2.10). Details can be found in the appendix, Section 2.7.3.

S Step: The unconstrained optimizers $(\widehat{\mathbf{B}},\widehat{\mathbf{V}},\widehat{\mathbf{\Sigma_f}},\widehat{\sigma^2_{\mathbf{e}}})$ in (2.11)–(2.14) typically satisfy the condition of Proposition 2.2.1. In this case, we can obtain the unique equivalent set of parameters that satisfy the identifiability conditions. In particular, we perform SVD on $\widehat{\mathbf{V}}\widehat{\mathbf{\Sigma_f}}\widehat{\mathbf{V}}^T$ to obtain the following eigen-decomposition:

$$\mathbf{V}^{(i+1)}\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}\mathbf{V}^{(i+1)^T} = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma_f}}\widehat{\mathbf{V}}^T,$$

where the columns of $\mathbf{V}^{(i+1)}$ are the orthonormal eigenvectors and the diagonal entries of the diagonal matrix $\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}$ are the eigenvalues. In practice, the eigenvalues are almost always positive and distinct, so that the matrices $\mathbf{V}^{(i+1)}$ and $\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}$ satisfy the identifiability conditions and are unique up to a column reordering. Then, we set $\mathbf{B}^{(i+1)} = \widehat{\mathbf{B}}\widehat{\mathbf{V}}^T\mathbf{V}^{(i+1)}$ and $\sigma_{\mathbf{e}}^{2(i+1)} = \widehat{\sigma^2_{\mathbf{e}}}$. It is easy to see that

$$\mathbf{B}^{(i+1)}\mathbf{V}^{(i+1)^T} = \widehat{\mathbf{B}}\widehat{\mathbf{V}}^T.$$

Lastly, we reorder the columns of $\mathbf{V}^{(i+1)}$, and accordingly the columns of $\mathbf{B}^{(i+1)}$ and the rows/columns of $\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}$, in order to ensure that the column norms of $\mathbf{X}\mathbf{V}^{(i+1)}$ are decreasing. As a result, we get parameter estimates $\theta^{(i+1)} = (\mathbf{B}^{(i+1)}, \mathbf{V}^{(i+1)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}, \sigma_{\mathbf{e}}^{2(i+1)})$ for the $(i+1)$th iteration.

Each step of the EMS algorithm has an analytical expression and can be computed efficiently. Our numerical studies indicate that the algorithm is insensitive to initial values. In practice, we use the naive estimates from SVD as the initial values. The following proposition guarantees convergence of the EMS algorithm to a local optimum.

Proposition 2.3.1. *In each iteration of the EMS algorithm, the log likelihood of the observed data $\mathcal{L}(\mathbf{X})$ is monotonically nondecreasing. Therefore, the EMS algorithm always converges to some stationary point (maybe local maximum).*

Algorithm 1 The EMS Algorithm for Parameter Estimation under the SupSVD Model

1: Set initial values for the parameters $(\mathbf{B}^{(0)}, \mathbf{V}^{(0)}, \mathbf{\Sigma}_{\mathbf{f}}^{(0)}, \sigma_{\mathbf{e}}^{2(0)})$;
2: **while** $\mathcal{L}(\mathbf{X}|\theta^{(i+1)}) - \mathcal{L}(\mathbf{X}|\theta^{(i)}) >$threshold **do**
3: **E Step:** Derive the conditional distribution (2.10) given $\theta^{(i)} = (\mathbf{B}^{(i)}, \mathbf{V}^{(i)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i)}, \sigma_{\mathbf{e}}^{2(i)})$;
4: **M Step:** Obtain the unconstrained optimizer $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma}_{\mathbf{f}}}, \widehat{\sigma_{\mathbf{e}}^2})$ from (2.11)-(2.14);
5: **S Step:** Standardize $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma}_{\mathbf{f}}}, \widehat{\sigma_{\mathbf{e}}^2})$ to get $\theta^{(i+1)} = (\mathbf{B}^{(i+1)}, \mathbf{V}^{(i+1)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}, \sigma_{\mathbf{e}}^{2(i+1)})$ that satisfy the identifiability conditions;
6: Set $i \leftarrow i + 1$.
7: **end while**

The presentation in this section assumes that the rank r is known. In practice, the rank has to be determined from the data. In the numerical studies of Section 2.5.1, we adopt a popular practice within the PCA literature: using the scree plot of a primary data matrix to determine a proper rank. The rationale is that we assume the rank of the underlying signal of a primary data matrix is inherent. Auxiliary information is used to help recover the underlying low-rank structure more accurately, without altering the rank. Other rank selection methods that have been studied in the PCA literature, e.g., the permutation assessment method in Buja and Eyuboglu (1992) and the bi-cross-validation method in Owen and Perry (2009), are also appropriate in our framework. The likelihood ratio test approach of Cook et al. (2010) could be used to select r in the SupSVD model as well.

2.4 Asymptotic Analysis

In this section, we state the consistency and asymptotic normality of the SupSVD parameter estimates. Since the SupSVD model is overparameterized, i.e., unidentifiable without side conditions, standard asymptotics from the maximum likelihood framework do not apply directly. Instead, we refer to the asymptotic results in Shapiro (1986) for overparameterized structural models. A similar treatment can be found in Cook et al. (2010).

Specifically, we first focus on the estimable functions $\boldsymbol{\beta} = \mathbf{B}\mathbf{V}^T$ and $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T + \sigma^2_\mathbf{e}\mathbf{I}$, which uniquely define the likelihood function. In order to fit our analysis into the framework of (Shapiro, 1986), we rewrite the parameters as

$$\boldsymbol{\phi} = \begin{pmatrix} \text{vec}(\mathbf{B}) \\ \text{vec}(\mathbf{V}) \\ \text{vech}(\boldsymbol{\Sigma}_\mathbf{f}) \\ \sigma^2_\mathbf{e} \end{pmatrix} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix},$$

where the operator vech$(\cdot)$ stacks the lower triangular part of a symmetric matrix into a vector. The estimable functions can then be expressed as

$$\mathbf{h}(\boldsymbol{\phi}) = \begin{pmatrix} \text{vec}(\boldsymbol{\beta}) \\ \text{vech}(\boldsymbol{\Sigma}) \end{pmatrix} = \begin{pmatrix} \text{vec}(\mathbf{B}\mathbf{V}^T) \\ \text{vech}(\mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T + \sigma^2_\mathbf{e}\mathbf{I}) \end{pmatrix} = \begin{pmatrix} h_1(\boldsymbol{\phi}) \\ h_2(\boldsymbol{\phi}) \end{pmatrix}. \tag{2.15}$$

For any $d \times d$ symmetric matrix $\boldsymbol{\Omega}$, we denote the $d(d+1)/2 \times d^2$ constant contraction matrix as $\mathbf{C}_d$, and the $d^2 \times d(d+1)/2$ constant expansion matrix as $\mathbf{E}_d$ to relate the operator vech$(\cdot)$ and vec$(\cdot)$, i.e., vech$(\boldsymbol{\Omega}) = \mathbf{C}_dvec(\boldsymbol{\Omega})$ and vec$(\boldsymbol{\Omega}) = \mathbf{E}_d$vech$(\boldsymbol{\Omega})$. Moreover, for any $l \times m$ matrix $\boldsymbol{\Gamma}$, we denote the $lm \times lm$ constant commutation matrix as $\mathbf{K}_{lm}$, i.e., vec$(\boldsymbol{\Gamma}^T) = \mathbf{K}_{lm}vec(\boldsymbol{\Gamma})$. We can obtain the following theorem, whose proof can be found in the appendix, Section 2.7.4.

Theorem 2.4.1. *Assume Model (2.2) and let* $\mathbf{h}(\cdot)$ *be as in (2.15). Denote* $\mathbf{H} = \partial\mathbf{h}(\boldsymbol{\phi})/\partial\boldsymbol{\phi}$, *and let* $\mathbf{J}$ *be the Fisher information of* $\mathbf{h}(\boldsymbol{\phi})$. *Let* $\widehat{\mathbf{h}}$ *be the maximum likelihood estimator of*

$\mathbf{h}$. *Then,*

$$\sqrt{n}(\widehat{\mathbf{h}} - \mathbf{h}) \to_d \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_h}), \tag{2.16}$$

where $\mathbf{\Sigma_h} = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$, *where* $\dagger$ *indicates the Moore-Penrose inverse. Specifically,*

$$\mathbf{H} = \begin{pmatrix} \mathbf{V} \otimes \mathbf{I}_q & (\mathbf{I}_p \otimes \mathbf{B})\mathbf{K}_{pr} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_p(\mathbf{V}\mathbf{\Sigma_f} \otimes \mathbf{I}_p) & \mathbf{C}_p(\mathbf{V} \otimes \mathbf{V})\mathbf{E}_r & vech(\mathbf{I}_p) \end{pmatrix}$$

and

$$\mathbf{J} = \begin{pmatrix} \mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma_Y} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{E}_p^T(\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1})\mathbf{E}_p \end{pmatrix}$$

where $\mathbf{\Sigma_Y} = \lim_{n\to\infty} \mathbf{Y}\mathbf{Y}^T/n$.

As a result, we know that $\sqrt{n}$ vec$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\sqrt{n}$ vech$(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})$ are jointly asymptotically normally distributed with mean zero. Moreover, under the identifiability conditions, we obtain the following asymptotic property for each parameter in $\widehat{\boldsymbol{\phi}}$.

Corollary 2.4.1. *Given (2.16), under the identifiability conditions,* $\sqrt{n}$ *vec*$(\widehat{\mathbf{B}}-\mathbf{B})$, $\sqrt{n}$ *vec*$(\widehat{\mathbf{V}}-\mathbf{V})$, $\sqrt{n}$ *diag*$(\widehat{\mathbf{\Sigma_f}} - \mathbf{\Sigma_f})$, *and* $\sqrt{n}$ $(\widehat{\sigma_\mathbf{e}^2} - \sigma_\mathbf{e}^2)$ *are asymptotically jointly normal with mean zero. The asymptotic covariance matrix of* $\sqrt{n}$ $(\widehat{\mathbf{v}}_i - \mathbf{v}_i)$, *where* $\widehat{\mathbf{v}}_i$ *and* $\mathbf{v}_i$ *are the* i*th columns of* $\widehat{\mathbf{V}}$ *and* $\mathbf{V}$ *respectively, is given in the appendix, Section 2.7.5.*

2.5 Numerical Examples

We compare SupSVD with SVD and RRR using extensive simulations (Section 2.5.1) and a real data example (Section 2.5.2). Section 2.5.1.1 compares the three methods with data simulated from each of the models respectively to show the adaptivity of SupSVD. Section 2.5.1.2 illustrates the performances of the methods under a spectrum of settings ranging from PCA to RRR. In Section 2.5.2, we illustrate SupSVD using the breast cancer data from The

Cancer Genome Atlas Network (2012). Additional simulation and real data examples can be found in the appendix, Section 2.7.6, 2.7.7, and 2.7.8.

2.5.1 Simulation Studies

2.5.1.1 Adaptivity of SupSVD

We consider three simulation examples where the data are generated from each one of the three models (SupSVD, PCA, RRR) respectively. In particular, the PCA example illustrates a situation where the "supervision" data are actually not related to the primary data; the RRR example illustrates a situation where the underlying structure of primary data is fully driven by supervision. For each simulated example, we apply all three methods to analyze the simulated data, and demonstrate the adaptivity of SupSVD under different settings. We have tried a range of parameter settings in each case and the results are concordant across settings. Below we choose to only present representative results in each example.

In all three examples, we set the sample size $n = 100$, the dimension of $\mathbf{X}$ as $p = 68$, and the dimension of the supervision data $\mathbf{Y}$ as $q = 4$. The rank of the underlying structure is set to be $r = 2$. We fill in the supervision data matrix $\mathbf{Y}$ with numbers generated from a standard normal distribution. The loading vectors in $\mathbf{V}$ are set to be the first two orthogonal loadings with unit norms estimated from the call center data in the appendix, Section 2.7.8. The intention is to make the simulation setting as realistic as possible. In particular, the primary data matrix $\mathbf{X}$ is generated in the following ways for different examples.

(1) **Case 1 (SupSVD): $\mathbf{X}$** is generated from the SupSVD model $\mathbf{X} = \mathbf{YBV}^T + \mathbf{FV}^T + \mathbf{E}$. The 4×2 fixed coefficient matrix $\mathbf{B}$ is standardized to have orthogonal columns with norm 3. The matrix $\mathbf{F}$ has i.i.d. rows from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma_f} = \text{diag}(9, 4)$. The matrix $\mathbf{E}$ has i.i.d. entries from $\mathcal{N}(0, 3)$.

(2) **Case 2 (PCA): $\mathbf{X}$** is generated from the PCA model $\mathbf{X} = \mathbf{FV}^T + \mathbf{E}$, where $\mathbf{F}$ is generated in the same way as in Case 1, and $\mathbf{E}$ has i.i.d. entries from a standard normal

distribution.

(3) **Case 3 (RRR): X** is generated from the RRR model $\mathbf{X} = \mathbf{Y}\mathbf{B}\mathbf{V}^T + \mathbf{E}$, where the 4×2 fixed coefficient matrix $\mathbf{B}$ is standardized to have orthogonal columns with norm 6 and 3 respectively. The error matrix $\mathbf{E}$ has i.i.d. entries from $\mathcal{N}(0, 3)$.

Performance Measures The three methods are compared in two aspects, *low rank structure recovery* and *parameter estimation*. The low rank recovery accuracy is measured by the mean square error (MSE) defined as

$$MSE_{\mathbf{U}\mathbf{V}^T} = \frac{1}{np}\|\mathbf{U}\mathbf{V}^T - \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T\|_{\mathbb{F}}^2,$$

where $\|\cdot\|_{\mathbb{F}}$ denotes the Frobenius norm, and $\mathbf{U}\mathbf{V}^T$ and $\widehat{\mathbf{U}}\widehat{\mathbf{V}}^T$ are the true and estimated low rank structures respectively. For SVD, $\widehat{\mathbf{U}} = \mathbf{X}\widehat{\mathbf{V}}_{SVD}$; for RRR, $\widehat{\mathbf{U}} = \mathbf{Y}\widehat{\mathbf{B}}_{RRR}$; for SupSVD, $\widehat{\mathbf{U}} = \left(\mathbf{Y}\widehat{\mathbf{B}}(\widehat{\sigma_{\mathbf{e}}^2}\widehat{\mathbf{\Sigma}_{\mathbf{f}}}^{-1}) + \mathbf{X}\widehat{\mathbf{V}}\right)\left(\mathbf{I}_r + \widehat{\sigma_{\mathbf{e}}^2}\widehat{\mathbf{\Sigma}_{\mathbf{f}}}^{-1}\right)^{-1}$, where $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma}_{\mathbf{f}}}, \widehat{\sigma_{\mathbf{e}}^2})$ is the parameter set estimated from the SupSVD approach. We also considered other matrix norms such as 1-norm and 2-norm (Golub and Van Loan, 2012), and obtained similar results.

For parameter estimation, only the loading matrix $\mathbf{V}$ and the noise variance $\sigma_{\mathbf{e}}^2$ are common across the three methods. We use the following performance measures:

$$MSE_{\mathbf{V}} = \frac{1}{pr}\|\mathbf{V} - \widehat{\mathbf{V}}\|_{\mathbb{F}}^2, \qquad MSE_{\sigma_{\mathbf{e}}^2} = (\sigma_{\mathbf{e}}^2 - \widehat{\sigma_{\mathbf{e}}^2})^2.$$

Moreover, since the columns of a loading matrix form a basis for a projection subspace, we also measure the largest principal angle (Golub and Van Loan (2012)) between the true subspace and the estimated subspace which is defined as

$$Angle_{\mathbf{V}} \;=\; \frac{180}{\pi}\arccos(\min \mathrm{eig}(\mathbf{V}^T\widehat{\mathbf{V}})),$$

where $\min \mathrm{eig}(\cdot)$ denotes the minimal eigenvalue.

Results For each case, we repeat the simulation 100 times and present in Table 2.1 the median and the median absolute deviations (MAD) of each performance measurement for the three

methods. The results clearly show that SupSVD performs favorably no matter which true model the data are generated from, while SVD and RRR only work well in their respective settings. This demonstrates that SupSVD, covering SVD and RRR as special cases, adapts to a wide range of practical situations. In practice, whenever additional information is available (whether it is truly supervision or not), SupSVD is always a good choice for dimension reduction. In these simulations, SupSVD always provides the best results, equivalent to (or better than) the method corresponding to the true data generative model.

		SupSVD	SVD	RRR
Case 1 (SupSVD)	$MSE_{\mathbf{UV}^T}$	**0.1289** (0.0082)	0.1830 (0.0128)	0.2487 (0.0154)
	$MSE_{\mathbf{V}}$	**0.0025** (0.0005)	0.0036 (0.0013)	0.0080 (0.0048)
	$MSE_{\sigma_e^2}$	**0.0104** (0.0066)	0.0357 (0.0127)	0.0075 (0.0060)
	$Angle_{\mathbf{V}}$	**23.1605** (1.3312)	**23.5571** (1.4463)	27.0765 (2.0600)
Case 2 (PCA)	$MSE_{\mathbf{UV}^T}$	**0.0497** (0.0035)	0.0606 (0.0036)	0.2066 (0.0138)
	$MSE_{\mathbf{V}}$	**0.0022** (0.0003)	**0.0022** (0.0003)	0.0199 (0.0027)
	$MSE_{\sigma_e^2}$	**0.0009** (0.0007)	0.0035 (0.0014)	0.0231 (0.0056)
	$Angle_{\mathbf{V}}$	**25.0287** (2.3239)	**24.9046** (2.1729)	77.1232 (7.0102)
Case 3 (RRR)	$MSE_{\mathbf{UV}^T}$	**0.0659** (0.0051)	0.1845 (0.0097)	**0.0635** (0.0055)
	$MSE_{\mathbf{V}}$	0.0032 (0.0014)	0.0024 (0.0003)	**0.0018** (0.0002)
	$MSE_{\sigma_e^2}$	**0.0082** (0.0064)	0.0329 (0.0130)	**0.0063** (0.0052)
	$Angle_{\mathbf{V}}$	**25.4285** (1.6554)	29.4099 (2.0742)	**25.2282** (1.5882)

Table 2.1: Median(MAD) for Low Rank Structure Recovery Accuracy and Parameter Estimation Accuracy.

Note that Table 2.1 also shows that the $MSE_{\mathbf{V}}$ of SupSVD is larger than the other two methods when the data are generated from the RRR model, i.e. in Case 3. We remark that this is due to the low identifiability of the SupSVD model when the true $\mathbf{\Sigma_f}$ is exactly zero. Numerically, SupSVD is still applicable but the estimated loading vectors are subject to an unstable orthogonal rotation. However, we comment that the estimated projection subspace of $\mathbf{V}$ (i.e., $Angle_{\mathbf{V}}$) and the low-rank recovery accuracy (i.e., $MSE_{\mathbf{UV}^T}$) are unaffected. Generally, it's very unlikely that the underlying structure of a primary data matrix is fully driven by supervision without any variations in practice. Therefore, we do not view this as a major drawback of SupSVD for practical use.

2.5.1.2 Comparison across A Spectrum

We now compare SupSVD, SVD and RRR across a spectrum of simulation settings ranging from the PCA model to the RRR model. For easy presentation, we set $n = 210$, $p = 68$, $q = 1$, and $r = 1$. Fill the 210×1 vector $\mathbf{Y}$ with standard normal random numbers. We simulate $\mathbf{X}$ from the SupSVD model, with the loading vector being the first column of $\mathbf{V}$ in Case 1 above, $\sigma_{\mathbf{e}}^2 = 16$, and $(\mathbf{B}, \mathbf{\Sigma_f}) \in \{(0, 36), (1, 25), (2, 16), (3, 9), (4, 0)\}$, corresponding to Setting 1 to 5, respectively. Therefore, the SupSVD model ranges from the PCA model $\mathbf{X} = 6\mathbf{Z}\mathbf{V}^T + \mathbf{E}$ (Setting 1; $\mathbf{Z}$ is a random vector with i.i.d. entries from standard normal distribution) to the RRR model $\mathbf{X} = 4\mathbf{Y}\mathbf{V}^T + \mathbf{E}$ (Setting 5). Again, under each setting, we run 100 simulations and summarize the results.

To avoid redundancy, we only show the median curves of $MSE_{\mathbf{U}\mathbf{V}^T}$, $MSE_{\mathbf{V}}$, and $Angle_{\mathbf{V}}$ for the methods in Figure 2.1. We observe that SupSVD is uniformly the best over the spectrum of settings, with similar performance with SVD when the true underlying model is PCA, and similar performance with RRR when the true underlying model is RRR. Again, the results illustrate that SupSVD is a robust method that adapts well over a wide range of data-generating models.

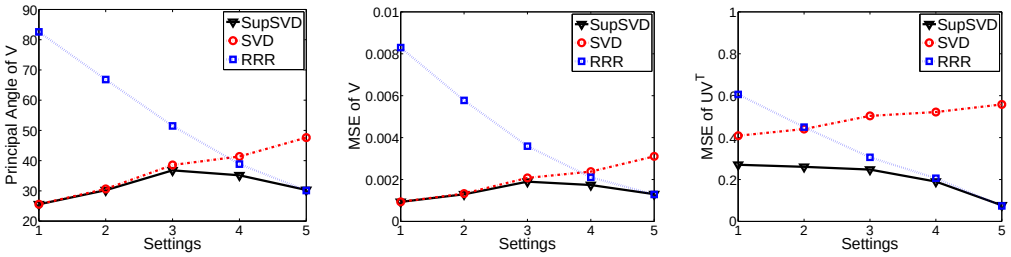


Figure 2.1: Median Curves for $Angle_{\mathbf{V}}$, $MSE_{\mathbf{V}}$, and $MSE_{\mathbf{U}\mathbf{V}^T}$ Based on 100 Simulation Runs.

2.5.2 Breast Cancer Data

We consider a real data set containing gene expression measurements from breast tumors, obtained from the The Cancer Genome Atlas (TCGA) project (The Cancer Genome Atlas Network, 2012). A pointer to the publicly available data is at https://tcga-data.nci.nih.gov/docs/publications/brca_2012/. A primary goal is to understand underlying patterns

of genetic variation among tumors. In this case, we have additional information of disease subtype for each tumor. We may regard cancer subtypes as a partial driver of the underlying structure of the gene expression data (Schadt et al., 2005). Samples from the same subtype will share common genetic variations. We use the subtype information as our supervision data and apply the SupSVD method.

The raw data set contains 17814 genes and 348 samples. Out of the 348 samples, there are 5 subtypes of breast cancer with different number of samples in each subtype: Basal (66), Her2 (42), LumA (154), LumB (81), and Normal (5). We preprocessed the data in the same way as in Lock and Dunson (2013). We first imputed missing values with the k-nearest neighbors algorithm ($k = 10$), then removed genes with low variations across samples (standard deviation smaller than 1.5), and finally mean centered each gene. The result is a column-centered data matrix $\mathbf{X}$ with 348 samples and 645 genes. Based on the scree plot of the singular values of $\mathbf{X}$, we select the rank of the underlying structure to be 3.

Figure 2.2 shows the scatter plots of the estimated SupSVD scores. The first score vector clearly separates the Basal subgroup from the rest. The second score vector captures variations within each subtype. The third score vector roughly separates the Her2, LumA, and LumB subgroups.

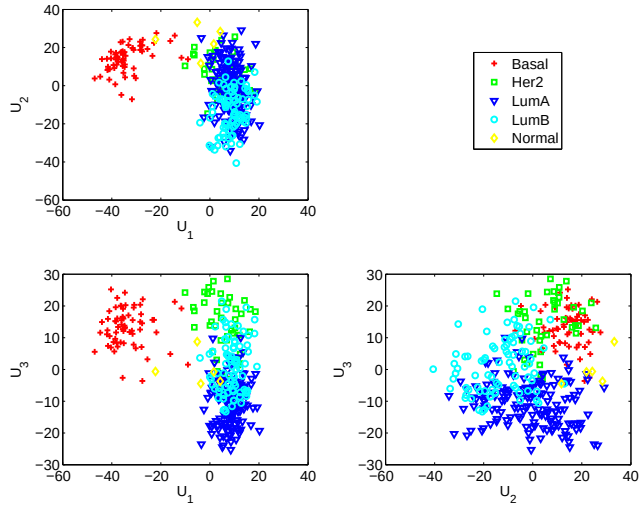


Figure 2.2: Breast Cancer Data - Scatter Plots of SupSVD Score Vectors. The 5 different subtypes are well separated by the first three score vectors.

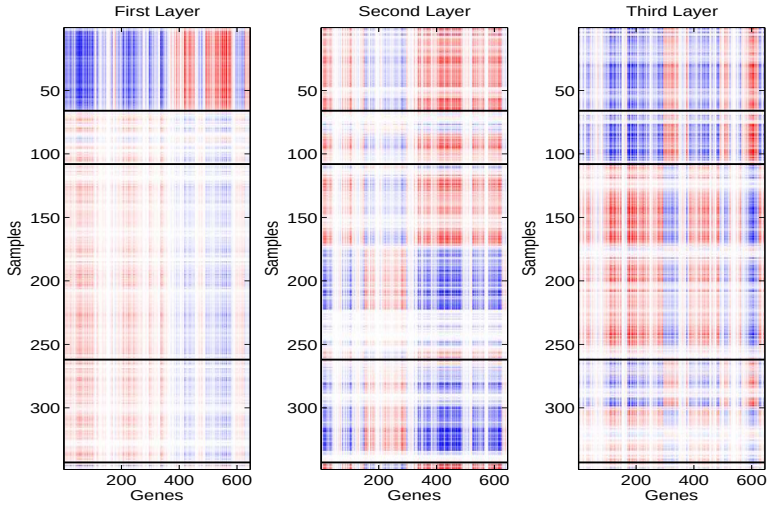


Figure 2.3: Breast Cancer Data - Heat Map of First Three Unit-rank SupSVD Structures of the Gene Expression Data. Blue is negative and red is positive. The samples are grouped in the order of Basal, Her2, LumA, LumB, Normal. The genes are reordered for better visualization.

Figure 2.3 presents the heat maps of the unit-rank structures from SupSVD. There are clear patterns driven by subtypes. For example, the first layer is dominated by the unique pattern in the Basal subgroup. The third layer shows patterns similar between Basal and Her2, but different among Her2, LumA and LumB. There are also within-group variations that are not driven by subtypes. For example, the LumA samples in the second layer clearly exhibit several different patterns. The SVD and RRR results are given in the appendix, Section 2.7.7. In comparison, SupSVD effectively captures important underlying patterns consisting of both between-group variations driven by the subtype information and within-group variations from unknown sources.

2.6 Discussion

In this chapter, we propose a supervised dimension reduction model, SupSVD that takes advantage of auxiliary information to better recover the underlying low-rank structure in the primary data of interest. We focused on recovering comprehensive low-rank structures from the data with the potential guidance of the supervision information. The SupSVD model

contains the PCA model and the RRR model as two extreme cases: when the supervision information is unrelated to the data of interest, SupSVD reduces to PCA; when the underlying structure is fully driven by the supervision information, SupSVD reduces to RRR. SupSVD automatically adjusts the amount of supervision used for dimension reduction without the use of tuning parameters. The proposed EMS algorithm for parameter estimation in SupSVD is computationally efficient. Asymptotic properties of SupSVD are derived for the resulting estimates. Simulation studies and real data applications clearly demonstrate the advantages and flexibility of the SupSVD method.

Dimension reduction is also useful in functional data analysis (FDA) to facilitate various subsequent analyses. For an overview of the FDA literature including recent advances, see Bongiorno et al. (2014); Ferraty and Vieu (2006); Horváth and Kokoszka (2012); Silverman and Ramsay (2005). We remark that our SupSVD method can be directly adapted to FDA through a basis approach. In particular, one can decompose discretized observations of functional data onto proper basis functions, obtain a coefficient matrix, and then apply SupSVD to the coefficient matrix. The low rank approximation obtained from SupSVD can then be converted back to the original functional space through the basis functions. Another approach is to first select important variables in discretized values of the function (Aneiros and Vieu, 2014), and then apply SupSVD to the dimension-reduced vectors. Alternatively, in the next chapter, we extend the recent regularization formulation of functional principal component analysis (Huang et al., 2008, 2009) to incorporate supervision for FDA. We impose both sparsity (Shen and Huang, 2008b) and roughness regularization to incorporate both high-dimensional multivariate data as well as infinite-dimensional functional data.

2.7 Appendix

2.7.1 Proof of Proposition 2.2.1

PROOF. Let $(\mathbf{B}, \mathbf{V}, \mathbf{\Sigma_f}, \sigma_e^2)$ be a parameter set such that $\mathbf{B}$ is a $q \times r$ matrix, $\mathbf{V}$ is a $p \times r$ matrix, $\mathbf{\Sigma_f}$ is a $r \times r$ positive definite matrix, and σ_e^2 is a positive scalar. Moreover, let the largest r eigenvalues of the $p \times p$ matrix $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_e^2\mathbf{I}$ to be distinct and greater than

the rest $p-r$ equal eigenvalues. It's equivalent to say that $\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T$ has r positive distinct eigenvalues. We have the eigen-decomposition of the $p \times p$ matrix $\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T$ as

$$\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma_f}}\widehat{\mathbf{V}}^T$$

where $\widehat{\mathbf{\Sigma_f}}$ is the $r \times r$ diagonal matrix containing the distinct eigenvalues, and $\widehat{\mathbf{V}}$ is the $p \times r$ orthonormal matrix containing the corresponding eigenvectors. Moreover, set $\widehat{\mathbf{B}} = \mathbf{B}\mathbf{V}^T\widehat{\mathbf{V}}$. Since $\mathbf{V}$ and $\widehat{\mathbf{V}}$ have the same column space, we know

$$\mathbf{B}\mathbf{V}^T = \widehat{\mathbf{B}}\widehat{\mathbf{V}}^T.$$

Therefore, the new parameter set $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma_f}}, \sigma_{\mathbf{e}}^2)$ is equivalent with the original parameter set in terms of Model (2.2), and satisfies the aforementioned identifiability conditions.

The uniqueness of the resulting parameter set is guaranteed by the uniqueness of the eigen-decomposition of the matrix with distinct eigenvalues.

2.7.2 Proof of Proposition 2.3.1

PROOF. Let $\boldsymbol{\theta}^{(i)} = (\mathbf{B}^{(i)}, \mathbf{V}^{(i)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i)}, \sigma_{\mathbf{e}}^{2(i)})$ denote the EMS parameter estimation from the ith iteration. From the algorithm we know it satisfies the identifiability conditions. Let $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})$ denote the conditional expectation of the joint log likelihood. Namely,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)}) &= \mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X}, \mathbf{U}|\boldsymbol{\theta})|\mathbf{X}, \boldsymbol{\theta}^{(i)}) \\ &= \mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{U}|\mathbf{X}, \boldsymbol{\theta})|\mathbf{X}, \boldsymbol{\theta}^{(i)}) + \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) \end{aligned}$$

Let $\widehat{\boldsymbol{\theta}}$ denote the unconstrained optimizer from the M step of EMS algorithm. Namely,

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})$$

Referring to the information inequality that $\mathrm{E}_g(\log f) \leq \mathrm{E}_g(\log g)$ for any densities f and g, we have

$$\mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{U}|\mathbf{X}, \widehat{\boldsymbol{\theta}})|\mathbf{X}, \boldsymbol{\theta}^{(i)}) \leq \mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{U}|\mathbf{X}, \boldsymbol{\theta}^{(i)})|\mathbf{X}, \boldsymbol{\theta}^{(i)})$$

Combining with the fact $Q(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}^{(i)}) \geq Q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i)})$, we know

$$\mathcal{L}(\mathbf{X}|\widehat{\boldsymbol{\theta}}) \geq \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}^{(i)})$$

Moreover, let $\boldsymbol{\theta}^{(i+1)}$ denote the equivalent parameter set that satisfies the identifiability conditions. We have

$$\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}^{(i+1)}) = \mathcal{L}(\mathbf{X}|\widehat{\boldsymbol{\theta}}) \geq \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}^{(i)})$$

Therefore, the likelihood of the observed data $\mathbf{X}$ is monotonically nondecreasing with iterations. If we assume the maximum likelihood exists, the EMS algorithm can always converge.

2.7.3 Details of Algorithm 1

In the paper, we propose the EMS algorithm, which is a modified version of EM algorithm, to efficiently estimate the SupSVD model parameters. The detailed calculations for each step in each iteration are described below. We use $(\mathbf{B}^{(i)}, \mathbf{V}^{(i)}, \boldsymbol{\Sigma}_{\mathbf{f}}^{(i)}, \sigma_{\mathbf{e}}^{2(i)})$ to denote the estimations from the ith iteration.

Initial estimation: Our numerical studies indicate the algorithm is not sensitive to initial values. In practice, we apply SVD to the matrix $\mathbf{X}$ to get the initial estimation. More specifically, we first find the rank-r approximation of $\mathbf{X}$ as

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^T \tag{2.17}$$

where $\mathbf{U}$ is the $n \times r$ semi-orthogonal matrix (i.e., the submatrix of the product of the left singular matrix and the diagonal singular value matrix), and $\mathbf{V}$ is the $p \times r$ matrix with

orthonormal columns (i.e., the submatrix of the right singular matrix). Here $\mathbf{V}$ is an initial estimation of $\mathbf{V}$ in our model. We treat $\mathbf{X} - \mathbf{U}\mathbf{V}^T$ as a random matrix with i.i.d. entries from $\mathcal{N}(0, \sigma_{\mathbf{e}}^2)$. Therefore we can get an initial estimation of $\sigma_{\mathbf{e}}^2$. Then we regress $\mathbf{U}$ on $\mathbf{Y}$ and assume that the multivariate residuals are i.i.d. with diagonal covariance structure. The regression coefficient matrix is an initial estimation of $\mathbf{B}$ and the diagonal covariance matrix is an initial estimation of $\mathbf{\Sigma_f}$.

E step: We have the conditional distribution (2.10) of $\mathbf{U}$ given $\mathbf{X}$ under the current parameter estimations. We can calculate the following quantities to be used in M step.

(1) First order conditional expectation:

$$\begin{aligned} \mathrm{E}_{\mathbf{U}}\left(\mathbf{U}|\mathbf{X}, \theta^{(i)}\right) &= \left(\mathbf{Y}\mathbf{B}\left(\sigma_{\mathbf{e}}^{2(i)}{\mathbf{\Sigma_f}^{(i)}}^{-1}\right) + \mathbf{X}\mathbf{V}^{(i)}\right)\left(\mathbf{I}_r + \sigma_{\mathbf{e}}^{2(i)}{\mathbf{\Sigma_f}^{(i)}}^{-1}\right)^{-1} \\ &\triangleq \Theta_{\mathbf{U}|\mathbf{X}}^{(i)} \end{aligned}$$

(2) Second order conditional expectation:

$$\mathrm{E}_{\mathbf{U}}\left(\mathbf{U}^T\mathbf{U}|\mathbf{X}, \theta^{(i)}\right) = n\Omega_{\mathbf{U}|\mathbf{X}}^{(i)} + {\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}}^T\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}$$

where $\Omega_{\mathbf{U}|\mathbf{X}}^{(i)} \triangleq \left({\mathbf{\Sigma_f}^{(i)}}^{-1} + \sigma_{\mathbf{e}}^{-2(i)}\mathbf{I}_r\right)^{-1}$.

(3) Conditional expectation of any quadratic form in $\mathbf{U}$:

$$\mathrm{E}_{\mathbf{U}}\left(\mathrm{tr}\left(\mathbf{U}\mathbf{\Delta}\mathbf{U}^T\right)|\mathbf{X}, \theta^{(i)}\right) = n\mathrm{tr}\left(\mathbf{\Delta}\Omega_{\mathbf{U}|\mathbf{X}}^{(i)}\right) + \mathrm{tr}\left(\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}\mathbf{\Delta}{\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}}^T\right)$$

where $\mathbf{\Delta}$ is any $r \times r$ symmetric matrix.

M step: We maximize the object function $\mathbb{E}_{\mathbf{U}}\left(\mathcal{L}(\mathbf{X}, \mathbf{U})|\mathbf{X}, \theta^{(i)}\right)$ without any constraints. Specifically, we set the partial derivatives of the conditional expectation with respect to all parameters to zero, and solve for the maximizer. Referring to the Leibniz's rule, we can

exchange partial derivative with conditional expectation. We have

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{X},\mathbf{U})}{\partial \mathbf{B}} &= 2(\mathbf{Y}^T\mathbf{U} - \mathbf{Y}^T\mathbf{Y}\mathbf{B})\mathbf{\Sigma}_\mathbf{f}^{-1} \\
\frac{\partial \mathcal{L}(\mathbf{X},\mathbf{U})}{\partial \mathbf{V}} &= 2\sigma_\mathbf{e}^{-2}(\mathbf{X}^T\mathbf{U} - \mathbf{V}\mathbf{U}^T\mathbf{U}) \\
\frac{\partial \mathcal{L}(\mathbf{X},\mathbf{U})}{\partial \mathbf{\Sigma}_\mathbf{f}} &= -n\mathbf{\Sigma}_\mathbf{f}^{-1} + \mathbf{\Sigma}_\mathbf{f}^{-1}(\mathbf{U}-\mathbf{Y}\mathbf{B})^T(\mathbf{U}-\mathbf{Y}\mathbf{B})\mathbf{\Sigma}_\mathbf{f}^{-1} \\
\frac{\partial \mathcal{L}(\mathbf{X},\mathbf{U})}{\partial \sigma_\mathbf{e}^2} &= -np\sigma_\mathbf{e}^{-2} + \sigma_\mathbf{e}^{-4}\mathrm{tr}\left((\mathbf{X}-\mathbf{U}\mathbf{V}^T)(\mathbf{X}-\mathbf{U}\mathbf{V}^T)^T\right)
\end{aligned}$$

By setting the conditional expectations of the above items to zero, we have

$$\begin{aligned}
\widehat{\mathbf{B}} &= (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbb{E}_\mathbf{U}(\mathbf{U}|\mathbf{X},\theta^{(i)}), \\
\widehat{\mathbf{V}} &= \mathbf{X}^T\mathbb{E}_\mathbf{U}(\mathbf{U}|\mathbf{X},\theta^{(i)})\left[\mathbb{E}_\mathbf{U}(\mathbf{U}^T\mathbf{U}\ |\mathbf{X},\theta^{(i)})\right]^{-1}, \\
\widehat{\mathbf{\Sigma}_\mathbf{f}} &= \frac{1}{n}\mathbb{E}_\mathbf{U}\left[(\mathbf{U}-\mathbf{Y}\widehat{\mathbf{B}})^T(\mathbf{U}-\mathbf{Y}\widehat{\mathbf{B}})\ |\mathbf{X},\theta^{(i)}\right], \\
\widehat{\sigma_\mathbf{e}^2} &= \frac{1}{np}\mathbb{E}_\mathbf{U}\left[\ \mathrm{tr}((\mathbf{X}-\mathbf{U}\widehat{\mathbf{V}}^T)(\mathbf{X}-\mathbf{U}\widehat{\mathbf{V}}^T)^T)\ |\mathbf{X},\theta^{(i)}\right].
\end{aligned}$$

By substituting the corresponding conditional expectations with the quantities obtained in E step, we have the following explicit expressions of all unconstrained maximizers

$$\begin{aligned}
\widehat{\mathbf{B}} &= (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\Theta_{\mathbf{U}|\mathbf{X}}^{(i)} \\
\widehat{\mathbf{V}} &= \mathbf{X}^T\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}\left(n\Omega_{\mathbf{U}|\mathbf{X}}^{(i)} + {\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}}^T\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}\right)^{-1} \\
\widehat{\mathbf{\Sigma}_\mathbf{f}} &= \frac{1}{n}\left(n\Omega_{\mathbf{U}|\mathbf{X}}^{(i)} + {\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}}^T\Theta_{\mathbf{U}|\mathbf{X}}^{(i)} + \widehat{\mathbf{B}}^T\mathbf{Y}^T\mathbf{Y}\widehat{\mathbf{B}} - \widehat{\mathbf{B}}^T\mathbf{Y}^T\Theta_{\mathbf{U}|\mathbf{X}}^{(i)} - {\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}}^T\mathbf{Y}\widehat{\mathbf{B}}\right) \\
\widehat{\sigma_\mathbf{e}^2} &= \frac{1}{np}\left[\mathrm{tr}\left(\mathbf{X}\mathbf{X}^T\right) - 2\mathrm{tr}\left(\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}\widehat{\mathbf{V}}^T\mathbf{X}^T\right) + n\mathrm{tr}\left(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}}\Omega_{\mathbf{U}|\mathbf{X}}^{(i)}\right)\right. \\
&\quad \left.+\mathrm{tr}\left(\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}\widehat{\mathbf{V}}^T\widehat{\mathbf{V}}{\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}}^T\right)\right]
\end{aligned}$$

where the parameters are estimated from previous iteration.

S step: We standardize the parameter set $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma}_\mathbf{f}}, \widehat{\sigma_\mathbf{e}^2})$ by decomposing $\widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}_\mathbf{f}}\widehat{\mathbf{V}}^T$ as $\mathbf{V}^{(i+1)}\mathbf{\Sigma}_\mathbf{f}^{(i+1)}{\mathbf{V}^{(i+1)}}^T$, and setting $\mathbf{B}^{(i+1)} = \widehat{\mathbf{B}}\widehat{\mathbf{V}}^T\mathbf{V}^{(i+1)}$ and $\sigma_\mathbf{e}^{2(i+1)} = \widehat{\sigma_\mathbf{e}^2}$, as in Section 2.7.1.

As a result, $(\mathbf{B}^{(i+1)}, \mathbf{V}^{(i+1)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}, \sigma_{\mathbf{e}}^{2(i+1)})$ is the set of parameter estimations from the current iteration that satisfies the identifiability conditions.

Stopping rule: As shown in Section 2.7.2, the log likelihoods of the observed data are monotonically nondecreasing with iterations. We evaluate the log likelihood at each iteration and terminate the algorithm when the increase between two iterations is below 10^{-5}.

2.7.4 Proof of Theorem 2.4.1

PROOF. The proof is similar with the proof of Theorem 5.1 in Cook et al. (2010).

The SupSVD model (2.2) can be written as a simple multivariate linear model

$$\mathbf{X} = \mathbf{Y}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where the coefficient matrix $\boldsymbol{\beta}$ and the noise matrix $\boldsymbol{\varepsilon}$ are equal to $\mathbf{B}\mathbf{V}^T$ and $\mathbf{F}\mathbf{V} + \mathbf{E}$ separately. Rows of the residual matrix is i.i.d. from $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Sigma}_{\mathbf{f}}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}$. Let $\mathbf{h} = \begin{pmatrix} \text{vec}(\boldsymbol{\beta}) \\ \text{vech}(\mathbf{\Sigma}) \end{pmatrix}$ denote the true parameters that satisfy the overparameterized structural constraints, and let $\widehat{\mathbf{h}}_{full}$ denote the unconstrained maximum likelihood estimation of the multivariate regression model. From classic asymptotic theories for maximum likelihood we know $\sqrt{n}(\widehat{\mathbf{h}}_{full} - \mathbf{h})$ is asymptotically normally distributed with mean equal to zero and covariance equal to $\mathbf{J}^{-1}$, i.e., the inverse of the Fisher information matrix of $\mathbf{h}$.

$$\mathbf{J} = \begin{pmatrix} \mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}_{\mathbf{Y}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{E}_p^T(\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1})\mathbf{E}_p \end{pmatrix}$$

where $\mathbf{\Sigma}_{\mathbf{Y}} = \lim_{n\to\infty} \mathbf{Y}\mathbf{Y}^T/n$ and $\mathbf{E}_p$ is the expansion matrix relating vech() with vec().

In order to apply Shapiro's theorem, we define a discrepancy function $F(\cdot, \cdot)$ as follows. It is proportional to the log likelihood difference between $\widehat{\mathbf{h}}_{full}$ and any parameter set $\mathbf{h}$ that

satisfies the overparameterized structural constraints.

$$\begin{aligned} F(\widehat{\mathbf{h}}_{full}, \mathbf{h}) &= \operatorname{tr}((\mathbf{X} - \mathbf{Y}\boldsymbol{\beta})^T(\mathbf{X} - \mathbf{Y}\boldsymbol{\beta})\boldsymbol{\Sigma}) + n \log |\boldsymbol{\Sigma}| \\ &\quad -\operatorname{tr}((\mathbf{X} - \mathbf{Y}\widehat{\boldsymbol{\beta}}_{full})^T(\mathbf{X} - \mathbf{Y}\widehat{\boldsymbol{\beta}}_{full})\widehat{\boldsymbol{\Sigma}}_{full}) - n \log |\widehat{\boldsymbol{\Sigma}}_{full}| \end{aligned}$$

It's straightforward to see that $F(\widehat{\mathbf{h}}_{full}, \mathbf{h})$ is nonnegative, equal to 0 if and only if $\mathbf{h} = \widehat{\mathbf{h}}_{full}$. Moreover, $F(\widehat{\mathbf{h}}_{full}, \mathbf{h})$ is twice continuously differentiable in terms of $\mathbf{h}$ and $\widehat{\mathbf{h}}_{full}$. Besides, from the regularity of the normal likelihood we know, there is no neighborhood of $\widehat{\mathbf{h}}_{full}$ such that $F(\widehat{\mathbf{h}}_{full}, \mathbf{h})$ is zero for all $\mathbf{h}$ in it. Therefore, from Shapiro's theorem, we know the minimizer of $F(\widehat{\mathbf{h}}_{full}, \cdot)$, or equivalently, the maximizer of the log likelihood function under the structural constraints, has the asymptotic normality. More specifically,

$$\sqrt{n}(\widehat{\mathbf{h}} - \mathbf{h}) \to_d \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{h}})$$

and $\boldsymbol{\Sigma}_{\mathbf{h}} = \mathbf{P}\boldsymbol{\Gamma}\mathbf{P}^T$, where $\mathbf{P} = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T\mathbf{J}$ is the projection matrix and $\boldsymbol{\Gamma}$ is the asymptotic covariance matrix for $\widehat{\mathbf{h}}_{full}$. Here the matrix $\mathbf{J}$ is the Fisher Information of $\mathbf{h}$ as n goes to infinity, and the matrix $\mathbf{H}$ is the Jacobian matrix of $\mathbf{h}$ with respect to the overparameterized model parameters. The symbol $\dagger$ denotes the Moore-Penrose inverse. Particularly, under normality we know that $\boldsymbol{\Gamma} = \mathbf{J}^{-1}$, so that the asymptotic covariance matrix $\boldsymbol{\Sigma}_{\mathbf{h}}$ can be simplified as $\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$. The derivation of the Jacobian matrix $\mathbf{H}$ follows from basic matrix calculus, which can also be found in Cook et al. (2010). Specifically,

$$\mathbf{H} = \begin{pmatrix} \mathbf{V} \otimes \mathbf{I}_q & (\mathbf{I}_p \otimes \mathbf{B})\mathbf{K}_{pr} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_p(\mathbf{V}\boldsymbol{\Sigma}_{\mathbf{f}} \otimes \mathbf{I}_p) & \mathbf{C}_p(\mathbf{V} \otimes \mathbf{V})\mathbf{E}_r & \operatorname{vech}(\mathbf{I}_p) \end{pmatrix}$$

where $\mathbf{C}_p$ is the $p(p+1)/2 \times p^2$ constant contraction matrix; $\mathbf{E}_r$ is the $r^2 \times r(r+1)/2$ constant expansion matrix; $\mathbf{K}_{pr}$ is the $pr \times pr$ constant commutation matrix.

2.7.5 Proof of Corollary 2.4.1

PROOF. We follow the procedure in Anderson (1963) to prove all parameter estimations are jointly asymptotically normal, and derive the asymptotic covariance for the estimated loading vectors $\sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i)$, where $i = 1, \cdots, r$.

First, we introduce some notations. We know from Theorem 2.4.1 that $\sqrt{n}(\text{vech}(\widehat{\mathbf{\Sigma}}) - \text{vech}(\mathbf{\Sigma})) \to_d \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$, where $\mathbf{\Sigma}_0$ is the $p(p+1)/2 \times p(p+1)/2$ lower corner submatrix of $\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$. We can decompose $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ as

$$\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Delta}}\widehat{\mathbf{\Gamma}}^T, \quad \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Delta}\mathbf{\Gamma}$$

where $\widehat{\mathbf{\Gamma}} = (\widehat{\mathbf{V}}, \widehat{\mathbf{V}}_\perp)$, $\widehat{\mathbf{\Delta}} = \begin{pmatrix} \widehat{\mathbf{\Sigma}_\mathbf{f}} + \widehat{\sigma^2_\mathbf{e}}\mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \widehat{\sigma^2_\mathbf{e}}\mathbf{I}_{p-r} \end{pmatrix}$, $\mathbf{\Gamma} = (\mathbf{V}, \mathbf{V}_\perp)$, $\mathbf{\Delta} = \begin{pmatrix} \mathbf{\Sigma}_\mathbf{f} + \sigma^2_\mathbf{e}\mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \sigma^2_\mathbf{e}\mathbf{I}_{p-r} \end{pmatrix}$. For notation purpose, we write $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_r)$, and $\text{diag}(\mathbf{\Sigma}_\mathbf{f}) = (\sigma^2_{\mathbf{f},1}, \cdots, \sigma^2_{\mathbf{f},r})$. The parameters $(\mathbf{V}, \mathbf{\Sigma}_\mathbf{f}, \sigma^2_\mathbf{e})$ and $(\widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma}_\mathbf{f}}, \widehat{\sigma^2_\mathbf{e}})$ satisfy the identifiability conditions. Following the idea in Anderson (1963), we denote

$$\mathbf{M} \triangleq \sqrt{n}(\mathbf{\Gamma}^T\widehat{\mathbf{\Sigma}}\mathbf{\Gamma} - \mathbf{\Delta}) = \sqrt{n}(\mathbf{\Gamma}^T\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Delta}}\widehat{\mathbf{\Gamma}}^T\mathbf{\Gamma} - \mathbf{\Delta}).$$

It's easily seen that $\mathbf{M}$ is asymptotically normally distributed with asymptotic mean $\mathbb{E}(m_{ij}) = 0$ and asymptotic covariance

$$\begin{aligned}
& \mathbb{E}(m_{ij}m_{gh}) \\
= \; & \mathbb{E}[\mathbf{v}_i^T\sqrt{n}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{v}_j\mathbf{v}_g^T\sqrt{n}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{v}_h] \\
= \; & \mathbb{E}[\mathbf{v}_j^T \otimes \mathbf{v}_i^T\sqrt{n}(\text{vec}(\widehat{\mathbf{\Sigma}}) - \text{vec}(\mathbf{\Sigma}))\mathbf{v}_h^T \otimes \mathbf{v}_g^T\sqrt{n}(\text{vec}(\widehat{\mathbf{\Sigma}}) - \text{vec}(\mathbf{\Sigma}))] \\
= \; & \mathbf{v}_j^T \otimes \mathbf{v}_i^T\mathbb{E}[\sqrt{n}(\text{vec}(\widehat{\mathbf{\Sigma}}) - \text{vec}(\mathbf{\Sigma}))\sqrt{n}(\text{vec}(\widehat{\mathbf{\Sigma}}) - \text{vec}(\mathbf{\Sigma}))^T]\mathbf{v}_h \otimes \mathbf{v}_g \\
= \; & \mathbf{v}_j^T \otimes \mathbf{v}_i^T\mathbf{E}_p\mathbf{\Sigma}_0\mathbf{E}_p^T\mathbf{v}_h \otimes \mathbf{v}_g
\end{aligned}$$

Moreover, we denote $\mathbf{T} \triangleq \mathbf{\Gamma}^T \widehat{\mathbf{\Gamma}}$ and partition it as

$$\mathbf{T} = \begin{pmatrix} \mathbf{v}_1^T \widehat{\mathbf{v}}_1 & \cdots & \mathbf{v}_1^T \widehat{\mathbf{v}}_r & \mathbf{v}_1^T \widehat{\mathbf{V}}_\perp \\ \vdots & & \vdots & \vdots \\ \mathbf{v}_r^T \widehat{\mathbf{v}}_1 & \cdots & \mathbf{v}_r^T \widehat{\mathbf{v}}_r & \mathbf{v}_r^T \widehat{\mathbf{V}}_\perp \\ \mathbf{V}_\perp^T \widehat{\mathbf{v}}_1 & \cdots & \mathbf{V}_\perp^T \widehat{\mathbf{v}}_r & \mathbf{V}_\perp^T \widehat{\mathbf{V}}_\perp \end{pmatrix} = \begin{pmatrix} t_{11} & \cdots & t_{1r} & T_{1\perp} \\ \vdots & & \vdots & \vdots \\ t_{r1} & \cdots & t_{rr} & T_{r\perp} \\ T_{\perp 1} & \cdots & T_{\perp r} & T_{\perp\perp} \end{pmatrix}.$$

Accordingly, we partition $\mathbf{M}$ as

$$\mathbf{M} = \begin{pmatrix} m_{11} & \cdots & m_{1r} & M_{1\perp} \\ \vdots & & \vdots & \vdots \\ m_{r1} & \cdots & m_{rr} & M_{r\perp} \\ M_{\perp 1} & \cdots & M_{\perp r} & M_{\perp\perp} \end{pmatrix}.$$

Following the proof verbatim in Section 2 of Anderson (1963), we know the diagonal values of $\sqrt{n}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})$ are asymptotically normally distributed. In other words, $\sqrt{n}\text{diag}(\widehat{\mathbf{\Sigma}_\mathbf{f}} - \mathbf{\Sigma_f})$ and $\sqrt{n}(\widehat{\sigma_\mathbf{e}^2} - \sigma_\mathbf{e}^2)$ are jointly asymptotically normal. The diagonal blocks of $\mathbf{T}$ have the limiting distribution $\sqrt{n}(t_{ii}^2 - 1) \to_d 0$ $(i = 1, \cdots, r)$ and $\sqrt{n}(T_{\perp\perp} T_{\perp\perp}^T - \mathbf{I}_{p-r}) \to_d \mathbf{0}$. For the off diagonal blocks of $\mathbf{T}$, $\sqrt{n} t_{ij}$ $(i, j = 1, \cdots, r; i \neq j)$ has the same limiting distribution as $m_{ij}/(\sigma_{\mathbf{f},i}^2 - \sigma_{\mathbf{f},j}^2)$; $\sqrt{n} T_{i\perp}$ $(i = 1, \cdots, r)$ has the same limiting distribution as $M_{i\perp}/\sigma_{\mathbf{f},i}^2$; and $\sqrt{n} T_{\perp j}$ $(j = 1, \cdots, r)$ has the same limiting distribution as $M_{\perp j}/\sigma_{\mathbf{f},j}^2$.

In order to get the limiting distribution of $\sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i)$ $(i = 1, \cdots, r)$, we notice that

$$\begin{aligned}
& \sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i) \\
= \; & \sqrt{n}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\widehat{\mathbf{v}}_i - \mathbf{v}_i) \\
= \; & \sqrt{n}(\sum_{j=1}^{r} \mathbf{v}_j\mathbf{v}_j^T\widehat{\mathbf{v}}_i + \mathbf{V}_\perp\mathbf{V}_\perp^T\widehat{\mathbf{v}}_i - \mathbf{v}_i) \\
= \; & \sqrt{n}(\mathbf{v}_i\mathbf{v}_i^T\widehat{\mathbf{v}}_i - \mathbf{v}_i) + \sqrt{n}\sum_{j\le r, j\neq i} \mathbf{v}_j\mathbf{v}_j^T\widehat{\mathbf{v}}_i + \sqrt{n}\mathbf{V}_\perp\mathbf{V}_\perp^T\widehat{\mathbf{v}}_i \\
= \; & \sqrt{n}\mathbf{v}_i(t_{ii} - 1) \\
& + \sqrt{n}(\mathbf{v}_1 \cdots \mathbf{v}_{i-1}, \mathbf{v}_{i+1} \cdots \mathbf{v}_r, \mathbf{V}_\perp)(t_{1i} \cdots t_{(i-1)i}, t_{(i+1)i} \cdots t_{ri}, T_{\perp i}^T)^T \\
= \; & \sqrt{n}\mathbf{v}_i(t_{ii} - 1) + \sqrt{n}\boldsymbol{\Gamma}_{-i}\mathbf{t}_{(-i)i}
\end{aligned}$$

where $\boldsymbol{\Gamma}_{-i}$ is the submatrix of $\boldsymbol{\Gamma}$ without the ith column and $\mathbf{t}_{(-i)i}$ is the ith column of $\mathbf{T}$ without the ith entry. The limiting distribution of the first term is 0. The limiting distribution of the second term can be substituted by the limiting distribution of corresponding $\mathbf{M}$ components. Therefore, we have the following two have the same limiting distribution.

$$\sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i) =_d \boldsymbol{\Gamma}_{-i}\boldsymbol{\Delta}_i\mathbf{m}_{(-i)i}$$

where $\boldsymbol{\Delta}_i$ is the $(p-1) \times (p-1)$ submatrix of $\boldsymbol{\Delta} - (\sigma_{\mathbf{f},i}^2 + \sigma_{\mathbf{e}}^2)\mathbf{I}_p$ without the ith row and ith column, and $\mathbf{m}_{(-i)i}$ is the ith column of $\mathbf{M}$ without the ith entry. From previous derivation, we know the limiting distribution of $\mathbf{m}_{(-i)i}$ is multivariate normal with mean $\mathbf{0}$ and covariance $(\mathbf{v}_i^T \otimes \boldsymbol{\Gamma}_{-i}^T)\mathbf{E}_p\boldsymbol{\Sigma}_0\mathbf{E}_p^T(\mathbf{v}_i \otimes \boldsymbol{\Gamma}_{-i})$. Therefore, we have

$$\sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i) \to_d \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{v}_i})$$

where $\boldsymbol{\Sigma}_{\mathbf{v}_i} = \boldsymbol{\Gamma}_{-i}\boldsymbol{\Delta}_i(\mathbf{v}_i^T \otimes \boldsymbol{\Gamma}_{-i}^T)\mathbf{E}_p\boldsymbol{\Sigma}_0\mathbf{E}_p^T(\mathbf{v}_i \otimes \boldsymbol{\Gamma}_{-i})\boldsymbol{\Delta}_i\boldsymbol{\Gamma}_{-i}^T$, for $i = 1, \cdots, r$.

Lastly, since $\mathbf{B} = \mathbf{B}\mathbf{V}^T\mathbf{V} = \boldsymbol{\beta}\mathbf{V}$ and $\mathbf{V}$ can be expressed as a function of $\boldsymbol{\Sigma}$, $\mathbf{B}$ can be expressed as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. According to the joint asymptotic normality of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, it's obvious that $\sqrt{n}\text{vec}(\widehat{\mathbf{B}} - \mathbf{B})$ is also asymptotically normal.

2.7.6 Two Motivating Examples

Example 1: Let $\mathbf{X}$ be a data matrix with 80 samples and 200 variables. The samples are divided into 4 equal-sized subgroups, which have different means in the first two dimension of $\mathbf{X}$. Specifically,

$$\mathbf{X} = \mathbf{u}_1\mathbf{v}_1^T + \mathbf{u}_2\mathbf{v}_2^T + \mathbf{E},$$

where $\mathbf{v}_1 = (1, 0, 0, \cdots, 0)^T$, $\mathbf{v}_2 = (0, 1, 0, \cdots, 0)^T$, $\mathbf{u}_1$=[rep(16,20),rep(-16,20), rep(0,40)]^T + $\boldsymbol{\varepsilon}_1$, and $\mathbf{u}_2$=[rep(0,40),rep(10,20),rep(-10,20)]^T + $\boldsymbol{\varepsilon}_2$. The notation rep$(a, b)$ denotes a row vector of length b whose entries are all equal to a. The random vectors $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ have i.i.d. entries from $\mathcal{N}(0, 4)$ and $\mathcal{N}(0, 9)$, respectively. The random matrix $\mathbf{E}$ has i.i.d. entries from $\mathcal{N}(0, 16)$. The supervision information $\mathbf{Y}$ is the subgroup index.

This setting simulates the situation where the true underlying structure is partially driven by the supervision information, and partially affected by variations from unknown sources. Figure 2.4 shows the scatter plot of the true score vectors in the first two dimensions as well as the score vectors estimated by the different methods, with the subgroups indicated by different colors and symbols. Clearly, the results from SupSVD are the closest to the underlying truth. SupSVD not only explains a large portion of variation in the data, but also separates the underlying subgroups well. The SVD, although explaining slightly more variations, mixes the subgroups together. In particular, the SVD fails to capture the subgroup structure in the data. The RRR scores, on the other hand, shrink the four subgroups into four points, and do not allow any sample-to-sample variation. This example shows that by incorporating the additional supervision information, SupSVD can better recover the true underlying structure.

Example 2: Let $\mathbf{X}$ be a 210×100 data matrix. The first two dimensions of $\mathbf{X}$ have 4 subgroups, each of which follows a bivariate normal distribution. Specifically, 105 samples are from $\mathcal{N}((-40, 30)^T, \text{diag}(40, 1560))$, and one third of the remaining samples are from $\mathcal{N}((10, -30)^T, \text{diag}(55, 35))$, $\mathcal{N}((40, -30)^T, \text{diag}(120, 120))$, and $\mathcal{N}((70, 0)^T, \text{diag}(60, 20))$ respectively. The other dimensions of $\mathbf{X}$ are i.i.d. $\mathcal{N}(0, 4)$. The supervision information $\mathbf{Y}$ is

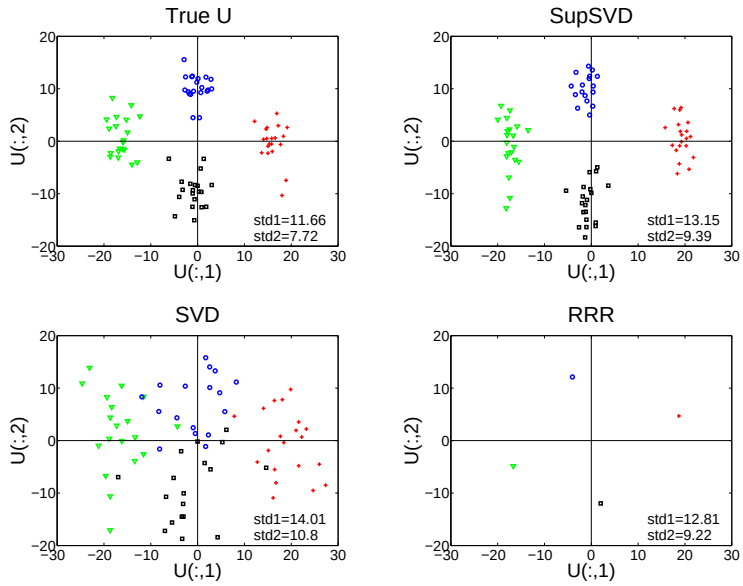


Figure 2.4: Example 1 - Scatter Plots of $\mathbf{U}_1$ and $\mathbf{U}_2$ from Different Methods. The standard deviations of two score vectors are given by std1 and std2.

the subgroup index.

Figure 2.5 shows the first two dimensions of $\mathbf{X}$ with the projected data onto the first two loading vectors obtained by each method. The projection from SupSVD is most similar to the truth, offering a good interpretation: the first direction separates the four subgroups; the second direction explains the variation within each subgroup. By contrast, both SVD and RRR have tilted loading directions. The variances explained by the first two components are similar among all three methods. This example indicates that SupSVD can provide improved interpretability by taking into account the supervision information.

2.7.7 Breast Cancer Data

In this section, we show additional analysis results of the breast cancer data from SVD and RRR. Figure 2.6 shows the scatter plots of the SVD scores. Compared with the SupSVD scores in Figure 2.2 of the main paper, the SVD scores are less interpretable, with samples from different subgroups tilted and intertwined. Figure 2.7 shows the heat maps of the unit-rank SVD approximations. Again, the information is less clear than that in the SupSVD results (Figure 2.3 of the main paper). Figures 2.8 and 2.9 depict the corresponding results

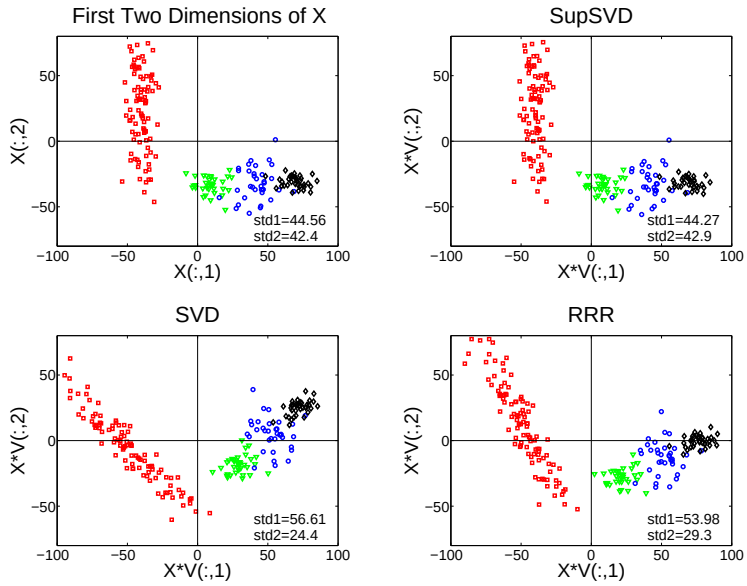


Figure 2.5: Example 2 - Scatter Plots of $\mathbf{XV}_1$ and $\mathbf{XV}_2$. The standard deviations of two projections are given by std1 and std2.

from RRR. Since RRR assumes that the underlying structures are fully driven by supervision, the resulting scores do not present any within-group variations. This is unrealistic and not informative for practical use.

2.7.8 Call Center Data

In this section, we provide an additional application of SupSVD to the call center data previously studied by Shen and Huang (2008a). The data record the number of agent-seeking calls to a banking call center during each 15-minute interval (from 7am to midnight) for 42 consecutive weeks. The goal is to understand the arrival pattern of calls and forecast future call volumes. The raw data can be found in Figure 2.10. We process the data set in the same way as in Shen and Huang (2008a), and focus on the 210 weekdays since the weekends have very different patterns. After imputing missing data, replacing outliers, and applying the square root transformation $\sqrt{N + 1/4}$ where N is the count data matrix, we get the data matrix $\mathbf{X}$ with 210 rows (days) and 68 columns (15-minute intervals). Moreover, we center each column of $\mathbf{X}$ to have mean zero. The scree plot of the singular values of $\mathbf{X}$ suggests the rank to be 4. The supervision data matrix $\mathbf{Y}$ for this case contains the dummy variables for

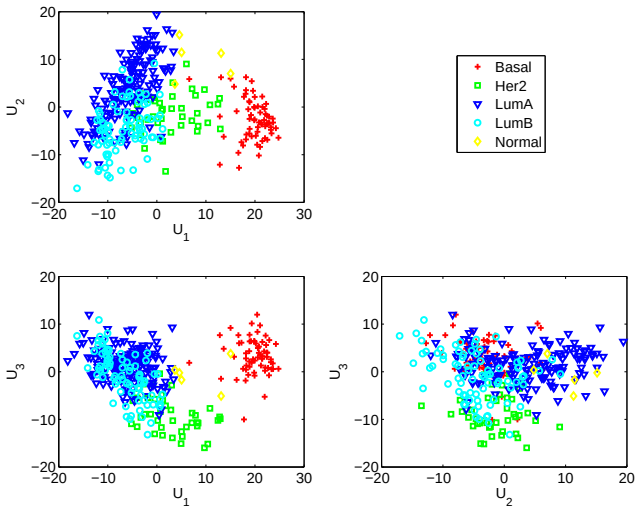


Figure 2.6: Breast Cancer Data - Scatter Plots of SVD Score Vectors.

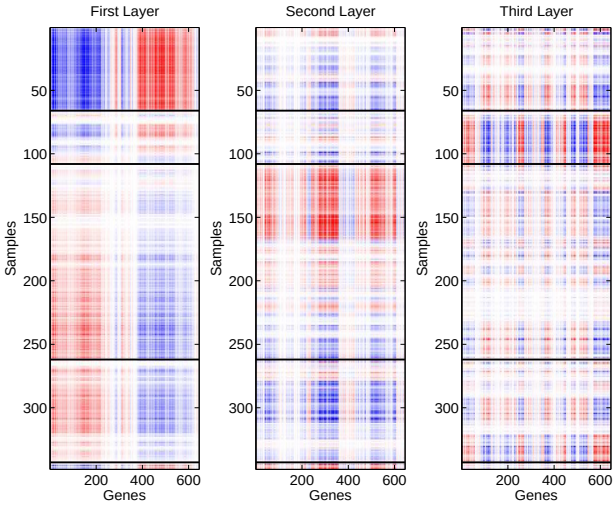


Figure 2.7: Breast Cancer Data - Heat Map of First Three Unit-rank SVD Structures of the Gene Expression Data. The genes are reordered for better visualization.

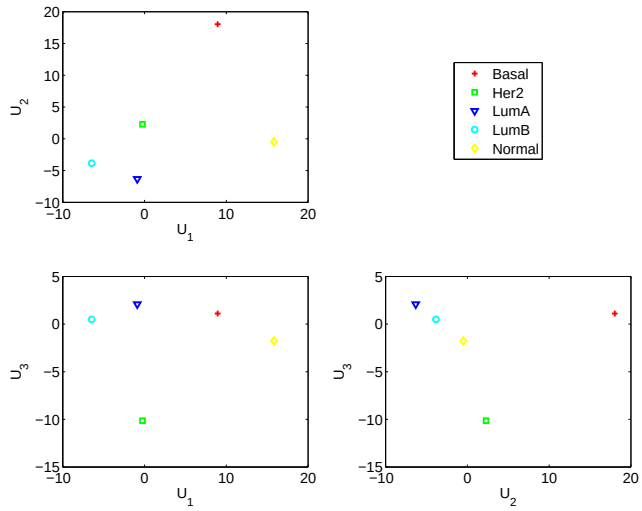


Figure 2.8: Breast Cancer Data - Scatter Plots of RRR Score Vectors.

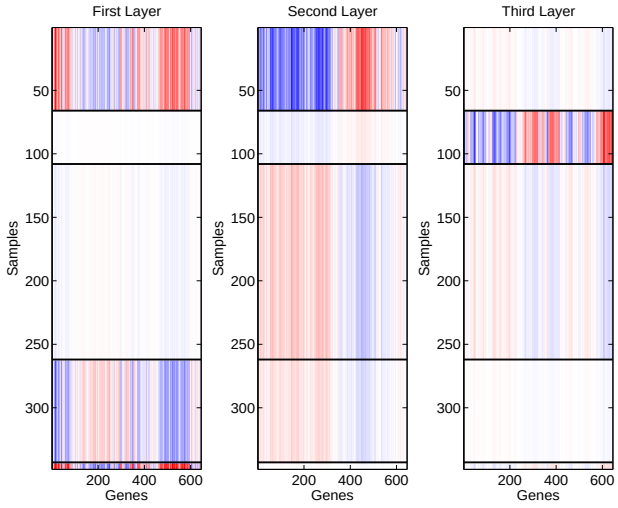


Figure 2.9: Breast Cancer Data - Heat Map of First Three Unit-rank RRR Structures of the Gene Expression Data. The genes are reordered for better visualization.

the day-of-week. Shen and Huang (2008a) point out that the entries of $\mathbf{X}$ are approximately normally distributed, and the weekday effect is a primary factor for the call volume patterns. Therefore, it makes sense to apply SupSVD in this case.

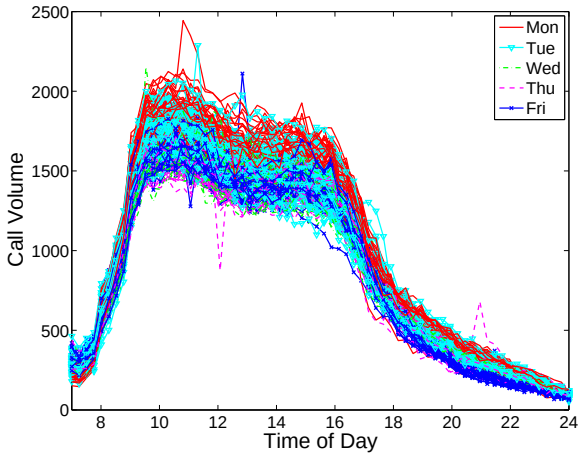


Figure 2.10: Call Center Data - Raw Data. Each curve corresponds to a day, with different markers representing different days of the week.

We apply SupSVD, SVD and RRR to the preprocessed data. The estimates of the first 4 loading vectors are shown in Figure 2.11. The SupSVD estimates have good interpretations. The first loading vector indicates the major variation of call volumes occurs in the daytime. The second loading captures the morning-afternoon contrast, as the curve crosses the zero line at about 12p.m.. The third loading vector changes signs at about 9a.m. and 5p.m., which coincide with common business hours. The forth loading vector explains more subtle variations in the early morning, late morning, afternoon, and late night. In contrast, SVD and RRR loadings do not have such clear interpretations. We remark the improvement of SupSVD is likely due to incorporating the supervision information of the day-of-week effect.

We follow the forecasting procedure proposed by Shen and Huang (2008a) (details can be found therein), but replacing their SVD with our SupSVD. We perform a rolling one-day-ahead forecasting scheme: use 150 days of data as the training set to predict the call volumes for the next day; then roll the forecasting window ahead for one day; repeat for 60 days. For each day, the forecasting accuracy is measured by the root mean squared error (RMSE) and

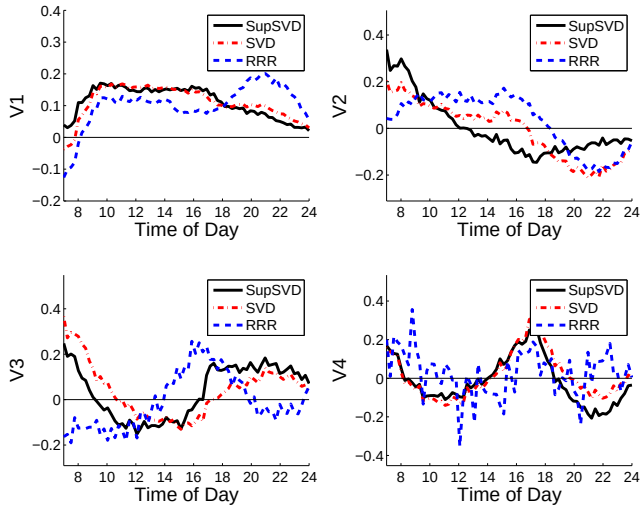


Figure 2.11: Call Center Data - Loading Vectors Estimates from Different Methods. Solid, dashdot and dashed lines represent SupSVD, SVD and RRR respectively.

the mean relative error (MRE) defined as

$$RMSE = \sqrt{\frac{1}{68}\sum_{i=1}^{68}(N_i - \widehat{N}_i)^2}, \qquad MRE = \frac{100}{68}\sum_{i=1}^{68}\frac{|N_i - \widehat{N}_i|}{N_i}$$

where N_i and $\widehat{N}_i$ are the true and predicted call volumes in the ith interval of the next day.

Table 2.2 presents the comparison of forecasting performance between SVD and SupSVD. Clearly, SupSVD outperforms SVD. By using the additional weekday information to guide the dimension reduction, SupSVD captures more essential patterns in the call volumes and has a greater forecasting power.

	RMSE			MRE		
	Q1	Median	Q3	Q1	Median	Q3
SupSVD	**41.1769**	**50.1403**	**60.1383**	**4.4082**	**5.2686**	**6.5211**
SVD	41.5895	50.6250	60.1436	4.4468	5.3321	6.6797

Table 2.2: Call Center Data - Comparison of Forecasting Accuracy between SVD and SupSVD. Results are based on one-day-ahead forecasting for 60 days.

CHAPTER 3: SUPERVISED REGULARIZED PRINCIPAL COMPONENT ANALYSIS

3.1 Introduction

PCA has been widely used in multivariate analysis to extract important features in data. PC loadings usually provide useful interpretation of major variations, while PC scores facilitate follow-up statistical analyses such as clustering and regression. It is a powerful tool for dimension reduction, pattern recognition, and visualization for big data.

This chapter concerns regularized PCA methods, which impose useful structural regularization on PCA, and have been extensively studied in the literature. Special structures like sparsity and smoothness are imposed on the loading vectors to model high-dimensional data with complex structure. For example, functional PCA is used to model functional observations such as temporal data or spatial data (cf. Rice and Silverman (1991), Silverman (1996), Huang et al. (2008), and references therein). In high dimensional situations where most variables are noise and only a few variables are important, sparse PCA is used to simultaneously select variables and capture major variations (cf. Zou et al. (2006), Shen and Huang (2008b), d'Aspremont et al. (2008), and references therein). More recently, some researchers studied two-way extensions of the above one-way regularized PCA methods (see Allen, 2013; Huang et al., 2009; Lee et al., 2010, for example).

Although powerful, the above regularized PCA methods have one limitation in common: they only make use of a single data set, and by default ignore any other measurements collected on the same set of samples. It is now increasingly common that multiple related data sets are available on the same set of samples. In such cases, borrowing information across data sets may lead to recovery of a more interpretable low rank structure. This is especially relevant when the additional measurements, referred to as *supervision information*, can potentially drive underlying patterns within the primary data. For example, in Section

3.5, we are interested in studying expression patterns of a number of yeast genes over two cell cycles. In addition to the gene expression data, we have extra binding information of transcription factors (TFs) for each gene. Since TFs regulate gene expressions biologically (Lee and Young, 2000; Nikolov and Burley, 1997), using TF binding information as supervision when studying expression patterns can lead to a more inherent and meaningful discovery. Another motivating example considered in Section 3.6.3 concerns daily arrival rates of patients to a hospital emergency room. It is of interest to understand patient arrival patterns to better allocate medical resources. In addition to the primary data, i.e., the arrival rates at different time of day over many days, we want to use the day-of-week index as *supervision* to extract day-of-week specific arrival patterns.

Motivated by these applications, we develop a supervised regularized PCA framework that makes use of extra supervision information when doing regularized PCA. We name it the *supervised sparse and functional PCA*, or *SupSFPC*. Supervision, subject to variable selection, directly affects the PC scores, while smooth and sparse structures are imposed on the PC loadings. The SupSFPC framework is very general and flexible. It unifies and generalizes many variants of PCA. In particular, without the supervision, it encompasses regularized PCA methods such as functional PCA and sparse PCA as special cases. Supervision and regularization complement each other under SupSFPC. By smoothing the loading vectors, our method can borrow strength across neighboring variables to reduce noise; with sparsity, the variation of the functional estimate is reduced; supervision indirectly affects the loading vectors to make them more interpretable. Overall, the proposed SupSFPC method can recover an interpretable and accurate low-rank approximation of a primary data set with potential guidance from supervision data.

SupSFPC is related to the SupSVD method in the previous chapter. However, SupSVD cannot accommodate special features of functional or high dimensional data. Incorporating smoothness and sparsity in such data reduces estimation variability and improves interpretability. Furthermore, SupSVD cannot achieve variable selection of the supervision set: when auxiliary data contain irrelevant information to the low-rank structure of the primary data, it is desirable to eliminate unimportant variables and identify crucial driving factors.

For example, in the yeast gene expression application of Section 3.5, researchers are also interested in identifying TFs that regulate cell cycles. The SupSFPC method addresses the above problems through regularization, and covers the method of Li et al. (2015) as a special case. The computational algorithm for SupSFPC is innovative. We combine the EM algorithm with several ascent algorithms, and embed tuning parameter selection in the iterative scheme. Numerical results show high computational efficiency and improvement in interpretability over existing methods.

The rest of the chapter is organized as follows. In Section 3.2, after reviewing the functional PCA model, we propose our new SupSFPC model, followed by the penalized likelihood framework. We then elaborate on connections of the SupSFPC framework to various regularized PCA and supervised PCA methods. In Section 3.3, we develop a computationally efficient algorithm to estimate the model parameters, and briefly discuss tuning parameter selection. We then demonstrate our method using comprehensive simulation studies in Section 3.4 and a real data example in Section 3.5. Additional technical details and numerical studies can be found in the appendix, Section 3.6.

3.2 Model and Likelihood

In this section, we first review the functional PCA model, and then develop the SupSFPC model and introduce a regularized likelihood approach for the model fitting.

3.2.1 Functional PCA Model

We assume that $Z_i(s)$ $(i = 1, \cdots, n)$ are independent realizations of a smooth random function $Z(s)$ with mean function $\mathbb{E}(Z(s)) = \mu(s)$ and covariance function $\text{cov}(Z(s), Z(s')) = G(s, s')$. The index variable s can represent any continuous measure such as time, spatial location, and so on. Its domain $\mathcal{S}$ is assumed bounded. The covariance function can be decomposed as

$$G(s, s') = \sum_{k=1}^{\infty} d_k V_k(s) V_k(s')$$

where $d_1 \geq d_2 \geq \cdots \geq 0$ are the eigenvalues and $V_k(s)$ $(k = 1, 2, \cdots)$ are the corresponding orthogonal unit-norm eigenfunctions. Consequently, by the Karhunen-Loève theorem, the random function $Z(s)$ can be expressed as a linear combination of the eigenfunctions as $Z(s) = \mu(s) + \sum_{k=1}^{\infty} u_k V_k(s)$ where $u_k = \int_{\mathcal{S}} Z(s)V_k(s)\mathrm{d}s$ $(k = 1, 2, \cdots)$ are uncorrelated random variables with mean zero and variance d_k. In particular, $Z_i(s)$ has the expression

$$Z_i(s) = \mu(s) + \sum_{k=1}^{\infty} u_{ik} V_k(s), \tag{3.1}$$

where u_{ik} $(i = 1, \cdots, n)$ are independent realizations of u_k. This is the classical functional PCA model where $(u_{1k}, \cdots, u_{nk})^T$ is the kth score vector, corresponding to the kth loading function $V_k(s)$, $k \geq 1$.

Researchers usually consider the above functional model with measurement errors added, as in for example Yao et al. (2005). Namely, each observed trajectory, denoted by $X_i(s)$, is expressed as

$$X_i(s) = Z_i(s) + e_i(s),$$

where $Z_i(s)$ is the latent random function given in (3.1), and $e_i(s)$ is a measurement error process that is assumed to be uncorrelated at each point with mean zero and variance $\sigma^2_{\mathbf{e}}$, independently identically distributed (i.i.d.) for different observations.

In practice, the majority of variations in data is contained in the subspace spanned by the first few PC loadings in (3.1). Namely, the first few layers of the latent function $Z(s)$ dominate and the rest are negligible. Hereafter, we consider the following rank-r functional PCA model:

$$X_i(s) = \mu(s) + \sum_{k=1}^{r} u_{ik} V_k(s) + e_i(s) = \mu(s) + \mathbf{u}_{(i)}^T \mathbf{V}(s) + e_i(s), \tag{3.2}$$

where $\mathbf{u}_{(i)} = (u_{i1}, \cdots, u_{ir})^T$ is the $r \times 1$ score vector for the ith observation, and $\mathbf{V}(s) = (V_1(s), \cdots, V_r(s))^T$ is the collection of r loading functions. In particular, the finite linear

combination $\mathbf{u}_{(i)}^T \mathbf{V}(s)$ is referred to as the low-rank approximation of the ith demeaned observation $X_i(s) - \mu(s)$.

3.2.2 SupSFPC Model

Let $X_i(s)$ be the ith functional observation from Model (3.2). Let $\mathbf{y}_{(i)}$ be an $q \times 1$ vector containing q auxiliary variables for the ith observation. We assume that $\mathbf{y}_{(i)}$, the supervision data, drives the low-rank structure of $X_i(s)$, the primary data, by directly affecting its PC score vector $\mathbf{u}_{(i)}$ in Model (3.2). In particular, we propose the following multivariate linear model for the scores:

$$\mathbf{u}_{(i)} = \boldsymbol{\beta}_0 + \mathbf{B}^T \mathbf{y}_{(i)} + \mathbf{f}_{(i)} \tag{3.3}$$

where $\boldsymbol{\beta}_0$ is an $r \times 1$ intercept vector, $\mathbf{B}$ is a $q \times r$ coefficient matrix with the rows corresponding to the supervision variables and the columns corresponding to the PC scores, and $\mathbf{f}_{(i)}$ is an independent realization of an $r \times 1$ random vector with mean zero and unknown covariance $\boldsymbol{\Sigma}_{\mathbf{f}}$. For example, in the genetic application of Section 3.5, $X_i(s)$ denotes the gene expression profile of the ith sample, while $\mathbf{y}_{(i)}$ are the corresponding transcription factors.

Model (3.3) consists of a fixed term $\boldsymbol{\beta}_0 + \mathbf{B}^T \mathbf{y}_{(i)}$ and a random term $\mathbf{f}_{(i)}$. The fixed term captures the variations in $\mathbf{u}_{(i)}$ that can be explained by the supervision data $\mathbf{y}_{(i)}$. The random term effectively collects the leftover variations driven by other (unknown) factors. Model (3.3) is flexible enough to adapt to different situations including those where the supervision information is indeed redundant, as we discuss later in Section 3.2.3.

Combining (3.2) and (3.3), we obtain the *supervised functional PCA model*. In particular, we substitute $\mathbf{u}_{(i)}$ in (3.2) with (3.3) and get the following equivalent expression of the model:

$$\begin{aligned} X_i(s) &= \mu(s) + (\boldsymbol{\beta}_0 + \mathbf{B}^T \mathbf{y}_{(i)} + \mathbf{f}_{(i)})^T \mathbf{V}(s) + e_i(s) \\ &= \left[\mu(s) + \boldsymbol{\beta}_0^T \mathbf{V}(s)\right] + \mathbf{y}_{(i)}^T \mathbf{B} \mathbf{V}(s) + \left[\mathbf{f}_{(i)}^T \mathbf{V}(s) + e_i(s)\right]. \end{aligned} \tag{3.4}$$

The first term, $\mu(s) + \boldsymbol{\beta}_0^T \mathbf{V}(s)$, is an intercept term. Without loss of generality, we assume that

$X_i(s)$ and $\mathbf{y}_{(i)}$ are centered at each variable so we can omit this intercept term. The second term, $\mathbf{y}_{(i)}^T\mathbf{B}\mathbf{V}(s)$, is a fixed term that incorporates the supervision information. The third term, $\mathbf{f}_{(i)}^T\mathbf{V}(s)+e_i(s)$, is a random term, with the covariance function being $\mathbf{V}(s)^T\mathbf{\Sigma_f}\mathbf{V}(s')+\sigma_{\mathbf{e}}^2\delta(s-s')$, where $\delta(\cdot)$ is the Dirac delta function. The recovery of the low-rank structure $\mathbf{y}_{(i)}^T\mathbf{B}\mathbf{V}(s)+\mathbf{f}_{(i)}^T\mathbf{V}(s)$ is of primary interest in dimension reduction.

We further generalize Model (3.4) by assuming that $\mathbf{B}$ and $\mathbf{V}(s)$ are potentially sparse. Consequently, we name Model (3.4) the *supervised sparse and functional PCA model*, or the SupSFPC model. Recall that $\mathbf{B}$ is a coefficient matrix to incorporate supervision. Sparsity on $\mathbf{B}$ can effectively identify auxiliary variables that do not provide supervision to the low-rank structure of the primary data. In particular, when $\mathbf{B}$ is a zero matrix, all auxiliary variables are irrelevant to the primary data, and the SupSFPC model reduces to the functional PCA model (3.2). The loading functions in $\mathbf{V}(s)$ can be sparse as well, in the sense that the support of each loading function may not be the entire domain $\mathcal{S}$. Similar to James et al. (2009) where the authors study a regression model with a (potentially sparse) functional predictor, we remark that sparse functions facilitate model interpretations by removing unnatural wiggles around zero. Overall, sparsity is usually a desirable (and sometimes necessary) feature in practice, especially when analyzing high-dimensional data.

As it stands, Model (3.4) is not identifiable. Because, for any $r\times r$ orthogonal matrix $\mathbf{Q}$, we have $\mathbf{B}\mathbf{Q}\mathbf{Q}^T\mathbf{V}(s)=\mathbf{B}\mathbf{V}(s)$ and $\mathbf{f}_{(i)}^T\mathbf{Q}\mathbf{Q}^T\mathbf{V}(s)=\mathbf{f}_{(i)}^T\mathbf{V}(s)$. Moreover, the columns of $\mathbf{B}$ and the entries of $\mathbf{V}(s)$ and $\mathbf{f}_{(i)}$ are subject to scale and order shifts. To rule out this kind of ambiguity, we impose the following identifiability constraints:

(1) The loading functions in $\mathbf{V}(s)$ form an orthonormal basis, i.e., $\int_{\mathcal{S}} V_i(s)V_j(s)\mathrm{d}s=\delta_{ij}$, where δ_{ij} is the Kronecker delta;

(2) The covariance matrix $\mathbf{\Sigma_f}$ is diagonal with distinct positive eigenvalues;

(3) The diagonal values of $\mathbf{\Sigma_f}$ are strictly decreasing.

The orthonormality constraint of the loading functions, also used in the functional PCA model (3.1), facilitates interpretation and rules out scale shift. The diagonality of the covariance matrix with distinct eigenvalues prevents random rotations. The order of the eigenvalues of

$\mathbf{\Sigma_f}$ determines the order of the loading functions in $\mathbf{V}(s)$ and the columns in $\mathbf{B}$. We remark that under the above conditions, the loading functions carry explicit interpretations: the first loading captures the direction where variation in the data from unknown sources is maximized; subsequent loadings are orthogonal to the previous ones and sequentially maximize variations from unknown sources. This is similar with functional PCA where there is no supervision and all variations are from unknown sources.

3.2.3 Penalized Likelihood

In reality, typically we do not observe an entire function but rather at discrete sampling points. In particular, we assume that there are p sampling points in domain $\mathcal{S}$ indexed by $s_1, \cdots, s_p$, which may not be evenly spaced. For notational simplicity, without special notice we generally use $i = 1, \cdots, n$ to index samples, use $j = 1, \cdots, p$ to index discretized points, and use $k = 1, \cdots, r$ to index PC layers.

The discrete observations of the functional data $X_i(s)$ $(i = 1, \cdots, n)$ are collected in an $n \times p$ matrix $\mathbf{X}$, where $x_{ij} = X_i(s_j)$. We discretize $\mathbf{V}(s)$ and $e_i(s)$ in Model (3.4) accordingly as a $p \times r$ loading matrix $\mathbf{V}$ with $v_{jk} = V_k(s_j)$ and an $n \times p$ error matrix $\mathbf{E}$ with $e_{ij} = e_i(s_j)$. We further denote $\mathbf{U} = (\mathbf{u}_{(1)}, \cdots, \mathbf{u}_{(n)})^T$ as an $n \times r$ score matrix, $\mathbf{Y} = (\mathbf{y}_{(1)}, \cdots, \mathbf{y}_{(n)})^T$ as an $n \times q$ supervision data matrix (viewed as fixed in the current context), and $\mathbf{F} = (\mathbf{f}_{(1)}, \cdots, \mathbf{f}_{(n)})^T$ as an $n \times r$ random error matrix. As a result, we obtain the following discretized version of the SupSFPC model (3.4):

$$\begin{cases} \mathbf{X} = \mathbf{U}\mathbf{V}^T + \mathbf{E} \\ \mathbf{U} = \mathbf{Y}\mathbf{B} + \mathbf{F} \end{cases}, \quad \text{or} \quad \mathbf{X} = \mathbf{Y}\mathbf{B}\mathbf{V}^T + \mathbf{F}\mathbf{V}^T + \mathbf{E}. \tag{3.5}$$

The identifiability conditions follow directly from those for the functional version of the model (3.4). Namely, $\mathbf{V}^T\mathbf{V}$ equals to an $r \times r$ identity matrix $\mathbf{I}_r$, and $\mathbf{\Sigma_f}$ is diagonal with positive decreasing eigenvalues.

To fit the SupSFPC model, we adopt a maximum likelihood approach. We assume normality for $\mathbf{E}$ and $\mathbf{F}$. In particular, we assume that e_{ij} is i.i.d. from an univariate normal

distribution $\mathcal{N}(0, \sigma_{\mathbf{e}}^2)$, and $\mathbf{f}_{(i)}$ is i.i.d. from a multivariate normal distribution $\mathcal{N}_r(\mathbf{0}, \mathbf{\Sigma_f})$. In addition, e_{ij} is independent of $\mathbf{f}_{(i)}$. From (3.5), we see that the observation vector $\mathbf{x}_{(i)}$ follows $\mathcal{N}_p(\mathbf{V}\mathbf{B}^T\mathbf{y}_{(i)}, \mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I})$, and different samples are independent. As a result, the log likelihood of the observed data matrix $\mathbf{X}$ is

$$\begin{aligned}\mathcal{L}(\mathbf{X}) = &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det\left(\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p\right) \\ &-\frac{1}{2}\mathrm{tr}\left((\mathbf{X} - \mathbf{Y}\mathbf{B}\mathbf{V}^T)(\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p)^{-1}(\mathbf{X} - \mathbf{Y}\mathbf{B}\mathbf{V}^T)^T\right).\end{aligned}$$

To impose desirable structures (i.e., smoothness and sparsity) on $\mathbf{V}$ and $\mathbf{B}$, we optimize a regularized log likelihood function to estimate the model parameters. Let $\theta \triangleq (\mathbf{B}, \mathbf{V}, \sigma_{\mathbf{e}}^2, \mathbf{\Sigma_f})$ denote the model parameter set and Θ be the parameter space under the identifiability conditions. We propose to solve the following optimization problem:

$$\max_{\theta\in\Theta}\{\mathcal{L}(\mathbf{X}) - \mathcal{P}_f(\mathbf{V}) - \mathcal{P}_s(\mathbf{V}) - \mathcal{P}_s(\mathbf{B})\}, \tag{3.6}$$

where $\mathcal{P}_f(\mathbf{V})$ is the roughness penalty ("f" stands for functionality) on columns of $\mathbf{V}$, and $\mathcal{P}_s(\mathbf{V})$ and $\mathcal{P}_s(\mathbf{B})$ are the sparsity-inducing penalties ("s" stands for sparsity) on entries of $\mathbf{V}$ and $\mathbf{B}$ respectively. We remark by imposing sparsity on $\mathbf{B}$ we also avoid overfitting in the multivariate linear model (3.3) when the dimension of supervision data is high ($q > n$). Therefore, SupSFPC does not have any restrictions on the order of n, p, and q, and is suitable for high dimensional data.

For sparsity, numerous penalties have been proposed and studied in the literature (cf. Fan and Li, 2001; Tibshirani, 1996; Tibshirani et al., 2005; Yuan and Lin, 2006). In this paper, we present out method using the LASSO penalty (Tibshirani, 1996). It can be easily generalized to incorporate other penalties as well. The sparsity-inducing penalties in (3.6) take the following form:

$$\mathcal{P}_s(\mathbf{V}) = \sum_{k=1}^{r}\lambda_k\|\mathbf{v}_k\|_1, \quad \mathcal{P}_s(\mathbf{B}) = \sum_{k=1}^{r}\gamma_k\|\mathbf{b}_k\|_1, \tag{3.7}$$

where $\mathbf{v}_k$ and $\mathbf{b}_k$ are the kth columns of $\mathbf{V}$ and $\mathbf{B}$ corresponding to the kth layer of the low rank structure respectively, and λ_k and γ_k are the corresponding layer-specific tuning parameters.

For smoothness, generalized ℓ_2 penalties are widely used in the literature. Here we consider the elliptical ℓ_2 penalty:

$$\mathcal{P}_f(\mathbf{V}) = \sum_{k=1}^{r} \alpha_k \mathbf{v}_k^T \Omega \mathbf{v}_k, \tag{3.8}$$

where α_k are the layer-specific tuning parameters, and Ω is a fixed $p \times p$ positive semi-definite matrix depending on the sampling points, with the quadratic form $\mathbf{v}_k^T \Omega \mathbf{v}_k$ penalizing differences among adjacent values in $\mathbf{v}_k$. Here we use the same formulation of Ω as in Green and Silverman (1994) which connects nicely with smoothing splines.

The penalized likelihood framework (3.6) is very general and it subsumes many existing methods as we now discuss. If $\mathcal{P}_f(\mathbf{V}) = \mathcal{P}_s(\mathbf{V}) = \mathcal{P}_s(\mathbf{B}) = 0$, i.e., without any structural constraints, it reduces to the SupSVD method of Li et al. (2015). When $\mathcal{P}_s(\mathbf{B}) = \infty$, i.e., $\mathbf{B} = \mathbf{0}$, it reduces to regularized PCA methods: if $\mathcal{P}_f(\mathbf{V}) \neq 0$ and $\mathcal{P}_s(\mathbf{V}) = 0$, it corresponds to functional PCA of Huang et al. (2008); if $\mathcal{P}_f(\mathbf{V}) = 0$ and $\mathcal{P}_s(\mathbf{V}) \neq 0$, it results in sparse PCA of Shen and Huang (2008b); if $\mathcal{P}_f(\mathbf{V}) \neq 0$ and $\mathcal{P}_s(\mathbf{V}) \neq 0$, one obtains the one-way situation of the sparse and functional PCA (SFPC) method of Allen (2013). We also note that the general framework includes many degenerated situations which have not been well studied before. For instance, when $\mathcal{P}_f(\mathbf{V}) = 0$ while $\mathcal{P}_s(\mathbf{V}) \neq 0$ and $\mathcal{P}_s(\mathbf{B}) \neq 0$, the framework reduces to a supervised PCA method with sparsity in $\mathbf{V}$ and $\mathbf{B}$.

We want to comment on situations where the sampling points are different for different samples. For example, in longitudinal studies, patients may follow up at different times and also have distinct time domains. Similar situations have been referred to as sparsely-observed data in functional data analysis(see James et al., 2000; Yao et al., 2005, for example). In such situations, we can think of two possible approaches. For the first one, we can find a set of common grid points that are finer than the irregular sampling points, and treat the functional observations as missing on those grids where no data are observed. Our estimation

algorithm can be extended to incorporate missing values. The second approach is to use basis expansion to interpolate the functional data, and then evaluate them on a set of common sampling points.

3.3 Computational Algorithm

In this section, we propose an algorithm for parameter estimation of the SupSFPC model. For the sake of clarity in describing the estimation algorithm, we first assume that all the tuning parameters, including the rank of the model, are given. We motivate and summarize the algorithm in Section 3.3.1, and derive the algorithm in more detail in Section 3.3.2. Then we briefly discuss the data-driven selection of tuning parameters in Section 3.3.3. Detailed derivation of the tuning parameter selection can be found in Section 3.6.1.

3.3.1 EM Algorithm

Directly optimizing the penalized log likelihood (3.6) with respect to the identifiability constraints is non-trivial. The model parameters are intertwined in the log likelihood $\mathcal{L}(\mathbf{X})$: both the mean and the covariance terms share the parameter matrix $\mathbf{V}$. In addition, the sparsity-inducing penalties are non-differentiable; the feasible region determined by the identifiability conditions is non-convex. We propose an algorithm that effectively combines the EM algorithm with proximal gradient ascent (Beck and Teboulle, 2009; Nesterov, 2005) and block coordinate descent (Ortega and Rheinboldt, 2000) to overcome these computational difficulties.

To motivate the EM formulation, we first note that the hierarchical Model (3.5) contains the PC scores $\mathbf{U}$ as latent variables. It is easily seen that $\mathbf{x}_{(i)}$ and $\mathbf{u}_{(i)}$ are jointly normally distributed, and different samples are independent. The joint log likelihood of the observed data $\mathbf{X}$ and the latent data $\mathbf{U}$ can be decomposed as:

$$\mathcal{L}(\mathbf{X}, \mathbf{U}) = \mathcal{L}(\mathbf{X}|\mathbf{U}) + \mathcal{L}(\mathbf{U}),$$

where the conditional log likelihood of $\mathbf{X}$ given $\mathbf{U}$ is

$$\mathcal{L}(\mathbf{X}|\mathbf{U}) \propto -np \log \sigma_{\mathbf{e}}^2 - \sigma_{\mathbf{e}}^{-2} \mathrm{tr}\left[(\mathbf{X} - \mathbf{U}\mathbf{V}^T)(\mathbf{X} - \mathbf{U}\mathbf{V}^T)^T\right], \tag{3.9}$$

which only depends on $\mathbf{V}$ and $\sigma_{\mathbf{e}}^2$, while the marginal log likelihood of $\mathbf{U}$ is

$$\mathcal{L}(\mathbf{U}) \propto -n \log \det \mathbf{\Sigma_f} - \mathrm{tr}\left[(\mathbf{U} - \mathbf{Y}\mathbf{B})\mathbf{\Sigma_f}^{-1}(\mathbf{U} - \mathbf{Y}\mathbf{B})^T\right], \tag{3.10}$$

only depending on $\mathbf{B}$ and $\mathbf{\Sigma_f}$. Therefore, an EM algorithm can effectively separate the parameter estimation into two parts to simplify the optimization.

The EM algorithm iterates between an E step and an M step. In the $(t+1)$th iteration, the E step is to calculate $\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}[\mathcal{L}(\mathbf{X}, \mathbf{U})]$, where the expectation is taken with respect to $\mathbf{U}$ given $\mathbf{X}$ and $\theta^{(t)} = (\mathbf{B}^{(t)}, \mathbf{V}^{(t)}, \sigma_{\mathbf{e}}^{2(t)}, \mathbf{\Sigma_f}^{(t)})$, the estimated parameter set obtained in the tth iteration. The M step is to maximize $\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}[\mathcal{L}(\mathbf{X}, \mathbf{U})] - \mathcal{P}_f(\mathbf{V}) - \mathcal{P}_s(\mathbf{V}) - \mathcal{P}_s(\mathbf{B})$ with respect to $\theta \in \Theta$, with the penalty terms as in (3.7) and (3.8). We denote the corresponding optimizer as $\theta^{(t+1)}$. After convergence, we obtain a local optimal solution for optimizing the regularized log likelihood (3.6).

Algorithm Summary: Before the detailed technical derivation, we summarize the algorithm with fixed tuning parameters below in Algorithm 2.

Algorithm 2 EM Algorithm for Fitting SupSFPC

1: Initialize model parameters $\theta^{(0)} = (\mathbf{B}^{(0)}, \mathbf{V}^{(0)}, \mathbf{\Sigma_f}^{(0)}, \sigma_{\mathbf{e}}^{2(0)})$;
2: Repeat until convergence:

(a) **E Step:**
- Get critical conditional expectations (3.13), (3.14), and (3.15);

(b) **M Step:**
- Estimate $\mathbf{v}_k^{(t+1)}$ for $k = 1, \cdots, r$ from (3.20);
- Estimate $\sigma_{\mathbf{e}}^{2(t+1)}$ from (3.18);
- Estimate $\mathbf{b}_k^{(t+1)}$ for $k = 1, \cdots, r$ from (3.22);
- Estimate $\mathbf{\Sigma_f}^{(t+1)}$ from (3.19);

3.3.2 Derivation of the EM Algorithm

Since $\mathbf{u}_{(i)}$ and $\mathbf{x}_{(i)}$ are jointly normally distributed, the conditional distribution of $\mathbf{u}_{(i)}$ given $\mathbf{x}_{(i)}$ and $\theta^{(t)}$ is easily derived as $\mathcal{N}_r\left(\boldsymbol{\mu}_{(i)}^{(t)}, \boldsymbol{\Psi}^{(t)}\right)$ where

$$\boldsymbol{\mu}_{(i)}^{(t)} = \left(\mathbf{I}_r + \sigma_{\mathbf{e}}^{2(t)}{\boldsymbol{\Sigma}_{\mathbf{f}}}^{(t)^{-1}}\right)^{-1}\left[\left(\sigma_{\mathbf{e}}^{2(t)}{\boldsymbol{\Sigma}_{\mathbf{f}}}^{(t)^{-1}}\right){\mathbf{B}^{(t)}}^T\mathbf{y}_{(i)} + {\mathbf{V}^{(t)}}^T\mathbf{x}_{(i)}\right], \tag{3.11}$$

$$\boldsymbol{\Psi}^{(t)} = \sigma_{\mathbf{e}}^{2(t)}\left(\mathbf{I}_r + \sigma_{\mathbf{e}}^{2(t)}\boldsymbol{\Sigma}_{\mathbf{f}}^{(t)^{-1}}\right)^{-1}. \tag{3.12}$$

We remark that the conditional expectation of the PC scores for the ith sample is a weighted average of ${\mathbf{B}^{(t)}}^T\mathbf{y}_{(i)}$ and ${\mathbf{V}^{(t)}}^T\mathbf{x}_{(i)}$, where the weight is determined by $\boldsymbol{\Sigma}_{\mathbf{f}}^{(t)}$ and $\sigma_{\mathbf{e}}^{2(t)}$. Namely, in SupSFPC, the PC scores are partially driven by the supervision effect $\mathbf{y}_{(i)}$, and partially affected by the observation $\mathbf{x}_{(i)}$ as in the ordinary PCA.

In the E step, given (3.11) and (3.12), we can derive the explicit expression of $\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}(\mathcal{L}(\mathbf{X},\mathbf{U}))$. As a matter of fact, we do not need to calculate the expectation of the entire joint log likelihood, but rather only the following three terms:

$$\text{first order term:} \quad \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}(\mathbf{U}) \triangleq \boldsymbol{\Gamma}^{(t)} = \left(\boldsymbol{\mu}_{(1)}^{(t)}, \cdots, \boldsymbol{\mu}_{(n)}^{(t)}\right)^T, \tag{3.13}$$

$$\text{second order term:} \quad \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}(\mathbf{U}^T\mathbf{U}) = n\boldsymbol{\Psi}^{(t)} + {\boldsymbol{\Gamma}^{(t)}}^T\boldsymbol{\Gamma}^{(t)}, \tag{3.14}$$

$$\text{quadratic form:} \quad \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}\left[\text{tr}\left(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\right)\right] = n\text{tr}\left(\boldsymbol{\Lambda}\boldsymbol{\Psi}^{(t)}\right) + \text{tr}\left(\boldsymbol{\Gamma}^{(t)}\boldsymbol{\Lambda}{\boldsymbol{\Gamma}^{(t)}}^T\right), \tag{3.15}$$

where $\boldsymbol{\Lambda}$ is any $r \times r$ symmetric matrix.

In the M step, we optimize $\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}[\mathcal{L}(\mathbf{X},\mathbf{U})] - \mathcal{P}_f(\mathbf{V}) - \mathcal{P}_s(\mathbf{V}) - \mathcal{P}_s(\mathbf{B})$ with respect to $\theta \in \Theta$. It is equivalent to the following two separate optimization problems

$$\max_{\mathbf{V},\sigma_{\mathbf{e}}^2:\ \mathbf{V}^T\mathbf{V}=\mathbf{I}} \quad \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}[\mathcal{L}(\mathbf{X}|\mathbf{U})] - \sum_{k=1}^{r}\lambda_k\|\mathbf{v}_k\|_1 - \sum_{k=1}^{r}\alpha_k\mathbf{v}_k^T\Omega\mathbf{v}_k, \tag{3.16}$$

$$\max_{\mathbf{B},\boldsymbol{\Sigma}_{\mathbf{f}}:\ \boldsymbol{\Sigma}_{\mathbf{f}}=\text{diag}(\boldsymbol{\Sigma}_{\mathbf{f}})} \quad \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}[\mathcal{L}(\mathbf{U})] - \sum_{k=1}^{r}\gamma_k\|\mathbf{b}_k\|_1, \tag{3.17}$$

where $\mathcal{L}(\mathbf{X}|\mathbf{U})$ is given by (3.9), and $\mathcal{L}(\mathbf{U})$ is given by (3.10). The notation, $\text{diag}(\boldsymbol{\Sigma}_{\mathbf{f}})$, represents a diagonal matrix whose diagonal entries are the diagonal entries of $\boldsymbol{\Sigma}_{\mathbf{f}}$.

Estimation of $\sigma_{\mathbf{e}}^2$ and $\mathbf{\Sigma_f}$

We take the first order derivative of (3.16) (or (3.17)) with respect to $\sigma_{\mathbf{e}}^2$ (or the diagonal entries of $\mathbf{\Sigma_f}$) and set them to zero, and obtain the analytical expressions

$$\sigma_{\mathbf{e}}^{2(t+1)} = \frac{1}{np}\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}\left\{\text{tr}\left[(\mathbf{X}-\mathbf{U}\mathbf{V}^{(t+1)^T})(\mathbf{X}^T-\mathbf{V}^{(t+1)}\mathbf{U}^T)\right]\right\}, \tag{3.18}$$

$$\mathbf{\Sigma_f}^{(t+1)} = \frac{1}{n}\text{diag}\left\{\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}\left[(\mathbf{U}-\mathbf{Y}\mathbf{B}^{(t+1)})^T(\mathbf{U}-\mathbf{Y}\mathbf{B}^{(t+1)})\right]\right\}, \tag{3.19}$$

where $\mathbf{V}^{(t+1)}$ and $\mathbf{B}^{(t+1)}$ are the optimizers of $\mathbf{V}$ and $\mathbf{B}$ for (3.16) and (3.17) respectively, to be discussed below. In particular, the conditional expectation terms of (3.18) and (3.19) can be obtained using (3.13) to (3.15) from the E step.

Estimation of V

Optimizing (3.16) with respect to $\mathbf{V}$ under the orthogonality constraint is formidable. Instead, we propose to drop the orthogonality and optimize the criterion with respect to the columns of $\mathbf{V}$, one at a time while fixing the others, mimicking a block coordinate descent algorithm. Since the conditional distribution (3.9) of $\mathbf{X}$ given $\mathbf{U}$ is identifiable even without the orthogonality condition, the optimization problem is still well defined. The scheme is similar to the deflation method used in the regularized PCA literature (see Allen, 2013; Hays et al., 2012; Huang et al., 2008; Shen and Huang, 2008b, for example). We remark that the greedy algorithm maintains orthogonality of the columns of $\mathbf{V}$ approximately throughout the EM iterations. Therefore, the column-by-column optimizers serve as a reasonable surrogate of the global optimizer of (3.16).

Given all the parameters except the kth column of $\mathbf{V}$, we can estimate $\mathbf{v}_k^{(t+1)}$ as

$$\mathbf{v}_k^{(t+1)} = \underset{\mathbf{v}_k:\|\mathbf{v}_k\|_2=1}{\arg\min}\ \frac{1}{2}\|\mathbf{v}_k-\boldsymbol{\beta}_k^{(t)}\|_2^2+\lambda_k^{(t)}\|\mathbf{v}_k\|_1+\frac{1}{2}\alpha_k^{(t)}\mathbf{v}_k^T\Omega\mathbf{v}_k, \tag{3.20}$$

where $\boldsymbol{\beta}_k^{(t)} = \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}\left[(\mathbf{X}^T-\mathbf{V}_{-k}^{(t)}\mathbf{U}_{-k}^T)\mathbf{u}_k\right]/c_k^{(t)}$, $\lambda_k^{(t)} = \sigma_{\mathbf{e}}^{2(t+1)}\lambda_k/(2c_k^{(t)})$, $\alpha_k^{(t)} = \sigma_{\mathbf{e}}^{2(t+1)}\alpha_k/c_k^{(t)}$, and $c_k^{(t)} = \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}(\mathbf{u}_k^T\mathbf{u}_k)$. The matrices $\mathbf{U}_{-k}$ and $\mathbf{V}_{-k}^{(t)}$ are the submatrices of $\mathbf{U}$ and $\mathbf{V}^{(t)}$ leaving out the kth column $\mathbf{u}_k$ and $\mathbf{v}_k^{(t)}$, respectively. This setup facilitates parallel computing for the different columns in $\mathbf{V}$. The constants $\boldsymbol{\beta}_k^{(t)}$ and $c_k^{(t)}$ can be calculated from (3.13)

and (3.14). The modified tuning parameters $\lambda_k^{(t)}$ and $\alpha_k^{(t)}$ can absorb the unknown constant $\sigma_{\mathbf{e}}^{2(t+1)}$, and be selected adaptively in a data-driven fashion in each iteration. For now, we treat them as known.

To solve (3.20), we adopt the *proximal gradient ascent* scheme studied in Nesterov (2005) and Beck and Teboulle (2009). We drop the subscripts and the superscripts in (3.20) for simplicity, and the optimization problem becomes $\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} f(\mathbf{v}) + \lambda\|\mathbf{v}\|_1$ where $f(\mathbf{v}) \triangleq \frac{1}{2}\|\mathbf{v} - \boldsymbol{\beta}\|_2^2 + \frac{1}{2}\alpha\mathbf{v}^T\Omega\mathbf{v}$. This optimization is solved by the iterative procedure

$$\mathbf{v}^{(l+1)} = \underset{\mathbf{v}:\|\mathbf{v}\|_2=1}{\arg\min} \left\{ \frac{1}{2} \left\| \mathbf{v} - \left(\mathbf{v}^{(l)} - \frac{1}{L}\nabla f(\mathbf{v}^{(l)}) \right) \right\|_2^2 + \frac{\lambda}{L}\|\mathbf{v}\|_1 \right\}, \tag{3.21}$$

where ∇f is the gradient of f, and L is the Lipschitz constant of ∇f such that $\|\nabla f(\mathbf{a}) - \nabla f(\mathbf{b})\|_2 \leq L\|\mathbf{a} - \mathbf{b}\|_2$ for every $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$. Since $\nabla f(\mathbf{v}) = -\boldsymbol{\beta} + (\mathbf{I} + \alpha\Omega)\,\mathbf{v}$, L is the largest eigenvalue of $\mathbf{I} + \alpha\Omega$. Note that l is the proximal gradient ascent iteration index for estimating one column of $\mathbf{V}$, not to be confused with the EM iteration index t. In particular, we solve (3.21) approximately through the following two steps

$$\begin{aligned} \mathbf{v}^\star &= \mathbf{thres}\left(\mathbf{v}^{(l)} - \frac{1}{L}\nabla f(\mathbf{v}^{(l)}) \;,\; \frac{\lambda}{L} \right), \\ \mathbf{v}^{(l+1)} &= \begin{cases} \dfrac{\mathbf{v}^\star}{\|\mathbf{v}^\star\|_2}, & \mathbf{v}^\star \neq \mathbf{0}, \\ \mathbf{0}, & \mathbf{v}^\star = \mathbf{0}, \end{cases} \end{aligned}$$

where $\mathbf{thres}()$ is a soft-thresholding function that $\mathbf{thres}(\boldsymbol{\beta}, \lambda) \triangleq \text{sign}(\boldsymbol{\beta})(|\boldsymbol{\beta}| - \lambda)_+$.

Estimation of B

To estimate $\mathbf{B}$, we can rewrite (3.17) as r independent unconstrained optimization problems, and obtain each column of $\mathbf{B}$ as

$$\mathbf{b}_k^{(t+1)} = \min_{\mathbf{b}_k} \; \frac{1}{2}\|\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}(\mathbf{u}_k) - \mathbf{Y}\mathbf{b}_k\|_2^2 + \gamma_k^{(t)}\|\mathbf{b}_k\|_1, \tag{3.22}$$

where $\gamma_k^{(t)} = {\sigma_{\mathbf{f},k}^2}^{(t+1)}\gamma_k/2$, with ${\sigma_{\mathbf{f},k}^2}^{(t+1)}$ being the kth diagonal entry of ${\mathbf{\Sigma}_\mathbf{f}}^{(t+1)}$. The unknown constant ${\sigma_{\mathbf{f},k}^2}^{(t+1)}$ is absorbed by the modified tuning parameter $\gamma_k^{(t)}$ that can be

adaptively selected. The vector $\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}(\mathbf{u}_k)$ can be calculated from (3.13).

The optimization problem (3.22) is an univariate LASSO problem with $\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}(\mathbf{u}_k)$ being the response vector, $\mathbf{Y}$ being the $n \times q$ design matrix, and $\mathbf{b}_k$ being the coefficient vector. In addition, both the response vector and the design matrix are column centered, so there is no intercept term. Many methods have been developed to solve (3.22) (cf. Efron et al., 2004; Friedman et al., 2010). Here we use the default *coordinate descent* algorithm in Matlab (Friedman et al., 2010).

3.3.3 Tuning Parameter Selection

The tuning parameters in (3.6) play an important role in balancing the likelihood and the penalties. Note that there are $3r$ tuning parameters in the model. Searching over a $3r$-dimensional grid and refitting the model (potentially multiple times, if one uses cross validation) for each tuning set can be a huge computational burden. Instead, we adopt a nested procedure of selecting tuning parameters introduced by Huang et al. (2009). In each iteration, we find the optimal tuning parameters $\lambda_k^{(t)}$ and $\alpha_k^{(t)}$ for $\mathbf{v}_k^{(t+1)}$ while solving (3.20), and find the best $\gamma_k^{(t)}$ for $\mathbf{b}_k^{(t+1)}$ while solving (3.22). Numerical results illustrate this nested procedure always converges. Theoretical justification of the convergence property of the scheme is an open question.

In particular, when selecting $\lambda_k^{(t)}$ and $\alpha_k^{(t)}$, we assume that they do not interfere with each other and select one while fixing the other as zero. As a result, the selection of $\alpha_k^{(t)}$ is equivalent to selecting the smoothing parameter in a smoothing spline problem. For this selection task, we use leave-one-out cross validation (LOOCV), since the LOOCV score has an analytical form from Green and Silverman (1994) that facilitates fast computation. To select $\lambda_k^{(t)}$, we set it at an asymptotical value since the problem is equivalent to a filtering problem studied in Yang et al. (2014). The asymptotical value can induce appropriate amount of sparsity in $\mathbf{v}_k^{(t+1)}$. The tuning parameter $\gamma_k^{(t)}$ in (3.22) is selected using BIC, which is a popular choice in LASSO problems (Chand, 2012; Wang et al., 2009). The degree of freedom is determined in the same way as in Tibshirani and Taylor (2012). As a result, the algorithm

is computationally efficient and scalable for high dimensional data. Numerical results in Sections 3.4 and 3.5 suggest that the scheme performs well. A more detailed derivation of tuning parameter selection can be found in Section 3.6.1.

So far we have assumed that the rank r of the model is known. In practice, the rank needs to be determined from data. It is reasonable to assume that the rank of the underlying signal of the primary data matrix is inherent. Therefore, all rank selection methods studied in the PCA literature may be used in our framework. In this paper, we adopt a popular approach of using the scree plot of the primary data matrix to determine a proper rank. One can also consider other methods, such as the permutation assessment method in Buja and Eyuboglu (1992) and the bi-cross-validation method in Owen and Perry (2009). More sophisticated rank selection methods for functional data and high dimensional data need further investigation and are beyond the scope of the current paper.

3.4 Simulations

In this section, we compare SupSFPC with SupSVD proposed by Li et al. (2015), one-way SFPC proposed by Allen (2013), and the PCA using comprehensive simulations.

Simulation Settings

Data are generated from the low rank model: $\mathbf{X} = \mathbf{YBV}^T + \mathbf{FV}^T + \mathbf{E}$, which connects to the SupSFPC, SupSVD, one-way SFPC, and PCA models respectively through specific choices of $\mathbf{B}, \mathbf{\Sigma_f}$ and $\mathbf{V}$. Throughout the section, we assume that each entry of $\mathbf{E}$ is i.i.d. standard normal (i.e., $\sigma_{\mathbf{e}}^2 = 1$).

Study I: We first consider an **unit-rank** setup where $n = 200$, $p = 100$, $q = 4$, $r = 1$. The 200×4 supervision matrix $\mathbf{Y}$ is filled with standard normal random numbers and then column centered. The 200×100 primary data matrix $\mathbf{X}$ is also column-centered after being generated. We focus on 4 settings where data are generated from each model respectively:

- **Case 1 (SupSFPC):** The loading vector $\mathbf{V}$ is shown in the left panel of Figure 3.1; the coefficient vector $\mathbf{B}$ is $(3, -3, 5, 0)^T$; $\mathbf{F}$ is a 200×1 random vector where each entry is i.i.d. standard normal (i.e., $\mathbf{\Sigma_f} = 1$).

- **Case 2 (SupSVD):** The parameters $\mathbf{B}$ and $\mathbf{\Sigma_f}$ are the same as in Case 1; the loading vector $\mathbf{V}$ is filled with standard normal random numbers and scaled to have norm one. Namely, there is no smoothness or sparsity in the loading.

- **Case 3 (SFPC):** The vector $\mathbf{V}$ is the same as in Case 1; the coefficient $\mathbf{B} = \mathbf{0}$, which eliminates the supervision effect; each entry of $\mathbf{F}$ is i.i.d. $\mathcal{N}(0, 9)$ (i.e., $\mathbf{\Sigma_f} = 9$).

- **Case 4 (PCA):** The parameters $\mathbf{B}$ and $\mathbf{\Sigma_f}$ are the same as in Case 3, and the loading vector $\mathbf{V}$ is obtained in the same way as in Case 2.

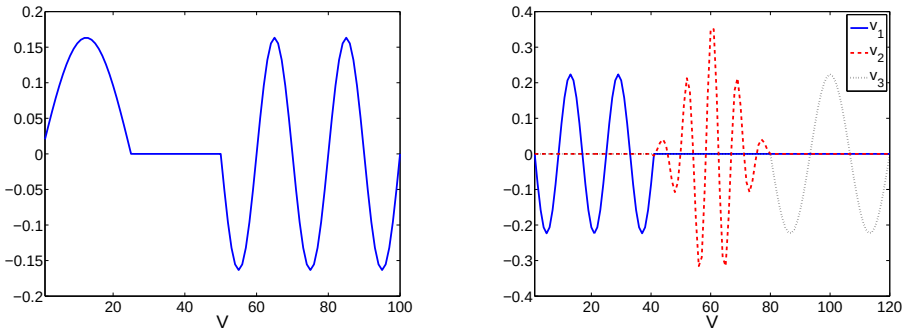


Figure 3.1: Smooth and sparse loading vectors. Left: the loading vector for Cases 1 and 3 in the unit-rank example; right: the loading vectors for Cases 5 and 7 in the rank-3 example.

Study II: We then consider a **multi-rank** setup, where $n = 100$, $p = 120$, $q = 10$, $r = 3$. Again, the 100×10 supervision matrix $\mathbf{Y}$ contains standard normal random numbers and column centered. The 100×120 primary data $\mathbf{X}$ is also column-centered after being generated. Similarly to the unit-rank setup, we consider the following 4 settings:

- **Case 5 (SupSFPC):** The loading vectors in $\mathbf{V}$ are shown in the right panel of Figure 3.1; the 10×3 coefficient matrix $\mathbf{B} = [3, -4, 2, -1, \text{rep}(0, 6); \text{rep}(0, 3), 2, -3, 1, 1, \text{rep}(0, 3); \text{rep}(0, 6), -1, 1, 1, 2]^T$, where rep(a,b) means repeat a b times; the 3×3 covariance matrix $\mathbf{\Sigma_f}$ is a diagonal matrix with diagonal values $(1, 3, 4)$.

- **Case 6 (SupSVD):** The parameters $\mathbf{B}$ and $\mathbf{\Sigma_f}$ are the same as in Case 5; the 120×3 loading matrix $\mathbf{V}$ is filled with standard normal random numbers and normalized to have orthonormal columns.

- **Case 7 (SFPC):** The loading matrix $\mathbf{V}$ is the same as in Case 5; the coefficient matrix $\mathbf{B} = \mathbf{0}$, which eliminates the supervision effect; the covariance matrix $\mathbf{\Sigma_f}$ is diagonal with diagonal values $(16, 9, 4)$.

- **Case 8 (PCA):** The parameters $\mathbf{B}$ and $\mathbf{\Sigma_f}$ are the same as in Case 7, and the loading vectors are obtained in the same way as in Case 6.

Performance Measures

We compare the methods in three aspects, *loading estimation*, *score prediction*, and *low-rank structure recovery*. To evaluate the loading estimation accuracy, we use two criteria, the *mean square error* and the *largest principal angle* (Golub and Van Loan, 2012):

$$MSE_{\mathbf{V}} = \frac{1}{pr}\|\mathbf{V} - \widehat{\mathbf{V}}\|_{\mathbb{F}}^2, \qquad Angle_{\mathbf{V}} = \frac{180}{\pi}\arccos(\min \mathrm{eig}(\mathbf{V}^T\widehat{\mathbf{V}})),$$

where $\|\cdot\|_{\mathbb{F}}$ denotes the Frobenius norm, and $\min \mathrm{eig}(\cdot)$ denotes the minimal eigenvalue. The former characterizes the entry-wise accuracy, and the latter captures the subspace-wise accuracy which is invariant to rotations. For evaluating score prediction and low-rank structure recovery, we use *mean squared prediction errors* defined as:

$$MSPE_{\mathbf{U}} = \frac{1}{nr}\|\mathbf{U} - \widehat{\mathbf{U}}\|_{\mathbb{F}}^2, \qquad MSPE_{\mathbf{U}\mathbf{V}^T} = \frac{1}{np}\|\mathbf{U}\mathbf{V}^T - \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T\|_{\mathbb{F}}^2,$$

where the true scores $\mathbf{U} = \mathbf{Y}\mathbf{B} + \mathbf{F}$, and the predicted $\widehat{\mathbf{U}}$ have different formulas for different methods. For SupSFPC and SupSVD, $\widehat{\mathbf{U}} = \mathbb{E}_{\mathbf{U}|\mathbf{X},\widehat{\theta}}(\mathbf{U})$ where $\widehat{\theta}$ is specific to respective methods; for SFPC and PCA, $\widehat{\mathbf{U}} = \mathbf{X}\widehat{\mathbf{V}}$ where $\widehat{\mathbf{V}}$ is method specific.

Results

For each case, we repeat the simulation 100 times and present the median and the median absolute deviation (MAD) of each performance measurement for all methods in Table 3.1. The results show that SupSFPC outperforms the other methods in all cases, in terms of the considered aspects. One explanation for the superior performance is that SupSFPC is a general framework unifying many existing methods. It automatically adapts to a wide range

of practical situations.

There are several interesting observations in Table 3.1. First, in Case 3 and Case 7 , SFPC surprisingly performs badly in all aspects. This is likely due to an inadequate tuning parameter selection procedure. The original SFPC paper did not provide any guidance on how to set tuning grids for BIC, which is a crucial issue in practice. We consulted with the author and used a suggested tuning grid here. Second, in Case 4 and Case 8, SupSFPC and SupSVD outperform SFPC and PCA in terms of score prediction and low-rank structure recovery. Since the auxiliary data are irrelevant in both cases, the improvement in score prediction must come from the shrinkage effect imposed by $\left(\mathbf{I}+\sigma_{\mathbf{e}}^2\boldsymbol{\Sigma}_{\mathbf{f}}^{-1}\right)^{-1}$ in (3.13). This has been studied from a random matrix point of view by Shabalin and Nobel (2013). Third, in Case 6 where the generating model is SupSVD, the medians of $MSPE_{\mathbf{U}}$ and $MSE_{\mathbf{V}}$ for SupSVD are larger than SupSFPC. In this case the only difference between SupSFPC and SupSVD is that the former does not require strict orthogonality in loading estimation, so we think the improvement comes from this extra flexibility. Nevertheless, both SupSVD and SupSFPC have similar medians of $MSPE_{\mathbf{U}\mathbf{V}^T}$ that are superior to SFPC and PCA. This suggests that the recovery of low-rank structures actually benefits from incorporating auxiliary data.

Note: In Case 2, Case 4, Case 6 and Case 8, we deliberately set the sparsity and smoothness parameters in SupSFPC and SFPC to zero to improve performances. Practically, we usually know when smoothness and sparsity are needed.

3.5 Real Data Example: Yeast Cell Cycle Data

In this section, we demonstrate the advantage of SupSFPC using a yeast cell cycle data set.Two additional real data examples, a government bond yield data set and a hospital emergency room visit data set, are considered in Sections 3.6.2 and 3.6.3.

We consider microarray expression measurements ($\mathbf{X}$) of yeast genes over a certain time period. About 800 cell cycle-related genes are identified in Spellman et al. (1998) through three independent synchronization methods. We consider the data from the α factor based experiment where mRNA levels were measured at every 7 minutes for 18 time points (about

			SupSFPC	SupSVD	SFPC	PCA
$r = 1$		$MSE_{\mathbf{V}}$	**.44e-4** (1.0-5)	1.0e-4 (.86e-5)	2.8e-4 (3.3e-5)	1.0e-4 (.86e-5)
	Case 1	$Angle_{\mathbf{V}}$	**3.8** (.45)	5.8 (.24)	9.6 (.55)	5.9 (.24)
	(SupSFPC)	$MSPE_{\mathbf{U}}$	**1.0** (.07)	**1.0** (.06)	2.0 (.14)	2.1 (.14)
		$MSPE_{\mathbf{UV}^T}$	**.72e-2** (.60e-3)	.99e-2 (.60e-3)	2.3e-2 (1.5e-3)	1.5e-2 (.80e-3)
		$MSE_{\mathbf{V}}$	**1.1e-4** (1.0e-5)	**1.1e-4** (1.0e-5)	**1.1e-4** (1.1e-5)	**1.1e-4** (1.1e-5)
	Case 2	$Angle_{\mathbf{V}}$	**5.8** (.29)	**5.8** (.28)	5.9 (.29)	5.9 (.29)
	(SupSVD)	$MSPE_{\mathbf{U}}$	**1.0** (.07)	**1.0** (.07)	2.0 (.11)	2.0 (.11)
		$MSPE_{\mathbf{UV}^T}$	**1.0e-2** (5.6e-4)	**1.0e-2** (5.6e-4)	1.5e-2 (7.3e-4)	1.5e-2 (7.3e-4)
		$MSE_{\mathbf{V}}$	**.20e-3** (.69e-4)	.60e-3 (.81e-4)	7.4e-3 (9.2e-4)	.60e-3 (.82e-4)
	Case 3	$Angle_{\mathbf{V}}$	**7.6** (1.5)	14 (0.9)	51 (3.5)	14 (1.0)
	(SFPC)	$MSPE_{\mathbf{U}}$	**1.9** (.18)	2.0 (.15)	4.5 (.65)	2.3 (.16)
		$MSPE_{\mathbf{UV}^T}$	**1.3e-2** (1.2e-3)	1.5e-2 (.80e-3)	6.4e-2 (3.7e-3)	1.7e-2 (.90e-3)
		$MSE_{\mathbf{V}}$	**5.8e-4** (7.1e-5)	**5.8e-4** (7.3e-5)	**5.8e-4** (7.1e-5)	**5.8e-4** (7.1e-5)
	Case 4	$Angle_{\mathbf{V}}$	**14** (.88)	**14** (.88)	**14** (.87)	**14** (.87)
	(PCA)	$MSPE_{\mathbf{U}}$	**2.0** (.17)	**2.0** (.16)	2.3 (.17)	2.3 (.17)
		$MSPE_{\mathbf{UV}^T}$	**1.5e-2** (1.0e-3)	**1.5e-2** (1.1e-3)	1.7e-2 (1.2e-3)	1.7e-2 (1.2e-3)
$r = 3$		$MSE_{\mathbf{V}}$	**.60e-3** (.10e-3)	3.7e-3 (1.4e-3)	2.0e-3 (.40e-3)	3.3e-3 (.90e-3)
	Case 5	$Angle_{\mathbf{V}}$	**15** (1.5)	18 (.87)	22 (1.5)	18 (.92)
	(SupSFPC)	$MSPE_{\mathbf{U}}$	**1.9** (.12)	6.0 (2.5)	2.7 (.19)	6.0 (1.5)
		$MSPE_{\mathbf{UV}^T}$	**3.0e-2** (2.2e-3)	4.9e-2 (2.0e-3)	5.8e-2 (4.7e-3)	6.0e-2 (2.6e-3)
		$MSE_{\mathbf{V}}$	**1.8e-3** (.20e-3)	3.9e-3 (1.6e-3)	3.6e-3 (1.0e-3)	3.6e-3 (1.0e-3)
	Case 6	$Angle_{\mathbf{V}}$	**18** (1.1)	**18** (1.0)	**18** (1.0)	**18** (1.0)
	(SupSVD)	$MSPE_{\mathbf{U}}$	**2.2** (.23)	6.6 (3.3)	6.4 (1.6)	6.4 (1.6)
		$MSPE_{\mathbf{UV}^T}$	**4.9e-2** (1.8e-3)	**4.9e-2** (1.7e-3)	6.0e-2 (2.0e-3)	6.0e-2 (2.0e-3)
		$MSE_{\mathbf{V}}$	**1.4e-3** (.20e-3)	5.3e-3 (.70e-3)	4.5e-3 (1.7e-3)	5.3e-3 (.60e-3)
	Case 7	$Angle_{\mathbf{V}}$	**19** (1.3)	32 (2.0)	35 (7.5)	32 (2.0)
	(SFPC)	$MSPE_{\mathbf{U}}$	**2.5** (.17)	3.9 (.68)	3.2 (.49)	4.5 (.58)
		$MSPE_{\mathbf{UV}^T}$	**3.6e-2** (.23e-2)	5.9e-2 (.26e-2)	6.5e-2 (1.2e-2)	6.8e-2 (.32e-2)
		$MSE_{\mathbf{V}}$	**5.3e-3** (7.2e-4)	**5.3e-3** (6.8e-4)	**5.3e-3** (7.1e-4)	**5.3e-3** (7.1e-4)
	Case 8	$Angle_{\mathbf{V}}$	**33** (2.0)	**33** (2.2)	**33** (2.1)	**33** (2.1)
	(PCA)	$MSPE_{\mathbf{U}}$	**3.7** (.51)	3.9 (.55)	4.3 (.49)	4.3 (.49)
		$MSPE_{\mathbf{UV}^T}$	**5.8e-2** (2.9e-3)	**5.8e-2** (3.0e-3)	6.8e-2 (3.0e-3)	6.8e-2 (3.0e-3)

Table 3.1: Median(MAD) of performance measurements for different settings based on 100 simulation runs.

2 hours) covering two cell cycles. In addition to the expression data, we also have ChIP-chip data (Lee et al., 2002) that contain binding information ($\mathbf{Y}$) of 106 TFs for the cell cycle-related genes. We exclude genes with missing values in either expression measurements or TF binding information as in Chen and Huang (2012) and Chun and Keleş (2010), and consider a subset of 542 genes. The data are publicly available in the R package "spls". Figure 3.2 shows the raw expression time series of the 542 cell cycle-related genes.

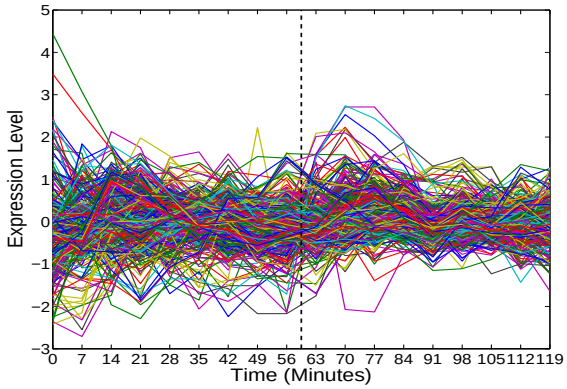


Figure 3.2: Raw expression curves for 542 cell-cycle related genes.

The goal of the yeast cell cycle data analysis is two-fold: 1) understanding the underlying expression patterns of cell cycle-related genes, and 2) identifying transcription factors (TFs) that regulate cell cycles. Below we address both topics simultaneously using SupSFPC. Zhao et al. (2004) primarily focus on the former by projecting the raw time series onto Fourier basis functions with even frequencies and carrying out principal component analysis of the projected data. Chun and Keleş (2010) and Chen and Huang (2012) study the latter by regressing the gene expression data onto TF data through sparse partial least square and sparse reduced rank regression respectively.

The primary data matrix $\mathbf{X}$ contains expression measurements of 542 genes at 18 time points. The supervision data matrix $\mathbf{Y}$ contains binding information of the same genes for 106 TFs. We mean center each time point in $\mathbf{X}$ and each TF in $\mathbf{Y}$. Based on the scree plot of singular values of the column-centered data matrix $\mathbf{X}$, we select the rank to be $r = 4$. The fitting procedure took about an hour on a standard desktop to reach relatively high accuracy (ℓ_2 difference between consecutive estimates of loading vectors below 10^{-3}).

Figure 3.3 compares the loading estimates from four methods: SupSFPC, SupSVD, SFPC and PCA. By taking into account the auxiliary binding information, the SupSFPC loadings are the most interpretable ones. The first and the forth loading vectors of SupSFPC effectively capture periodic patterns of cell cycles without referring to a priori knowledge of true cyclic information as in Zhao et al. (2004). In addition, the second loading mainly presents the variation in the first cell cycle, and the third loading reflects the contrast of the two cycles. The fourth loading also emphasizes the variation in the second cycle.

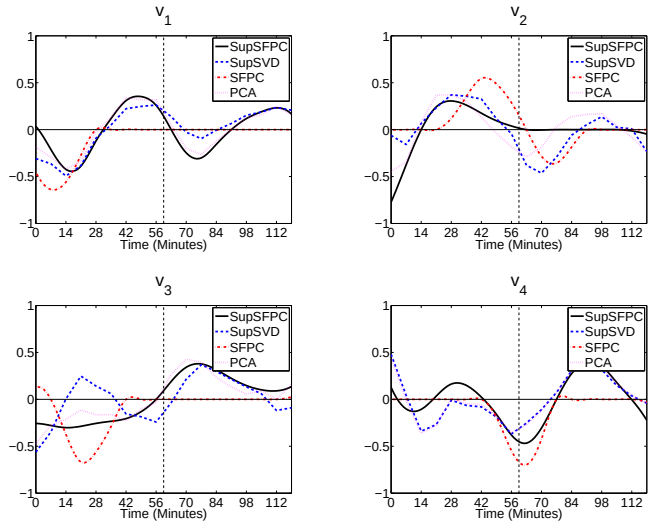


Figure 3.3: The first 4 loading vectors estimated from SupSFPC, SupSVD, SFPC, and PCA.

Figure 3.4 shows the clustering results of the 542 cell cycle-related genes based on SupSFPC scores. We apply a 5-mean clustering approach, where the number of clusters is suggested by Zhao et al. (2004). Different clusters contain genes with different periodic phases. In particular, the genes in the 2nd-5th clusters clearly exhibit different cyclic patterns, similar to the results in Zhao et al. (2004). The genes in the first cluster, on the other hand, do not show strong periodicity, which may need further investigation.

We also investigate the TF activities. Active TFs correspond to the nonzero rows of the estimated supervision coefficient matrix $\widehat{\mathbf{B}}$. Out of the 106 TFs, we identify 32 to be active, with 13 of them being among the 21 experimentally confirmed TFs in Wang et al. (2007). The TF activities for those discovered by SupSFPC are shown in Figure 3.5. Most of the confirmed TFs have clear periodic behavior; among the unconfirmed ones, DOT6, MET4, SFL1, and

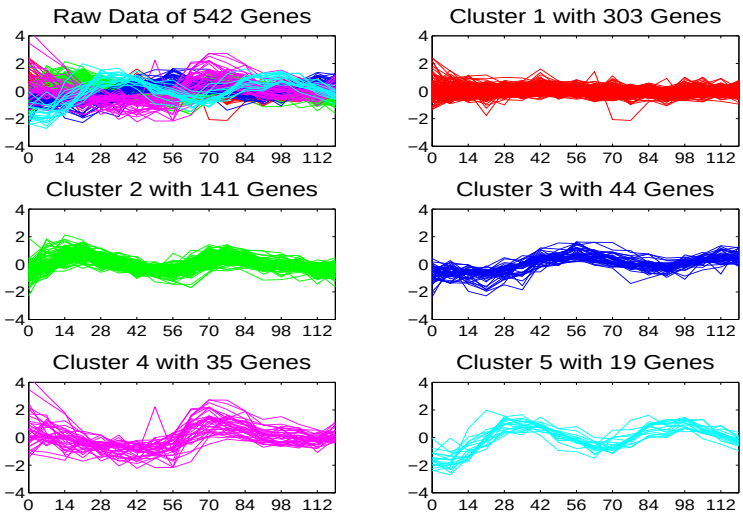


Figure 3.4: Raw gene expression curves clustered into 5 groups based on SupSFPC scores.

YAP5 have the most significant cyclic patterns, which may provide useful guidance to further investigate the regulation effect of TFs on yeast cell cycle.

3.6 Appendix

3.6.1 Tuning Parameter Selection

In this section, we elaborate on the tuning parameter selection procedures for SupSFPC, which are briefly discussed in Section 3.3.3 of the main paper. For computational efficiency, we embed the selection procedures in each EM iteration, as in Huang et al. (2009) and Allen (2013). Before presenting more technical details, we summarize the comprehensive SupSFPC algorithm in Algorithm 3.

3.6.1.1 Select α and λ

The optimization (3.20) involves two tuning parameters: $\alpha_k^{(t)}$ and $\lambda_k^{(t)}$. They control the smoothness and the sparsity of the kth estimated loading vector $\mathbf{v}_k^{(t+1)}$, respectively. To select the best values for both tuning parameters simultaneously, one may search over a

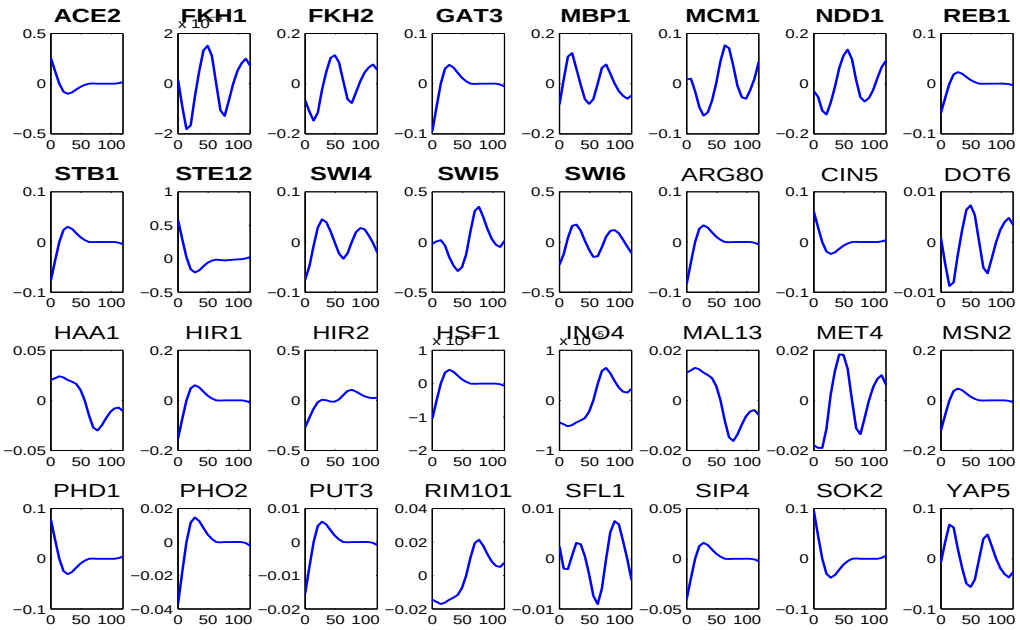


Figure 3.5: TF activities identified by SupSFPC that are related to yeast cell cycles. The first 13 (with bold titles) are experimentally confirmed TFs that are related to cell cycles.

2-dimensional tuning grid and use cross validation methods (Zou and Hastie, 2005) or information theoretic criteria (Allen, 2013). However, the searching procedure is computationally intensive, especially when we do not have a good knowledge of the range of different tuning parameters and have to search over a large grid. Moreover, we need to repeat the procedure for different PC layers in every EM iteration. The overall computational cost can be huge.

As a remedy, we propose to select $\alpha_k^{(t)}$ and $\lambda_k^{(t)}$ separately. In particular, we omit the sparsity penalty (i.e., set $\lambda_k^{(t)} = 0$) when selecting the smoothness parameter $\alpha_k^{(t)}$, and vice versa. An advantage of this approach is that the optimization (3.20) reduces to two well-studied problems: a smoothing spline problem (when $\lambda_k^{(t)} = 0$) and a penalized least square problem (when $\alpha_k^{(t)} = 0$). For each respective problem, the other tuning parameter can be selected adaptively using some computationally efficient methods. We drop the subscripts and the superscripts in (3.20) for simplicity and discuss in more detail below.

Algorithm 3 EM Algorithm for SupSFPC with Adaptive Tuning Selection

1: Initialize model parameters $\theta^{(0)} = (\mathbf{B}^{(0)}, \mathbf{V}^{(0)}, \mathbf{\Sigma_f}^{(0)}, \sigma_{\mathbf{e}}^{2(0)})$;
2: Repeat until convergence:

(a) **E Step:**

- Get critical conditional expectations (3.13), (3.14), and (3.15);

(b) **M Step:**

- **for** $k = 1 \cdots r$ **do**
  - Select $\alpha_k^{(t)}$ from (3.24);
  - Set $\lambda_k^{(t)}$ to be (3.25);
  - Estimate $\mathbf{v}_k^{(t+1)}$ from (3.20);
- **end for**
- Estimate $\sigma_{\mathbf{e}}^{2(t+1)}$ from (3.18);
- **for** $k = 1 \cdots r$ **do**
  - Select $\gamma_k^{(t)}$ from (3.26);
  - Estimate $\mathbf{b}_k^{(t+1)}$ from (3.22);
- **end for**
- Estimate $\mathbf{\Sigma_f}^{(t+1)}$ from (3.19);

When $\lambda = 0$, (3.20) becomes

$$\min_{\mathbf{v}} \ \|\boldsymbol{\beta} - \mathbf{v}\|_2^2 + \alpha \mathbf{v}^T \Omega \mathbf{v} \tag{3.23}$$

where Ω has an expression that is the same as that in smoothing splines (Green and Silverman, 1994). Therefore, (3.23) is a smoothing spline problem. For a given $\alpha > 0$, the closed-form solution of (3.23) is $\widehat{\mathbf{v}}_\alpha = \mathbf{H}_\alpha \boldsymbol{\beta}$, where $\mathbf{H}_\alpha = (\mathbf{I} + \alpha\Omega)^{-1}$ is a $p \times p$ hat matrix. Leave-one-out cross validation (LOOCV) is commonly used to select the smoothing parameter α in smoothing splines. Given an α, we leave out one entry of $\boldsymbol{\beta}$ at a time, and solve (3.23) to get a smooth estimate of $\mathbf{v}$; then we calculate the squared difference between the left-out value in $\boldsymbol{\beta}$ and the corresponding interpolated value in $\mathbf{v}$; we repeat the procedure for all entries of $\boldsymbol{\beta}$ and sum up the squared differences as the LOOCV score for this tuning parameter α. In a candidate tuning set, the one that has the smallest LOOCV score is the optimal tuning parameter.

Solving (3.23) multiple times for each α can be computationally expensive. However,

Green and Silverman (1994) show that the LOOCV score for smoothing spline problems can be obtained analytically by solving the full optimization problem once as

$$\text{LOOCV}(\alpha) = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{\beta_j - \widehat{v}_{\alpha,j}}{1-h_{\alpha,jj}}\right)^2, \tag{3.24}$$

where β_j and $\widehat{v}_{\alpha,j}$ are the jth entry of $\boldsymbol{\beta}$ and $\widehat{\mathbf{v}}_\alpha$, and $h_{\alpha,jj}$ is the jth diagonal entry of $\mathbf{H}_\alpha$. Therefore, LOOCV is an efficient method for tuning parameter selection in smoothing spline. We adopt LOOCV for selecting α in our algorithm. In practice, we can search over a wide range of candidate values at rather low cost.

Given $\alpha = 0$, (3.20) reduces to a penalized least square problem:

$$\min_{\mathbf{v}} \ \frac{1}{2}\|\boldsymbol{\beta} - \mathbf{v}\|_2^2 + \lambda\|\mathbf{v}\|_1,$$

which has an explicit solution $\widehat{\mathbf{v}}_\lambda = \mathbf{thres}(\boldsymbol{\beta}, \lambda)$. Namely, $\lambda > 0$ is the shrinkage amount imposed on $\boldsymbol{\beta}$. Given $\mathbf{U}$, we know from the definition that the vector $\boldsymbol{\beta} = (\mathbf{E}^T + \mathbf{v}\mathbf{u}^T)\mathbf{u}/\|\mathbf{u}\|_2^2 = \mathbf{v} + \mathbf{E}^T\mathbf{u}/\|\mathbf{u}\|_2^2$, where $\mathbf{E}$ is the measurement error matrix in Model (3.5) with i.i.d. entries from $\mathcal{N}(0, \sigma_{\mathbf{e}}^2)$. Namely, $\boldsymbol{\beta}$ can be viewed as the true sparse vector $\mathbf{v}$ plus a noise vector with i.i.d. entries from $\mathcal{N}(0, \sigma_{\mathbf{e}}^2/\|\mathbf{u}\|_2^2)$. To accurately estimate the zero entries in $\mathbf{v}$, a proper threshold is the asymptotically tight upper bound of the expectation of infinity norm of the noise vector, which is $\sqrt{2\log(p)\sigma_{\mathbf{e}}^2/\|\mathbf{u}\|_2^2}$ (Yang et al., 2014). In practice, since both $\|\mathbf{u}\|_2^2$ and $\sigma_{\mathbf{e}}^2$ are unknown, we substitute them with estimates from the previous EM iteration. In particular, the approximate optimal value for $\lambda_k^{(t)}$ is

$$\lambda_k^{(t)} = \sqrt{2\log(p)\sigma_{\mathbf{e}}^{2(t)}/c_k^{(t)}}, \tag{3.25}$$

where $c_k^{(t)} = \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}(\mathbf{u}_k^T\mathbf{u}_k)$. Numerical studies indicate this constant works well.

3.6.1.2 Select γ

The tuning parameter $\gamma_k^{(t)}$ in (3.22) is a LASSO sparsity parameter as we formulate and solve (3.22) as a LASSO problem. Selection of sparsity parameter in a LASSO problem has been well studied in the literature. See, for example, Wang et al. (2009) and Chand (2012). Among data-driven approaches, BIC is a favorable method due to its theoretical merit and fast computation. In particular, since the coordinate descent algorithm can recover the entire solution path efficiently, using BIC to tune LASSO roughly has the same cost as fitting LASSO with a known parameter. Therefore, the BIC procedure for selecting $\gamma_k^{(t)}$ is suitable to be embedded in the EM iteration. For simplicity, we drop the subscripts and the superscripts in the discussion below.

In (3.22), the BIC score for a given tuning parameter γ is defined as

$$\mathrm{BIC}(\gamma) = n \log(\mathrm{MSE}_\gamma) + \mathrm{df}_\gamma \log(n), \tag{3.26}$$

where MSE_γ is the mean residual sum of squares, and df_γ is the degree of freedom of the fitted model corresponding to the tuning parameter γ. The degree of freedom of a LASSO fit has been studied in Zou et al. (2007) for a full-column-rank design matrix, and in Tibshirani and Taylor (2012) for general design matrices. In our case, the design matrix is $\mathbf{Y}$ where columns are potentially linearly dependent. Therefore, we estimate df_γ according to Tibshirani and Taylor (2012) as

$$\widehat{\mathrm{df}}_\lambda = \mathrm{rank}(\mathbf{Y}_{\mathcal{A}(\gamma)}),$$

where $\mathcal{A}(\gamma)$ is a column index set corresponding to nonzero LASSO estimates at γ, and $\mathbf{Y}_{\mathcal{A}(\gamma)}$ is a submatrix of $\mathbf{Y}$ with columns in $\mathcal{A}(\gamma)$. The value that leads to the smallest BIC score is the selected tuning parameter.

3.6.2 Government Bond Yield Data

In this section, we consider the application of SupSFPC to the government bond yield data also studied in Diebold and Li (2006) and Hays et al. (2012). We use the example to illustrate that when auxiliary data are *irrelevant* to the primary data of interest, SupSFPC can adaptively ignore the supervision effect and perform as well as an unsupervised method.

The primary data contain the end-of-month price quotes for U.S. Treasuries, from January 1985 to December 2000 (192 months). For each month, we consider yields on zero coupon bonds of 18 fixed maturities (imputed if missing) of 1.5, 3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, 60, 72, 84, 96, 108, 120 months. Each month is a sample ($n = 192$) and each maturity is a variable ($p = 18$), resulting in a 192×18 primary data matrix $\mathbf{X}$. The 192 raw yield curves of different maturities are shown in Figure 3.6, with random coloring. For each sample, we also have the auxiliary monthly index information, which may or may not influence the underlying structure of $\mathbf{X}$. In particular, we treat the monthly indices (converted to dummy variables) for the 192 months as supervision data $\mathbf{Y}$.

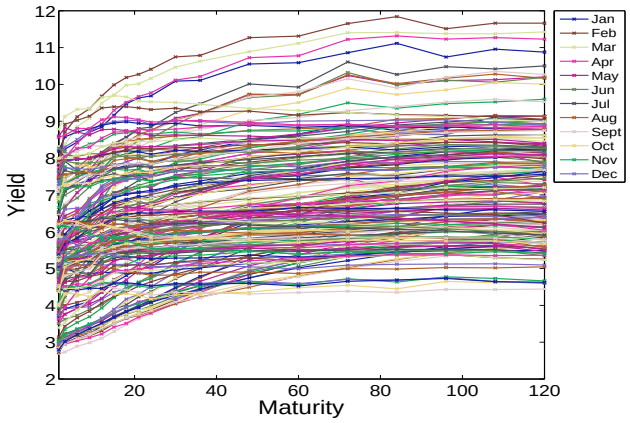


Figure 3.6: Raw yield curves of different maturities from January 1975 to December 2000.

Each column of $\mathbf{X}$ and $\mathbf{Y}$ is centered before applying SupSFPC. We set the rank $r = 2$ as the first 2 principal components of the column-centered $\mathbf{X}$ explain over 99% of the total variation. Then we estimate the SupSFPC model parameters from the data. The fitting procedure took less than 1 second to converge. The estimated supervision coefficient matrix $\widehat{\mathbf{B}}$ is a zero matrix, meaning the auxiliary monthly index data are not relevant to the underlying

structure of the yield data. Namely, the yield curves do not present any strong monthly patterns. This is concordant with our observation from the raw data in Figure 3.6.

We also compare the loading vectors estimated from SupSFPC with those obtained deterministically from the dynamic Nelson-Siegel (DNS) model (Diebold and Li, 2006), which is designed under prior economic theory guidance. Figure 3.7 shows the comparison results. The first panel shows the mean yield curve from the data versus the first loading from the DNS model, both representing a long-term factor. The deviance indicates that the constant loading of the DNS model may not be adequate to capture the overall yield trend at different maturities. The other two panels show the comparison between the 2nd and 3rd loading vectors between SupSFPC and DNS, respectively. From an economic point of view, the two pre-specified DNS loadings possess the interpretation of medium-term and short-term effects respectively. The two SupSFPC loadings have similar shapes with the respective DNS loadings, meaning that SupSFPC captures similar yield curve patterns as in the DNS model. However, we note that SupSFPC only uses information in the data without referring to any economic prior knowledge. Namely, SupSFPC is flexible enough to adapt to the dominant features in the data.

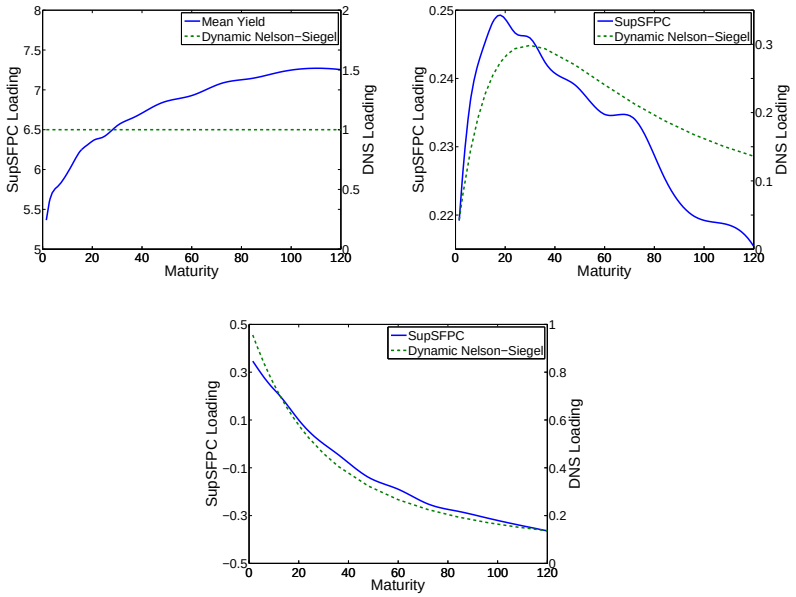


Figure 3.7: Loadings estimated from SupSFPC (solid line) and the pre-specified loadings from the dynamic Nelson-Siegel model (dashed line).

3.6.3 Emergency Room Visit Data

We now analyze the patient arrival rate data from Armony et al. (2011), that contain hourly number of patients arriving to the emergency room (ER) of the Rambam Hospital, Israel for 417 consecutive days (from September 10th, 2006 to October 31th, 2007). The goal is to understand underlying patient arrival patterns to better allocate human and medical resources. The 417 raw arrival rate curves are shown in the 1st panel of Figure 3.8.

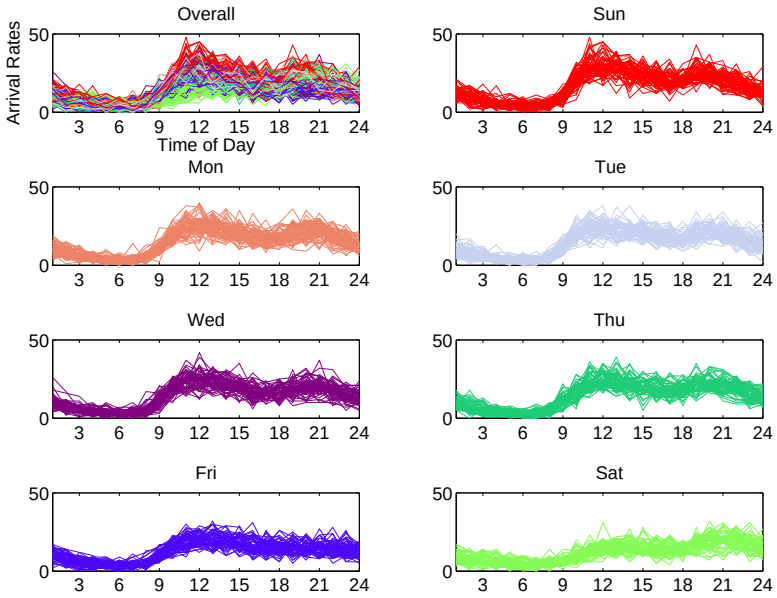


Figure 3.8: Raw arrive rate curves of the hospital ER visit data. The first panel shows the overall curves for 417 consecutive days; the other panels show arrival curves on different days of the week respectively.

Other than the hourly arrival rates, we also know the day-of-week index of each day. In particular, the 2nd-8th panels in Figure 3.8 show arrival curves grouped by the day-of-week index. It can be seen that different days of a week have distinct arrival patterns. Namely, the day-of-week index may be treated as supervision information as it partially drives the underlying structure of the arrival rates.

Each row of the 417×24 primary data matrix contains hourly arrival rates of a day. We apply a square root transformation to the arrival rate data (i.e., $\sqrt{\text{arrival rate} + 1/4}$) to achieve approximate normality (Brown et al., 2005). Then we column center the data

matrix and denote it as $\mathbf{X}$. The supervision data matrix $\mathbf{Y}$ contains 417 day-of-week indices (converted to dummy variables and column centered). The rank is set to be 4 based on the scree plot of the singular values of $\mathbf{X}$.

We apply different methods (SupSFPC, SupSVD, one-way SFPC and PCA) to the data. The fitting procedure of SupSFPC took about 1 minute to converge. Figure 3.9 shows the loading vectors estimated from the methods. By taking into account the auxiliary day-of-week information and allowing regularization, SupSFPC loadings have superior interpretability. The four loadings of SupSFPC capture major variabilities of arrival data from unknown sources after separating the day-of-week effect. They represent large variations of arrival rates at noon, in the evening, overnight, and in the morning, respectively. The day-of-week structure identified by SupSFPC is shown in Figure 3.10. To get the curves in the figure, we transform $\mathbf{Y}\widehat{\mathbf{B}}\widehat{\mathbf{V}}^T$ (where $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{V}}$ are the SupSFPC parameters estimated from the data) back into the original scale by adding the column mean of $\mathbf{X}$ and applying a square transformation. The recovered low-rank structures resemble the (smoothed) average arrival patterns for different weekdays in Figure 3.8.

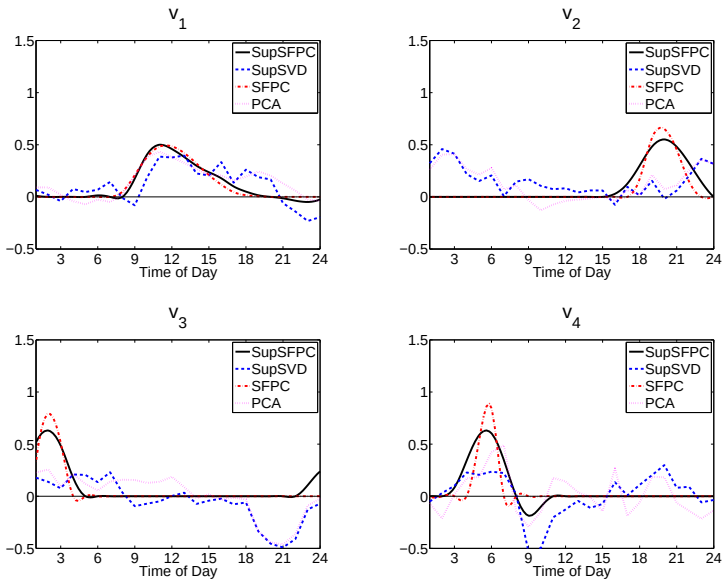


Figure 3.9: The first 4 loading vectors estimated from SupSFPC, SupSVD, SFPC, and PCA.

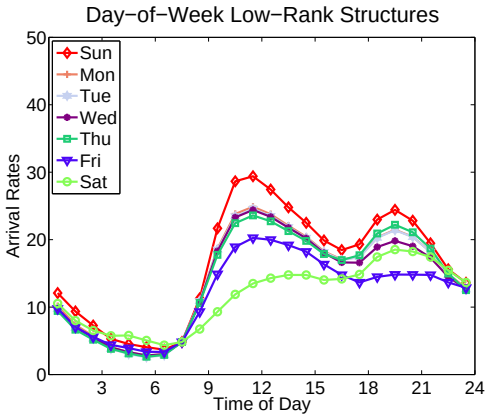


Figure 3.10: The Day-of-Week Structure Identified by SupSFPC.

CHAPTER 4: AN EMPIRICAL BAYES APPROACH FOR MULTIPLE TISSUE EQTL ANALYSIS

4.1 Introduction

In this chapter we introduce and study a multivariate, hierarchical Bayesian model for the simultaneous analysis of eQTLs in multiple tissues, which we call MT-eQTL. The dimension of the MT-eQTL model is equal to the number of tissues. Importantly, we do not seek to describe the full joint relationship between expression and genotype across tissues. Instead, we directly model the vector $\mathbf{z}$ of Fisher transformed correlations between expression and genotype across tissues, after appropriate scaling to account for different degrees of freedom in each tissue. The entries of $\mathbf{z}$ are z-statistics for testing the association between genotype and expression in each tissue. Working with the test statistics on the transformed scale facilitates modeling and interpretation. The upper panel of Figure 4.1b shows a density-based scatter plot of the $\mathbf{z}$-vectors for nerve and skin tissue in the December 2012 release of the GTEx data. The lower panel illustrates the results of the MT-eQTL model: vectors close to the origin for which no eQTLs are detected have been removed, resulting in the central white area; detected eQTLs are colored according to whether an eQTL is detected in both tissues (blue points) or a single tissue (red and green points).

The MT-eQTL model can be expressed in an equivalent, mixture form in which each component corresponds to a binary configuration indicating the presence (1) or absence (0) of an eQTL in each tissue. We adopt an empirical Bayes approach, fitting the MT-eQTL model by maximum likelihood using an EM based algorithm. Throughout we restrict attention to local (sometimes referred to as 'cis') gene-SNP pairs, for which the SNP is within a fixed genomic distance of the coding region of the gene.

We briefly describe some of the key features of the MT-eQTL model. A detailed description is given in Section 4.2. The model explicitly captures patterns of variation in the presence

or absence of eQTLs, as well as the heterogeneity of effect sizes across tissues. In complex multi-tissue data like that from GTEx, the number of samples can vary substantially from tissue to tissue, and the sets of donors for different tissues can exhibit different degrees of overlap. The MT-eQTL model is rich enough to accommodate both of these features. Another important aspect of complex multi-tissue data is that effect sizes in different tissues may be correlated. Correlations in effect sizes arise from biological factors (for example, the underlying relationships among tissues), and are reflected in the correlation structure of the vector $\mathbf{z}$. The correlation structure of $\mathbf{z}$ also reflects experimental factors such as donor overlap among tissues. The MT-eQTL model explicitly accounts for both sources of correlation in an identifiable way. Lastly, the MT-eQTL model has the desirable property of being marginally consistent: roughly speaking, the mixture model for a subset of tissues can be obtained from the full mixture model via marginalization.

Fitting of the MT-eQTL model from the $\mathbf{z}$-vectors of local gene-SNP pairs is carried out via empirical Bayes using an approximate EM algorithm. Fitting is fast enough to accommodate the full analysis of real data sets on a desktop computer. After fitting, the MT-eQTL model provides, for any given $\mathbf{z}$-vector, posterior probabilities for every binary configuration of eQTL absence (0) or presence (1) across tissues. Using the fitted model, we define the local false discovery rate of a gene-SNP pair to be the posterior probability of the zero configuration (no eQTL in any tissue) given its vector of $\mathbf{z}$-statistics. We test for gene-SNP pairs having an eQTL in some tissue by adaptive thresholding of the local false discovery rates. Assessment of tissue specificity can be obtained from the posterior probabilities of non-zero configurations. The procedure is readily generalized to more general hypothesis testing settings.

4.1.1 Related Work

Research on multi-tissue eQTLs is relatively new, with early published work dating from 2007. Most existing multi-tissue analyses extract eQTLs individually from each tissue and then apply post-hoc procedures to assess commonality and specificity. Dimas et al. (2009) and Heinzen et al. (2008) consider the simple pairwise overlap of single tissue eQTL discoveries.

Ding et al. (2010) proposed a procedure to measure eQTL overlap that accounts for differences in statistical power between data sets for individual tissues. Fu et al. (2012) proposed a resampling based procedure to assess the tissue-specificity of cis-eQTLs. Bullaughey et al. (2009) examined the gene-SNP associations in five human primary tissues of eQTLs with large effect sizes in lymphoblastoid cell lines. A similar idea is implemented in Nica et al. (2011): given a set of gene-SNP pairs with small p-values in one tissue, the p-values of these same pairs are examined in other tissues to assess enrichment of significant associations. In addition, several meta-analysis based approaches have been applied to integrate eQTL results for different tissues, *cf.* Brown et al. (2013) and Xia et al. (2012).

The papers cited above provide exploratory studies of eQTLs in multiple tissues, or pairwise conditional analysis of eQTLs declared significant in an initial tissue. However, they do not address the *ab-initio* statistical analysis of multi-tissue data in a manner that fully utilizes the data. Gerrits et al. (2009) used an ANOVA model to jointly analyze gene-SNP associations across tissues, with eQTL configurations assigned according to effect sizes in different tissues. Petretto et al. (2010) proposed a sparse Bayesian regression model in which gene expression in different tissues is treated as a multivariate response, and SNPs are treated as predictors; the presence and specificity of eQTLs are captured by a sparse coefficient matrix. Following Wen et al. (2014), Flutre et al. (2013) proposed a Bayesian framework for the joint analysis of eQTLs across tissues. They use a linear model to capture gene-SNP association in each tissue, and place a prior distribution on the coefficients subject to a latent indicator of whether or not it is an eQTL. Each of these methods uses permutation based procedures to control and calculate false discovery rates., which is computationally burdensome when dealing with millions of gene-SNP pairs and multiple tissues. In addition, these methods assume that each tissue has samples from an identical set of individuals; as noted above, in many cases the set and number of donors varies from tissue to tissue.

In recent work, Sul et al. (2013) proposed a"Meta-Tissue" method that combines linear mixed models and meta-analysis. The linear mixed model captures gene-SNP correlations across tissues and accounts for partial overlap among donors. Meta-analysis is used to address detection of eQTLs in multiple tissues, but the model does not use an explicit indicator vector

for eQTLs across tissues, making assignment of tissue specificity less straightforward than with other methods. Moreover, their hypothesis testing procedure does not make direct use of the alternative distribution, which may lead to a reduction in statistical power.

4.1.2 Outline

The MT-eQTL model is described in the next section. The modified EM algorithm used to fit the model is described in detail in Section 4.3. Section 4.4 describes the application of MT-eQTL to multi-tissue inference, including eQTL detection (both in all tissues and in a subset of tissues) using the local false discovery rate, and the determination of tissue specificity. Section 4.5 presents the results of a simulation study with four tissues. Section 4.6 is devoted to the analysis of new data from the GTEx initiative consisting of nine human tissues with sample sizes ranging from 83 to 156. Technical proofs can be found in the appendix, Section 4.8.

4.2 The MT-eQTL Model

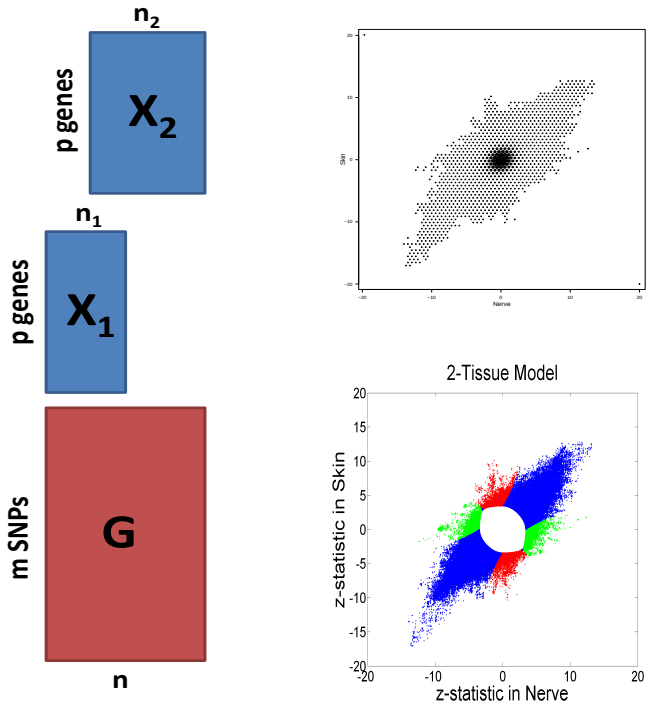
In this section we describe the MT-eQTL model in detail, beginning with a general description of multi-tissue data, and a detailed account of the multivariate z-statistics on which the model is based.

4.2.1 Format of Multi-Tissue eQTL Data

The general data format for the multi-tissue eQTL problem is as follows. For each of n donors we have full genotype information, and measurements of gene expression in at least one of K tissues. We assume that the same array platform is used for measurements of genotype, and similarly for expression.

Let $\mathbf{G}$ be an $m \times n$ matrix containing the measured genotype of each donor in the study at m genetic loci that are the sites of single nucleotide polymorphisms (SNPs). Each column of $\mathbf{G}$ corresponds to a donor, and each row corresponds to a locus/SNP. The measured transcript levels for tissue k are contained in a $p \times n_k$ matrix $\mathbf{X}_k$, where p is the number of

measured transcripts, and $n_k \leq n$ is the number of donors from which samples of tissue k are available. Each column of $\mathbf{X}_k$ has an identifier indicating the donor associated with the measurements in that column. In general, the number of donors n_k can vary widely among tissues, and even if two tissues have similar numbers of samples, they may have relatively few common donors. The data available for the purposes of multi-tissue eQTL analysis has the form $(\mathbf{G}, \mathbf{X}_1, \ldots, \mathbf{X}_K)$. Figure 4.1a gives an illustration of the typical data format with two tissues.



(a) Raw data format

(b) MT-eQTL input and output

Figure 4.1: (a) Illustration of the typical data format with two tissues. Genotype data G is available for m SNPs and each of n samples. Expression measurements are available for p genes; sample sets for different tissues may not be the same. (b) Scatter plots of z-statistics for nerve and skin: for all local gene-SNP pairs (top), and for significant local gene-SNP pairs with tissue specificity by color (bottom).

4.2.1.1 Data Preprocessing and Covariate Adjustment

In most cases eQTL analysis is preceded by several preprocessing steps and covariate adjustment. The genotype data matrix $\mathbf{G}$ consists of values 0, 1, and 2, typically coded as the number of minor allele variants; SNPs with too few minor allele instances are often discarded. Expression measurements may be obtained from array-based platforms or from RNA-Seq tag counts. Lowly expressed genes are typically dropped from the analysis.

Genotype and expression data may contain confounding factors. Some confounders, such as gender, are observed, while others are of unknown technical or biological origin. To identify the unknown confounding factors, most studies use principal components, surrogate variables (Leek and Storey, 2007), or PEER cofactors (Stegle et al., 2012) as covariates. We assume that the expression data and genotype data have been residualized for the confounders, so the comparison of these residualized quantities are partial correlations adjusted for covariates. The degrees of freedom lost in fitting the covariates is accounted for in computing the association between expression and genotype.

4.2.2 Multivariate z-Statistic from Single Tissue Correlations

Denote a measured transcript by $i \in \{1, \ldots, p\}$ and a measured genotype by $j \in \{1, \ldots, m\}$. We focus on a subset Λ of the full index set $\{1, \ldots, p\} \times \{1, \ldots, m\}$ that consists of pairs (i, j) such that SNP j is located within a fixed distance (usually 100 Kilobases or 1 Megabase) of the transcription start site of gene i.

Let $\lambda = (i, j)$ be a gene-SNP pair of interest, and let k be a tissue for which measurements of transcript i are available. Let $r_{\lambda k}$ and $\rho_{\lambda k}$ denote, respectively, the sample and population correlation of transcript i and SNP j in tissue k. Note that the sample correlation $r_{\lambda k}$ depends only on the n_k measurements from donors of tissue k. The vector of correlations $\mathbf{r}_\lambda = (r_{\lambda 1}, \ldots, r_{\lambda K})$ captures the association between the expression of transcript i and the value of genotype j in each of the K tissues. Relationships between different tissues will be reflected in correlations between the entries of $\mathbf{r}_\lambda$. These features make $\mathbf{r}_\lambda$ a natural starting point for a multi-tissue eQTL model.

In order to construct a multivariate model for the correlations $\mathbf{r}_\lambda$, it is convenient to work in a Gaussian setting. To this end, let

$$\mathbf{h}(\mathbf{r}_\lambda) \;=\; \big(h(r_{\lambda 1}), \ldots, h(r_{\lambda K})\big)$$

be the vector obtained by applying the Fisher transformation

$$h(r) = \frac{1}{2}\log\Big(\frac{1+r}{1-r}\Big)$$

to each component of $\mathbf{r}_\lambda$. Let

$$\mathbf{d}^{1/2} := (\sqrt{d_1 - 3}, \ldots, \sqrt{d_K - 3})$$

be a scaling vector, where d_k is the degrees of freedom for $\mathbf{X}_k$ and $\mathbf{G}$, equal to n_k minus the number of covariates used to correct genotype and expression for samples in tissue k. Finally, define the vector

$$\mathbf{z}_\lambda \;=\; \mathbf{d}^{1/2} \cdot \mathbf{h}(\mathbf{r}_\lambda) \tag{4.1}$$

where $\mathbf{u} \cdot \mathbf{v}$ denotes the Hadamard (entry-wise) product of vectors $\mathbf{u}$ and $\mathbf{v}$.

Consider a random vector $\mathbf{Z}_\lambda$ derived in the same fashion as $\mathbf{z}_\lambda$ from random data $(\mathbf{G}, \mathbf{X}_1, \ldots, \mathbf{X}_K)$. We assume that the expression measurements $\mathbf{X}_k$ are approximately normal. Standard arguments for the Fisher transformation (Winterbottom, 1979) show that $h(r_{\lambda k})$ is approximately normal with mean $h(\rho_{\lambda k})$ and variance $(d_k - 3)^{-1}$. By a routine multivariate extension of this fact, $\mathbf{Z}_\lambda$ is approximately normally distributed with mean

$$\boldsymbol{\mu}_\lambda \;=\; \mathbf{d}^{-1/2} \cdot \mathbf{h}(\boldsymbol{\rho}_\lambda).$$

The variance stabilizing property of the Fisher transformation and our choice of scaling ensures that the variance of each entry $Z_{\lambda k}$ of $\mathbf{Z}_\lambda$ is close to one, regardless of $\boldsymbol{\rho}_\lambda$. In particular, if the true correlation $\rho_{\lambda k}$ between transcript i and SNP j for tissue k is zero, then $Z_{\lambda k}$ is approximately standard normal. Thus the k-th entry of the observed vector $\mathbf{z}_\lambda$ is a z-statistic

for testing $\rho_{\lambda k} = 0$ vs. $\rho_{\lambda k} \neq 0$. Importantly, the components of $\mathbf{Z}_\lambda$ need not be independent, even when all the true correlations $\rho_{\lambda k}$ are zero. Capturing this dependence is a key feature of the MT-eQTL model, which is described in detail below.

4.2.3 Hierarchical Model

Let $\lambda = (i, j)$ be a gene-SNP pair in Λ. MT-eQTL is a multivariate, hierarchical Bayesian model for the random vector $\mathbf{Z}_\lambda$. In detail, we assume that

$$\begin{aligned} \mathbf{Z}_\lambda \,|\, \boldsymbol{\mu}_\lambda &\sim \mathcal{N}_K\left(\boldsymbol{\mu}_\lambda, \Delta\right) && (4.2) \\ \boldsymbol{\mu}_\lambda &= \boldsymbol{\Gamma}_\lambda \cdot \boldsymbol{\alpha}_\lambda && (4.3) \\ \boldsymbol{\Gamma}_\lambda &\sim \mathbf{p} \text{ on } \{0,1\}^K && (4.4) \\ \boldsymbol{\alpha}_\lambda &\sim \mathcal{N}_K(\boldsymbol{\mu}_0, \Sigma), \text{ independent of } \boldsymbol{\Gamma}_\lambda && (4.5) \end{aligned}$$

The mean vector $\boldsymbol{\mu}_\lambda$ contains the effect sizes for the relationship between transcript i and SNP j in each tissue. The $K \times K$ covariance matrix Δ is constrained to have diagonal entries equal to one, reflecting the variance stabilization of the Fisher transformation, and the scaling in (4.1). The off-diagonal entries of Δ capture correlations among the entries of $\mathbf{Z}_\lambda$ that are due to commonalities among tissues that arise from the underlying sampling process, for example, correlations resulting from shared donors among a pair of tissues.

We assume that the mean vector $\boldsymbol{\mu}_\lambda$ of $\mathbf{Z}_\lambda$ is equal to the entrywise product of a multinormal random vector $\boldsymbol{\mu}_\lambda$ and a vector $\boldsymbol{\Gamma}_\lambda$ with binary entries. The indicator vector $\boldsymbol{\Gamma}_\lambda$ determines the presence ($\Gamma_{\lambda k} = 0$) or absence ($\Gamma_{\lambda k} = 1$) of an association between transcript i and SNP j in tissues $k = 1, \ldots, K$. The strength of an association, when present, is determined by the corresponding component of $\boldsymbol{\alpha}_\lambda$. The covariance matrix Σ of $\boldsymbol{\alpha}_\lambda$ captures tissue specific variation in effect sizes, and correlations among effect sizes that reflect biological commonalities between tissues. The mean vector $\boldsymbol{\mu}_0$ of $\boldsymbol{\alpha}_\lambda$ captures the average effect sizes across tissues. In practice we usually set $\boldsymbol{\mu}_0 = \mathbf{0}$ because high expression levels of a gene can be associated with either the major or minor allele with roughly equal probability, resulting

in average effect sizes to be approximately zero across tissues. We have noticed little effect of this setting on numerical results. The final parameter of the model is a probability mass function $\mathbf{p}$ on $\{0,1\}^K$ that assigns probabilities to each of the 2^K possible configurations of $\mathbf{\Gamma}_\lambda$. In particular, $p_{\mathbf{0}}$ (i.e., $p_{(0,\cdots,0)}$) is the prior probability that transcript i and SNP j have no association in any tissue.

4.2.4 Mixture Model

The hierarchical model (4.2)-(4.5) describing the distribution of $\mathbf{Z}_\lambda$ is fully specified by $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$, which consists of $2^K + K^2 + K - 1$ real-valued parameters. Estimation of, and inference from, the hierarchical model is based on an equivalent mixture representation that we now discuss.

If $\mathbf{U}$ is distributed as $\mathcal{N}_K(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{\gamma}$ is a fixed vector in $\{0,1\}^K$, then one may readily verify that the entrywise product $\mathbf{U} \cdot \boldsymbol{\gamma}$ is distributed as $\mathcal{N}_K\big(\boldsymbol{\mu} \cdot \boldsymbol{\gamma}, \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T\big)$. A straightforward argument then shows that the hierarchical model (4.2)-(4.5) is equivalent to a mixture distribution of the form

$$\mathbf{Z}_\lambda \;\sim \sum_{\boldsymbol{\gamma} \in \{0,1\}^K} p_{\boldsymbol{\gamma}}\, \mathcal{N}_K\big(\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma},\, \Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T\big). \tag{4.6}$$

We adopt an empirical Bayes approach for performing inference from the model (4.6). Specifically, the parameters $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$ of the hierarchical model are estimated from the observed z-statistics $\{\mathbf{z}_\lambda : \lambda \in \Lambda\}$ by approximately maximizing a pseudo-likelihood derived from (4.6); see Appendix 4.3 for more details. Beginning with the work of Newton et al. (2001) and Efron et al. (2001), empirical Bayes approaches have been applied to hierarchical models in a number of genetic applications, most notably the study of differential expression and co-expression in gene expression microarrays, cf. Kendziorski et al. (2003), Newton et al. (2004), Smyth (2004) and Efron (2008), and Dawson and Kendziorski (2012).

The mixture model (4.6) is readily interpretable. Each component of the model corresponds to a unique configuration $\boldsymbol{\gamma}$, or equivalently, a unique pattern of tissue specificity. The model component corresponding to $\boldsymbol{\gamma} = \mathbf{0}$ represents the case in which there are no

eQTLs in any tissue, and has associated (null) distribution $\mathcal{N}_K(\mathbf{0}, \Delta)$. The model component corresponding to $\boldsymbol{\gamma} = \mathbf{1}$ represents the case in which there are eQTLs in every tissue, and has associated distribution $\mathcal{N}_K(\boldsymbol{\mu}_0, \Delta + \Sigma)$. Other values of $\boldsymbol{\gamma}$ represent intermediate cases in which there are eQTLs in some tissues (those with $\gamma_k = 1$) and not in others (those with $\gamma_k = 0$).

4.2.5 Marginal Consistency

In eQTL studies with multiple tissues, it is likely that some subsets of the tissues are of particular interest. From the point of view of model fitting and model interpretation, it is desirable if the model for any subset of tissues is consistent with the full model in the sense that it can be obtained from the full model (or any model on a superset of tissues) via marginalization. We refer to this property as *marginal consistency*.

To elaborate, let $S \subseteq \{1, \ldots, K\}$ be a subset of r tissues, with $1 \leq r \leq K$. The mixture model (4.6) has two important compatibility properties: (i) the marginalization of the full model to S has the same general form as the model derived from S alone; and (ii) the parameters of the marginal model are obtained by restricting the parameters of the full model to S. The following definition and lemma makes these statements precise. A proof of the lemma is given in the appendix.

Definition: Let $S \subseteq \{1, \ldots, K\}$ with cardinality $|S| = r$. For each vector $\mathbf{u} \in \mathbb{R}^K$ let $\mathbf{u}_S = (u_k : k \in S) \in \mathbb{R}^r$ be the vector obtained by restricting $\mathbf{u}$ to the entries in S. Similarly, for each matrix $A \in \mathbb{R}^{K \times K}$ let $A_S = \{a_{kl} : k, l \in S\}$ be the $r \times r$ matrix obtained by retaining only the rows and columns with indices in S. Note that if A is non-negative (positive) definite, then A_S is non-negative (positive) definite as well.

Lemma 4.2.1. *If $\mathbf{Z} \in \mathbb{R}^K$ be a random vector having the mixture distribution (4.6), then*

$$\mathbf{Z}_S \;\sim \sum_{\boldsymbol{\zeta} \in \{0,1\}^r} p_{S,\boldsymbol{\zeta}}\, \mathcal{N}_r\big(\boldsymbol{\mu}_{0S} \cdot \boldsymbol{\zeta},\, \Delta_S + \Sigma_S \cdot \boldsymbol{\zeta}\boldsymbol{\zeta}^T\big) \tag{4.7}$$

where $(p_{S,\mathbf{0}}, \cdots, p_{S,\mathbf{1}})$ is the probability mass function on $\{0,1\}^r$ obtained by marginalizing $\mathbf{p}$

to S, i.e., $p_{S,\boldsymbol{\zeta}} = \sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}_S=\boldsymbol{\zeta}} p_{\boldsymbol{\gamma}}$.

Remark: Suppose that the parameters $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$ of the full mixture model (4.6) are estimated from the z-statistic vectors $\{\mathbf{z}_\lambda : \lambda \in \Lambda\}$, and let $S \subseteq \{1, \ldots, K\}$ be a set containing r tissues. Lemma 4.2.1 describes the model θ_S obtained by marginalizing the full model to the tissue set S.

4.3 Model Fitting and Parameter Estimation

4.3.1 Matrix eQTL

The set of correlations $r_{\lambda k}$ for all transcript-SNP pairs λ and tissues $k = 1, \ldots, K$ can be conveniently calculated using the R package Matrix eQTL by Shabalin (2012). The package is designed for fast eQTL analysis in individual tissues. Matrix eQTL accounts for covariates and can filter transcript-SNP pairs by the distance between their genomic locations. Once Matrix eQTL is applied separately for each tissue, the t-statistics it reports can be transformed into correlations using the simple transformation

$$r_{\lambda k} = \frac{t_{\lambda k}}{\sqrt{d_k + t_{\lambda k}^2}}$$

where d_k is the number of degrees of freedom in the tests for tissue k which is defined in Section 4.2.2 and is also reported by Matrix eQTL. The set of correlations can then be combined in a single matrix with rows $\mathbf{r}_\lambda$.

4.3.2 Modified EM Algorithm

We wish to estimate the parameter $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$ from the observed z-statistics $\{\mathbf{z}_\lambda : \lambda \in \Lambda\}$, which are computed directly from the sample correlations $r_{\lambda k}$ obtained from Matrix eQTL. In order to make the estimation of θ tractable, we assume that the random vectors $\mathbf{Z}_\lambda$ are independent. The likelihood of the model then has a simple product form, depending

only on the unknown parameter θ, and the observed z-statistics $\{\mathbf{z}_\lambda\}$:

$$L(\{\mathbf{z}_\lambda\}|\theta) = \prod_{\lambda\in\Lambda} \sum_{\boldsymbol{\gamma}\in\{0,1\}^K} p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z}_\lambda \,|\, \theta), \tag{4.8}$$

where $f_{\boldsymbol{\gamma}}(\cdot \,|\, \theta)$ is the probability density function of the $\mathcal{N}_K\big(\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma}, \Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T\big)$ distribution.
Remark: It is important to note that the parameter θ concerns only the (common) marginal distribution of the random vectors $\mathbf{Z}_\lambda$, and is unaffected by their dependence. The assumption that the random vectors $\mathbf{Z}_\lambda$ are independent facilitates estimation of θ, but does not impose any constraints on the marginal dependence structure of $\mathbf{Z}_\lambda$.

We estimate the parameter θ by seeking to maximize the logarithm of the likelihood (4.8). The log-likelihood is not concave, and there appears to be no closed form solution to the maximization problem. Thus one must to rely on iterative algorithms that produce a sequence of parameters $\theta^{(t)}$ converging to a (local) maximum of the likelihood. A direct approach employing a generic software routine for numerical maximization of the likelihood function would be computationally intensive, as each iteration would require multiple (at least 2^K) calculations of the likelihood function around the estimate obtained at the previous iteration. A much faster convergence can be achieved by applying a modification of Expectation Maximization (EM) algorithm. Details are given below.

We treat the unobserved tissue-specificity information vector $\boldsymbol{\Gamma}_\lambda \in \{0,1\}^K$ as a latent variable. The joint likelihood of both observed and latent variables is:

$$L(\mathbf{z}, \boldsymbol{\gamma} \,|\, \theta) \;=\; p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z} \,|\, \theta).$$

The EM algorithm operates in an iterative fashion. Let $\theta^{(t)} = (\boldsymbol{\mu}_0^{(t)}, \Delta^{(t)}, \Sigma^{(t)}, \mathbf{p}^{(t)})$ be the estimate of the model parameters after t iterations. The estimate $\theta^{(t+1)}$ is defined by

$$\theta^{(t+1)} \;=\; \arg\max_{\theta} Q(\theta : \theta^{(t)}),$$

where

$$Q(\theta : \theta^{(t)}) = \sum_{\lambda} \mathbb{E}_{\mathbf{\Gamma}_\lambda | \mathbf{z}_\lambda, \theta^{(t)}} \big[\log L(\mathbf{z}_\lambda, \mathbf{\Gamma}_\lambda | \theta) \big].$$

The expectation of the log-likelihood is calculated with respect to the conditional distribution of $\mathbf{\Gamma}_\lambda$ given the observed vector of correlations $\mathbf{z}_\lambda$ and the model parameters $\theta^{(t)}$.

Consider the conditional expectation appearing in $Q(\theta : \theta^{(t)})$. Let $p(\boldsymbol{\gamma} \,|\, \theta)$ denote the probability of the configuration $\boldsymbol{\gamma}$ under the probability mass function $\mathbf{p}$ associated with the parameter θ, and define

$$p(\boldsymbol{\gamma} \,|\, \mathbf{z}, \theta) \;=\; \mathbb{P}(\mathbf{\Gamma}_\lambda = \boldsymbol{\gamma} \,|\, \mathbf{z}, \theta) \;=\; \frac{p(\boldsymbol{\gamma} \,|\, \theta) f_{\boldsymbol{\gamma}}(\mathbf{z} \,|\, \theta)}{\sum_{\boldsymbol{\gamma}'} p(\boldsymbol{\gamma}' \,|\, \theta) f_{\boldsymbol{\gamma}'}(\mathbf{z} \,|\, \theta)}$$

The objective function $Q(\theta : \theta^{(t)})$ then has the form

$$Q(\theta : \theta^{(t)}) = \sum_{\lambda} \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \big[\log p(\boldsymbol{\gamma} \,|\, \theta) + \log f_{\boldsymbol{\gamma}}(\mathbf{z}_\lambda \,|\, \theta) \,\big]$$

Maximization of Q with respect to θ leads to the explicit formula

$$p(\boldsymbol{\gamma} \,|\, \theta^{(t+1)}) \;=\; \sum_{\lambda} p(\boldsymbol{\gamma} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \Big/ |\Lambda|$$

where $|\Lambda|$ is the number of gene-SNP pairs under consideration. There appears to be no closed form solution for the iterates of $\boldsymbol{\mu}_0^{(t)}$, $\Sigma^{(t)}$ and $\Delta^{(t)}$. However, in practice, most of the probability mass of $\mathbf{p}$ is concentrated at the two extreme cases $\boldsymbol{\gamma} = \mathbf{0}$ and $\boldsymbol{\gamma} = \mathbf{1}$, reflecting the fact that most transcript-SNP pairs are associated in no tissues or all tissues. Approximating $Q(\cdot)$ by restricting the second sum to $\boldsymbol{\gamma} = 0, 1$ leads to explicit (approximate) estimates of

$\boldsymbol{\mu}_0$, Σ and Δ via the following first order conditions:

$$\begin{aligned}
\Delta^{(t+1)} &= \sum_{\lambda} p(\mathbf{0} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \mathbf{z}_\lambda \mathbf{z}_\lambda^T \Big/ \sum_{\lambda} p(\mathbf{0} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \\
\boldsymbol{\mu}_0^{(t+1)} &= \sum_{\lambda} p(\mathbf{1} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \mathbf{z}_\lambda \Big/ \sum_{\lambda} p(\mathbf{1} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \\
\Sigma^{(t+1)} + \Delta^{(t+1)} &= \sum_{\lambda} p(\mathbf{1} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) (\mathbf{z}_\lambda - \boldsymbol{\mu}_0^{(t+1)}) (\mathbf{z}_\lambda - \boldsymbol{\mu}_0^{(t+1)})^T \Big/ \sum_{\lambda} p(\mathbf{1} \,|\, \mathbf{z}_\lambda, \theta^{(t)})
\end{aligned}$$

At some iterations the estimates $\Sigma^{(t+1)}$ may fail to be non-negative definite. In such cases we force $\Sigma^{(t+1)}$ to be non-negative definite by calculating its singular value decomposition and dropping terms with negative coefficients (negative eigenvalues).

Starting with an initial parameter value $\theta^{(0)}$, we perform sequential updates in the manner described above until the change in the likelihood falls below a pre-set threshold. To assess the reliability of the estimate one may run the algorithm multiple times using distinct starting points. In our experiments the algorithm tends to converge to the same estimate regardless of the starting point.

4.4 Multi-Tissue eQTL Inference

Once fit, the mixture model (4.6) provides the basis for inference about eQTLs across tissues. In practice, we expect that θ will be well-estimated due to the large number of available gene-SNP pairs; we therefore regard θ as fixed and known. For data sets with small sample sizes, approximate standard errors can be obtained from the likelihood via the observed information matrix.

In most applications the covariance matrix Δ will be positive definite, and we assume this is the case here. With this assumption, the distribution $\mathcal{N}_K\big(\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma},\, \Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T\big)$ associated with the configuration $\boldsymbol{\gamma} \in \{0,1\}^K$ has a density, which we denote by $f_{\boldsymbol{\gamma}}$. Thus under the mixture model (4.6) the random vector $\mathbf{Z}_\lambda$ has density

$$f(\mathbf{z}) \;=\; \sum_{\boldsymbol{\gamma}} p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z}) \qquad \mathbf{z} \in \mathbb{R}^K. \tag{4.9}$$

In view of this expression and the hierarchical model (4.2)-(4.5), one may regard $\mathbf{Z}_\lambda$ as one element of a jointly distributed pair $(\mathbf{\Gamma}_\lambda, \mathbf{Z}_\lambda)$, where

$$\mathbf{\Gamma}_\lambda \sim \mathbf{p} \;\text{ and }\; \mathbf{Z}_\lambda \,|\, \mathbf{\Gamma}_\lambda \;\sim\; f_{\boldsymbol{\gamma}}. \tag{4.10}$$

We carry out multi-tissue eQTL analysis based on the posterior distribution of the configuration $\mathbf{\Gamma}_\lambda$ given the observed vector of z-statistics $\mathbf{z}_\lambda$. Two inference problems are of central interest to us. The first is eQTL detection, in all tissues and in a subset of tissues. The second is assessing the tissue specificity of eQTLs in transcript-SNP pairs where an eQTL is present in at least one tissue.

4.4.1 Detection of eQTLs Using the Local False Discovery Rate

A primary goal of multi-tissue analysis is testing each transcript-SNP pair for the presence of an eQTL in at least one tissue. This can be formulated as a multiple testing problem, namely testing

$$\mathrm{H}_{0,\lambda} : \mathbf{\Gamma}_\lambda = \mathbf{0} \;\text{ versus }\; \mathrm{H}_{1,\lambda} : \mathbf{\Gamma}_\lambda \neq \mathbf{0} \;\text{ for }\; \lambda \in \Lambda. \tag{4.11}$$

For $\lambda = (i, j) \in \Lambda$ the null hypothesis $H_{0,\lambda}$ asserts that there is no eQTL between transcript i and SNP j in any tissue, while the alternative $H_{1,\lambda}$ asserts that there is an eQTL between i and j in at least one tissue.

The null hypotheses can also be expressed in the form $\mathrm{H}_{0,\lambda} : \mathbf{Z}_\lambda \sim \mathcal{N}_K\big(\mathbf{0},\, \Delta\big)$. It is possible to derive a p-value for $\mathbf{z}_\lambda$ directly from the null distribution, and then control the overall false discovery rate in (4.11) using a step-up procedure like that of Benjamini and Hochberg (1995). However, this type of analysis ignores relevant information about the distribution of $\mathbf{Z}_\lambda$ under the alternative that is contained in the mixture model.

We address the multiple testing problem (4.11) using the local false discovery rate introduced by Efron et al. (2001) in the context of an empirical Bayes analysis of differential expression in microarrays. Other applications of the local false discovery rate to genomic

problems can be found in Newton et al. (2004), Efron (2007), and Efron (2008). To simplify notation in what follows, let $(\mathbf{\Gamma}, \mathbf{Z})$ denote a generic pair distributed as $(\mathbf{\Gamma}_\lambda, \mathbf{Z}_\lambda)$.

Definition: The *local false discovery rate* of an observed z-statistic vector $\mathbf{z}$ under the model (4.6) is defined by

$$\eta(\mathbf{z}) \;:=\; \mathbb{P}(\mathbf{\Gamma} = \mathbf{0} \,|\, \mathbf{Z} = \mathbf{z}) \;=\; \frac{p_0 f_0(\mathbf{z})}{f(\mathbf{z})}. \tag{4.12}$$

Let $\alpha \in (0, 1)$ be a target false discover rate (FDR) for the multiple testing problem (4.11). Vectors $\mathbf{z}$ for which the local false discovery rate $\eta(\mathbf{z})$ is small provide evidence for the alternative $\mathbf{\Gamma} \neq \mathbf{0}$. We carry out testing of gene-SNP pairs using a simple step-up procedure that is applied to the running average of the ordered local false discover rates. The procedure, which is described below, appears in essentially the same form in Newton et al. (2004), Sun and Cai (2007), and Cai and Sun (2009).

Local FDR Step-Up Procedure: Target FDR $= \alpha$

1. Given: Observed z-statistic vectors $\{\mathbf{z}_\lambda : \lambda \in \Lambda\}$.
2. Enumerate the elements of Λ as $\lambda_1, \ldots, \lambda_N$ so that $\eta(\mathbf{z}_{\lambda_1}) \leq \cdots \leq \eta(\mathbf{z}_{\lambda_N})$.
3. Reject hypotheses $\mathrm{H}_{0,\lambda_1}, \ldots, \mathrm{H}_{0,\lambda_L}$ where L is the largest integer such that $L^{-1} \sum_{l=1}^{L} \eta(\mathbf{z}_{\lambda_l}) \leq \alpha$.

In order to better understand the local FDR step-up procedure, and to assess its performance, it is useful to express the procedure in an equivalent form. As noted by Efron et al. (2001), the false discovery rate associated with a rejection region $R \subseteq \mathbb{R}^k$ for the multiple testing problem (4.11) is given by $\mathbb{P}(\mathbf{\Gamma} = \mathbf{0} \,|\, \mathbf{Z} \in R)$. They establish the following elementary fact, which exhibits a connection between the false discovery rate and the local false discovery rate.

Proposition 4.4.1. *If* $R \subseteq \mathbb{R}^k$ *is such that* $\mathbb{P}(\mathbf{Z} \in R) > 0$, *then* $\mathbb{P}(\mathbf{\Gamma} = \mathbf{0} \,|\, \mathbf{Z} \in R) = \mathbb{E}(\eta(\mathbf{Z}) \,|\, \mathbf{Z} \in R)$.

As noted above, vectors $\mathbf{z}$ for which $\eta(\mathbf{z})$ is small provide evidence against $\mathbf{\Gamma} = \mathbf{0}$, so it is natural to reject $\mathrm{H}_{0,\lambda}$ when $\eta(\mathbf{z}_\lambda)$ falls below an appropriate threshold. Consider rejection regions of the form $R(t) = \{\mathbf{z} : \eta(\mathbf{z}) \le t\}$ for $t \in (0,1)$. Given a target false discovery rate α, we wish to find t such that $\alpha = \mathbb{P}(\mathbf{\Gamma} = \mathbf{0} \,|\, \mathbf{Z} \in R(t))$. By Proposition 4.4.1 this is equivalent to finding $t \in (0,1)$ such that $F(t) = \alpha$, where

$$F(t) \;:=\; \mathbb{E}(\eta(\mathbf{Z}) \,|\, \eta(\mathbf{Z}) \le t) \;=\; \frac{\mathbb{E}[\eta(\mathbf{Z})\, \mathbb{I}(\eta(\mathbf{Z}) \le t)]}{\mathbb{P}(\eta(\mathbf{Z}) \le t)}. \tag{4.13}$$

The empirical analog of $F(t)$ is the ratio

$$\hat{F}(t) \;=\; \frac{\sum_{\lambda \in \Lambda} \eta(\mathbf{z}_\lambda)\, \mathbb{I}(\eta(\mathbf{z}_\lambda) \le t)}{\sum_{\lambda \in \Lambda} \mathbb{I}(\eta(\mathbf{z}_\lambda) \le t)},$$

which depends only on $\eta(\cdot)$ and the observed vectors $\{\mathbf{z}_\lambda\}$. It is easy to see that the local FDR step-up procedure is equivalent to the rule

$$\text{Reject } \mathrm{H}_{0,\lambda} \text{ if and only if } \eta(\mathbf{z}_\lambda) \le \sup\{t : \hat{F}(t) \le \alpha\}. \tag{4.14}$$

We show in Proposition 4.8.1 that $F(t)$ is strictly increasing and continuous. Thus if $F(t)$ and $\hat{F}(t)$ were equal, the local FDR step-up procedure and the idealized threshold procedure would coincide. In general, $F(t)$ and $\hat{F}(t)$ will be different, but multiplying the numerator and denominator of $\hat{F}(t)$ by $|\Lambda|^{-1}$ it is evident that the two functions will be close if $|\Lambda|$ is large and the dependence among the observed z-vectors is not extreme. Asymptotic control of the false discovery rate by the local FDR step-up procedure is established in Theorem 4.4.1 below.

Let $\Lambda^* \subseteq \mathbb{N} \times \mathbb{N}$ be an infinite index set, and let $\Lambda_1, \Lambda_2, \ldots \subseteq \Lambda^*$ be a sequence of finite subsets of Λ^*. Let $\alpha \in (0,1)$ be a target FDR that is less than the maximum value of $\eta(\mathbf{z})$. For each $n \ge 1$ let $\{(\mathbf{\Gamma}_\lambda, \mathbf{Z}_\lambda) : \lambda \in \Lambda_n\}$ be jointly distributed pairs having the same distribution as $(\mathbf{\Gamma}, \mathbf{Z})$. In order to assess the performance of the local FDR step-up procedure on the observed z-statistic vectors $\{\mathbf{Z}_\lambda : \lambda \in \Lambda_n\}$ we consider the equivalent rule (4.14), which rejects $H_{0,\lambda}$

when $\eta(\mathbf{Z}_\lambda) \leq \hat{\theta}_n = \sup\{t : \hat{F}_n(t) \leq \alpha\}$ where

$$\hat{F}_n(t) \;=\; \frac{\sum_{\lambda\in\Lambda_n} \eta(\mathbf{Z}_\lambda)\, \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leq t)}{\sum_{\lambda\in\Lambda_n} \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leq t)} \qquad 0 < t < 1.$$

The number of false discoveries and total discoveries for the local FDR step-up procedure are equal, respectively, to

$$M_n \;=\; \sum_{\lambda\in\Lambda_n} \mathbb{I}(\mathbf{\Gamma}_\lambda = 0)\, \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leq \hat{\theta}_n) \quad \text{and} \quad N_n \;=\; \sum_{\lambda\in\Lambda_n} \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leq \hat{\theta}_n).$$

Theorem 4.4.1. *Let* $(\mathbf{\Gamma}, \mathbf{Z})$ *have joint distribution given by the mixture model (4.10) with parameters* $(\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$*. Assume that* Δ *is positive definite and that the diagonal entries of* Σ *are positive. If* $\hat{F}_n(t) \to F(t)$ *in probability for each* $t \in (0,1)$ *then* $\mathbb{E}M_n/\mathbb{E}N_n \to \alpha$ *as* n *tends to infinity.*

The proof of Theorem 4.4.1 is given in the appendix, Section 4.8.2. The ratio of expectations $\mathbb{E}M_n/\mathbb{E}N_n$ is sometimes referred to as the marginal false discovery rate (m-FDR). Sun and Cai (2007) and Cai and Sun (2009) established optimality properties and m-FDR control of several local FDR based testing procedures, including the step-up procedure used here, under independence and monotonicity assumptions. However, these assumptions are typically violated in the setting of interest to us here. The monotonicity assumption, which in the present case involves the relationship between the distributions of the local FDR $\eta(\mathbf{Z}_\lambda)$ under $H_{0,\lambda}$ and $H_{1,\lambda}$, does not appear to hold. Moreover, in eQTL data there are typically significant correlations between nearby SNPs (linkage disequilibrium), leading to to complex, non-stationary correlations between the gene-SNP based vectors $\mathbf{Z}_\lambda$.

Theorem 4.4.1 makes no explicit assumptions on the joint distribution of the vectors $\mathbf{Z}_\lambda$; instead it relies on the relatively weak condition that $\hat{F}_n(t) \to F(t)$ in probability. This condition holds, for example, under the (very mild) assumption that the variance of the numerator and denominator of $\hat{F}_n(t)$ is equal to $o(|\Lambda_n|^2)$. While strong correlations between nearby SNPs will be present, gene-SNP pairs that are well separated will have little or no correlation, so the variance decay assumption is reasonable in practice. When the variance

decay assumption holds, the conclusion of the theorem may be strengthened to $M_n/N_n = \alpha + o_P(1)$.

In regards to the proof of Theorem 4.4.1, the assumption that Δ be positive definite is only needed to ensure the existence of the the densities f_γ; the assumption that the diagonal entries of Σ are positive is reasonable in practice, but can likely be weakened. The proof makes use of the properties of the multivariate normal, specifically the normality of conditional distributions and the fact that normal densities are analytic functions, but could likely be extended to more general exponential families with additional work.

4.4.2 Analysis for Subsets of Tissues

In some problems, a subset $S \subseteq \{1, \ldots, K\}$ of the available tissues may be of primary interest. The multiple testing framework described above can be adapted to the tissues in S in two primary ways. The first is to construct a model based only on the tissues in S and use the resulting local FDR to identify multi-tissue eQTLs. However, this approach does not make use of the available data from tissues outside S and as such it does not borrow strength from commonalities among tissues. As an alternative, one may use the *marginal local FDR* for S, defined by

$$\eta_S(\mathbf{z}) \;:=\; \mathbb{P}(\mathbf{\Gamma}_S = \mathbf{0} \,|\, \mathbf{Z} = \mathbf{z}) \;=\; \frac{\sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}_S=\mathbf{0}} p_{\boldsymbol{\gamma}} f_{\boldsymbol{\gamma}}(\mathbf{z})}{f(\mathbf{z})}. \tag{4.15}$$

Here $\mathbf{\Gamma}_S$ and $\boldsymbol{\gamma}_S$ denote, respectively, the restriction of the vectors $\mathbf{\Gamma}$ and $\boldsymbol{\gamma}$ to the tissues in S, while $p_{\boldsymbol{\gamma}}$, $f_{\boldsymbol{\gamma}}$ and f correspond to the full model (4.6). We emphasize that the marginal local FDR $\eta_S(\mathbf{z})$ is a function of the complete vector of z-statistics, and therefore depends on the fitted model for the full set of tissues.

4.4.3 Assessments of Tissue Specificity

Testing gene-SNP pairs is typically the first step in multi-tissue eQTL analysis. Rejection of $\mathrm{H}_{0,\lambda}$ is based on evidence that λ is an eQTL in at least one of the available tissues. More detailed statements about the pattern of eQTLs across tissues can be made using information

about the full configuration vector $\mathbf{\Gamma}_\lambda$. If the hypothesis $\mathrm{H}_{0,\lambda}$ is rejected, a natural estimate of $\mathbf{\Gamma}_\lambda$ is the maximum a-posteriori (MAP) configuration defined by

$$\widehat{\boldsymbol{\gamma}}_\lambda \;=\; \underset{\boldsymbol{\gamma}\in\{0,1\}^K\setminus\mathbf{0}}{\arg\max}\; p(\boldsymbol{\gamma}\,|\,\mathbf{z}_\lambda) \;=\; \underset{\boldsymbol{\gamma}\in\{0,1\}^K\setminus\mathbf{0}}{\arg\max}\; p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z}_\lambda).$$

The MAP rule is investigated in the simulation section below. As an alternative, one may compute the marginal posterior probability of an eQTL in each tissue k, namely

$$p(\mathbf{\Gamma}_{\lambda,k}=1|\mathbf{z}_\lambda) \;=\; \sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}_k=1} p(\boldsymbol{\gamma}|\mathbf{z}_\lambda) \;=\; \sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}_k=1} p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z}_\lambda)/f(\mathbf{z}_\lambda),$$

and declare an eQTL in tissue k if this marginal probability exceeds a predefined threshold. Both MAP and thresholding of the marginal posterior extend to subsets of tissues.

4.4.4 Testing a Family Configurations

The goal of the multiple testing problem (4.11) is to determine whether the configuration $\mathbf{\Gamma}_\lambda$ of a gene-SNP pair is equal to $\mathbf{0}$ or belongs to the complementary set $\{0,1\}^K\setminus\{\mathbf{0}\}$. More generally, one may test membership of $\mathbf{\Gamma}_\lambda$ in any fixed subset $T\subseteq\{0,1\}^K$ of configurations. The associated testing problem can be written as

$$\mathrm{H}^T_{0,\lambda}:\mathbf{\Gamma}_\lambda\in T^c \;\;\text{versus}\;\; \mathrm{H}^T_{1,\lambda}:\mathbf{\Gamma}_\lambda\in T, \qquad \lambda\in\Lambda. \tag{4.16}$$

A test statistic for (4.16) can be obtained by marginalizing the full local FDR (4.12), which yields

$$\eta_T(\mathbf{z}) \;:=\; \mathbb{P}(\mathbf{\Gamma}\in T^c\,|\,\mathbf{Z}=\mathbf{z}) \;=\; \frac{\sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}\in T^c} p_{\boldsymbol{\gamma}} f_{\boldsymbol{\gamma}}(\mathbf{z})}{f(\mathbf{z})}.$$

The local FDR step-up procedure can then be applied to the values $\{\eta_T(\mathbf{z}_\lambda)\}$ in order to control the overall FDR in (4.16).

4.5 Simulation Study

In this section, we examine the performance of MT-eQTL through a simulation study. As the basis of the model and subsequent inferences is the collection of z-statistic vectors derived from the observed genotype and transcript data, we directly simulate the z-statistic vectors themselves.

4.5.1 Simulation Setting

For $K = 4$ tissues we simulate 10 million vectors $\mathbf{z}_\lambda$ independently from the mixture model (4.6) using parameters $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$ obtained from eQTL analysis of data from the GTEx initiative. Specifically, we consider the tissues blood, lung, muscle, and thyroid, which we denote by a, b, c, and d, respectively. Sample sizes, sample overlap, and degrees of freedom after covariate correction are given in Table 4.1. See Section 4.6 for more details.

	a	b	c	d	Degree of Freedom
a	156	104	122	90	137
b		119	100	84	100
c			138	88	119
d				105	86

Table 4.1: Sample sizes (diagonal), sample overlap (off-diagonal), and degrees of freedom for different tissues in the simulation.

For computational simplicity, $\boldsymbol{\mu}_0$ is set to zero in the simulations and model fitting. The generating parameters Δ and Σ based on the GTEx data are as follows:

$$\Delta = \begin{pmatrix} 1.0000 & 0.1347 & 0.0805 & 0.1089 \\ 0.1347 & 1.0000 & 0.1204 & 0.1794 \\ 0.0805 & 0.1204 & 1.0000 & 0.1288 \\ 0.1089 & 0.1794 & 0.1288 & 1.0000 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 6.5699 & 5.3098 & 4.4683 & 4.7126 \\ 5.3098 & 5.9752 & 4.7906 & 5.5778 \\ 4.4683 & 4.7906 & 5.5263 & 4.6493 \\ 4.7126 & 5.5778 & 4.6493 & 6.0178 \end{pmatrix}.$$

The generating parameter $\mathbf{p}$ can be found in Table 4.2. We simulated each vector $\mathbf{z}_\lambda$ from (4.6) in a two-step fashion: first drawing $\boldsymbol{\gamma} \in \{0, 1\}^4$ from $\mathbf{p}$, and then drawing $\mathbf{z}_\lambda$ from $f_{\boldsymbol{\gamma}}(\mathbf{z})$. Access to the true configurations $\boldsymbol{\gamma}$ enables us to assess false discovery rates associated with

inferences from the fitted model.

4.5.2 Model Fit

The approximate EM procedure was used to fit the full 4-tissue model, as well as all possible 1-, 2-, and 3-tissue models. We terminated EM updates when the difference between log likelihoods in two consecutive iterations was less than 0.01. The number of iterations until convergence of the EM procedure varied from 40 to 132, with average equal to 80. The running time of the EM procedure depended on the number of tissues in the model, ranging from about 1 second per iteration for the 1-tissue models to about 40 seconds per iteration for full 4-tissue model. Fitting of the 4-tissue model based on the simulated data took slightly more than one hour.

As expected, the parameters estimated from the simulated data are very close to those used to generate the data. For the 4-tissue model, the relative error of each entry of Σ is less than 0.3%, while the relative error for each entry of Δ is less than 0.7%. For the probability mass vector $\mathbf{p}$, thirteen of sixteen entries had relative error less than 1%, with the remaining relative errors equal to 1.45%, 1.66% and 4.31%. These results confirm that the EM procedure works well on the simulated data.

4.5.3 Results

We applied the adaptive thresholding procedure to the full 4-tissue model with FDR threshold $\alpha = 0.05$ in order to identify gene-SNP pairs that are eQTLs in at least one tissue. For all models considered in the simulation, the true false discovery rates were slightly below 0.05.

Table 4.2 shows results from the 4-tissue model with MAP estimates of the configuration $\boldsymbol{\gamma}$. The TS-config column enumerates the 16 possible configurations according to the tissues in which eQTLs are present. The True column shows the true numbers of transcript-SNP pairs with the specified configuration in the simulated data. The Discoveries column shows the number of transcript-SNP pairs in the simulation estimated to have the specified

configuration. The Intersection column shows cardinality of the intersection of true and discovered transcript-SNP pairs with the specified configuration. The Proportion column gives the proportion of true discoveries.

For each configuration, only a modest fraction (about 1/4) of the true eQTLs with that configuration are detected by the local FDR procedure. This does not imply that the local FDR procedure is under-powered, but instead reflects features of the data generation process that we believe are representative of real data. In detail, the multi-tissue z-statistics of each gene-SNP pair are generated from a mixture multivariate Gaussian distribution centered at zero. As a result, the majority of alternative gene-SNP pairs have z-statistics near zero; these z-statistics are not readily distinguishable from those generated under the null.

For most configurations the proportion of true discoveries relative to total discoveries (the Proportion column) are above 60 percent. This is relatively high, given that distinguishing between nearby configurations (those with 1's in all but one of same positions) as well as the null configuration can be difficult.

TS-config	100*$\mathbf{p}$	True	Discoveries	Intersection	Proportion
0	77.24	7720693	8961544	7669320	0.86
a	1.96	196868	52070	33128	0.64
b	1.04	103866	23786	17070	0.72
c	1.88	189859	45253	28738	0.64
d	2.05	202925	53716	37600	0.70
a-b	0.29	29516	4592	3035	0.66
a-c	0.08	7835	446	313	0.70
a-d	0.09	9507	1280	870	0.68
b-c	0.10	9552	1448	903	0.62
b-d	0.33	32552	5196	2997	0.58
c-d	0.37	36738	6382	4294	0.67
a-b-c	0.19	19022	1730	1258	0.73
a-b-d	0.86	85418	9115	6194	0.68
a-c-d	0.09	8614	951	731	0.77
b-c-d	1.08	107405	14031	9445	0.67
a-b-c-d	12.34	1239630	818460	640847	0.78

Table 4.2: eQTL analysis results from the 4-tissue model for the simulation data.

In order to assess how the use of multiple tissues increases statistical power in the context of the simulation, we fit models for tissue sets $\{a\}$, $\{a, b\}$, $\{a, b, c\}$, and $\{a, b, c, d\}$ and only focused on eQTL detection in tissue a. In each case we applied the adaptive thresholding

procedure to the marginal local FDR defined in (4.15). Figure 4.2 shows that the number of discoveries for tissue a increases steadily with the number of auxiliary tissues. The realized false discovery rates were all controlled at the specified 0.05 level.

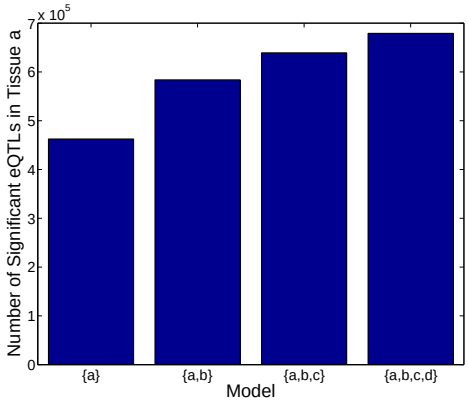


Figure 4.2: The number of significant discoveries in tissue a from the model for $\{a\}$, $\{a, b\}$, $\{a, b, c\}$, and $\{a, b, c, d\}$ respectively.

4.6 GTEx Data Analysis

In this section, we apply the MT-eQTL model and inference procedures to the December 2012 data freeze from the GTEx initiative. The data is publicly available from http://www.broadinstitute.org/gtex/.

4.6.1 Data Preprocessing

We focus on nine primary tissues having between 80 and 160 samples: adipose, artery, blood, heart, lung, muscle, nerve, skin, and thyroid. In what follows, tissues will be ordered alphabetically. In total, there are 175 genotyped individuals with expression data in at least one of these tissues. Figure 4.3 shows the sample sizes and the donor overlaps for all nine tissues.

Each entry of the genotype data matrix $\mathbf{G}$ records the minor allele frequency (MAF) of one donor at one SNP locus. Any missing value at a locus was imputed by the average MAF of that locus across donors. Loci with MAF less than 5% in all genotyped individuals

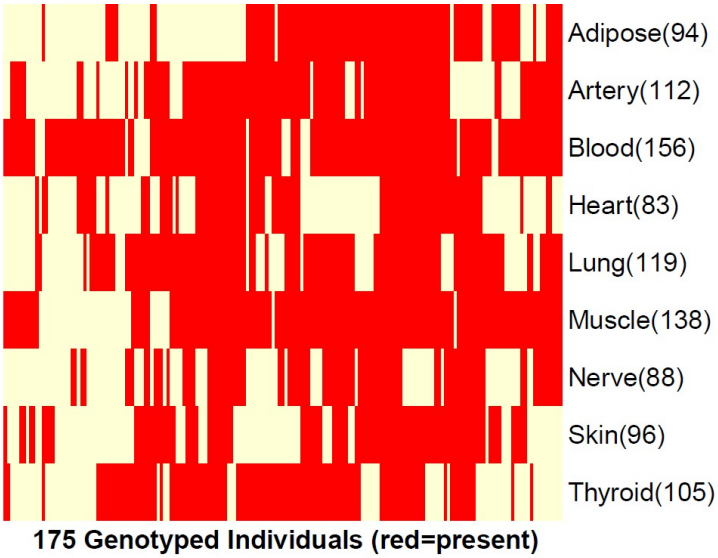


Figure 4.3: Sample information of the GTEx data. Each column represents a genotyped individual with expression measurements in at least one tissue; each row corresponds to a tissue. Red means the individual is a donor of the corresponding tissue.

were discarded, resulting in slightly less than 7 million SNPs. The expression level for each gene in each tissue and sample is measured by the number of mapped reads per kilobase per million reads (RPKM). Genes having less than 10 samples with RPKM greater than 0.1 in some tissue were discarded, leaving slightly more than 20 thousand common genes. In order to improve robustness, the expression values of each gene across the samples in a tissue are inverse quantile normalized.

Fifteen PEER factors were identified from the expression data from each tissue, and three principal components were identified from the genotype data. With an additional covariate for gender, we obtained nineteen covariates in total. For each tissue, the confounding effects were adjusted by residualizing the expression data and the corresponding genotype data on nineteen covariates respectively. Consequently, the degree of freedom for each tissue is equal to the sample size in that tissue minus 19.

4.6.2 Model Fit

We focus on testing of cis-eQTLs, restricting our attention to SNPs that lie within 100 kilobases of the transcription start site of a gene, yielding roughly 10 million gene-SNP pairs of

interest. The z-statistic vectors that act as input for the MT-eQTL model were obtained from the Matrix eQTL package. Subsequently, the 9-tissue MT-eQTL model was fit using the EM algorithm described in Section 4.3.2, with the parameter $\boldsymbol{\mu}_0$ set to zero. Fitting of the model took less than 24 hours, and required less than 8 gigabytes of RAM, on a desktop computer with 2.93GHz Intel Xeon CPU. Timing results for sub-models based on alphabetically ordered tissues are given in Table 4.3. (We note that fitting sub-models of MT-eQTL is unnecessary in practice, as one can obtain them through marginalization of the full model.)

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
Time	15 min	30 min	50 min	1.5 hr	2.5hr	6hr	11hr	16 hr	24 hr

Table 4.3: Approximate timing result for fitting a k-tissue model using the GTEx data.

In what follows we denote the estimated model parameters by $\theta = (\boldsymbol{\mu_0}, \Delta, \Sigma, \mathbf{p})$, which are given in the appendix, Section 4.8.3. The off-diagonal values of Δ are all positive but small in scale (between 0.07 and 0.2), suggesting that donor overlap among tissues and other features of the experimental design have a weak but positive effect on the correlations of effect sizes across tissues. The diagonal values of Σ indicate modest heterogeneity of effect size variation across tissues. The off-diagonal values of Σ reflect positive, often large, correlation of effect sizes arising from commonalities among tissues. The fitted probability mass function $\mathbf{p}$ assigns probabilities to each of the 2^9 possible eQTL configurations. The most likely configuration is $\mathbf{0}$ with $p_{\mathbf{0}} = 0.6808$, indicating that about 68% of the gene-SNP pairs do not have an eQTL in any tissue. To summarize $\mathbf{p}$ we compute the overall probability of seeing an eQTL in k tissues, where k ranges from 0 to 9. The results are shown in the blue curve of Figure 4.4. The curve indicates that the most likely configurations are eQTLs in no tissue, in a single tissue, or in all tissues, and that the least likely configurations are those with eQTLs in roughly half the tissues.

In order to assess the fit of the MT-eQTL model to the data, we compared the marginal distribution of the z-statistics for each tissue under the fitted model with the empirical distribution of the z-statistics for that tissue using a Q-Q plot. The results are given in Figure 4.5. Overall, the marginal two component Gaussian mixture for each tissue fits the bulk of the

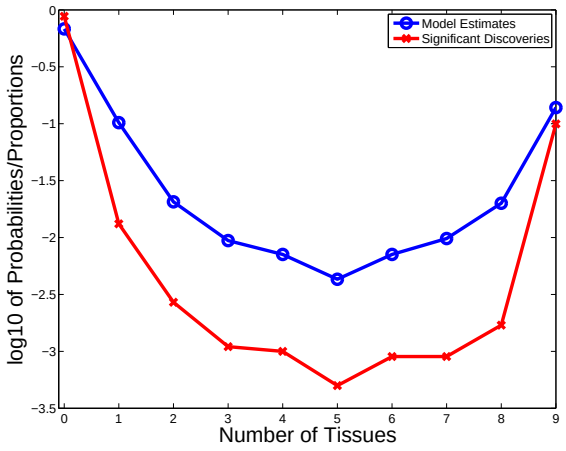


Figure 4.4: The probability of seeing an eQTL in k tissues based on the estimated mass function $\mathbf{p}$ (blue circle), and the proportion of the gene-SNP pairs with an eQTL in k tissues based on the analysis results from the 9-tissue model (red cross), where k ranges from 0 to 9.

observed data well, though we note that there is room for more complex models that might better capture the tail behavior of the data.

4.6.3 Results

Applied to the full 9-tissue model with FDR threshold $\alpha = .05$, the local FDR step-up procedure identified roughly 1.2 million gene-SNP pairs (roughly 12% of the total) with an eQTL in at least one tissue. We subsequently applied the MAP rule to each significant discovery in order to assess tissue specificity. The results, which are summarized in the red curve of Figure 4.4, are in broad agreement with the those derived from the underlying configuration probability $\mathbf{p}$. The downward shift in the red curve results from the fact that many eQTL have small effect sizes and are not detectable by the local FDR procedure at the specified value of α. See also the discussion in Section 4.5.

To better visualize eQTL discoveries and assessments of tissue specificity derived from the MT-eQTL model, it is useful to consider the simple case of two tissues. Figure 4.1b shows scatter plots of z-statistics for nerve and skin, while Figure 4.6 shows scatter plots of z-statistics for adipose and muscle. The black and white plot shows the density of the

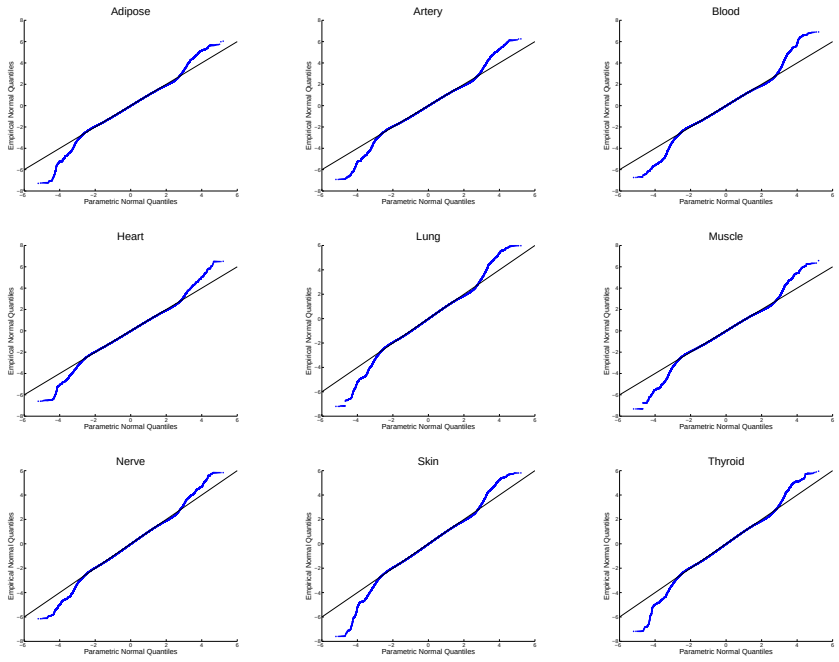


Figure 4.5: Q-Q plots of the observed z-statistics versus the marginal distribution of the fitted model for each tissue.

observed z-statistic vectors, while the companion plot shows the results of inference based on the fitted two-dimensional MT-eQTL model. In the companion plot, z-statistic vectors deemed not to be significant are omitted, leading to the white space at the center of the plot. The remaining points (corresponding to eQTLs) are colored according to their assessed tissue specificity: green represents the configuration $(1, 0)$ in which there is an eQTL in tissue 1 but not tissue 2; red represents the configuration $(0, 1)$ in which there is an eQTL in tissue 2 but not tissue 1; and blue represents the configuration $(1, 1)$ in which there is an eQTL in both tissues.

The overall shape of each plot is a tilted ellipse, with extreme values along the main diagonal and, to a lesser extent, along the coordinate axes. As expected, significant points close to one of the coordinate axes show evidence for an eQTL in a single tissue (tissue specific eQTL), while those along the positive diagonal show evidence for eQTLs in both tissues (common eQTL). In all other pairs of tissues (not shown), we observe similar results. In Figure 4.6, we also observe some discoveries along the anti-diagonal. For anti-diagonal pairs there is significant correlation between genotype and expression in each tissue, but the

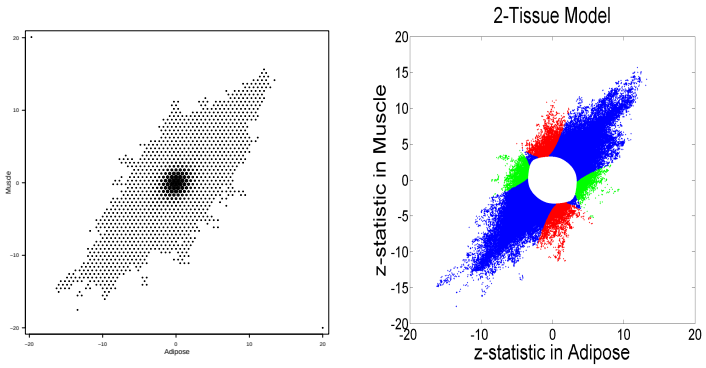


Figure 4.6: Scatter plots of z-statistics for adipose and muscle. Density-based scatter plot for all gene-SNP pairs (left), and significant eQTL discoveries with tissue specificity assessments from the fitted two-dimensional MT-eQTL model (right).

correlation is positive in one tissue, and negative in the other. The model reasonably identifies these points as common eQTLs. Similar phenomena are observed in Fu et al. (2012). Better biological understanding of anti-diagonal points is the subject of ongoing research.

To investigate how the use of auxiliary tissues increases statistical power of the analysis of subsets of tissues, we study a sequence of nested MT-eQTL models and focus on eQTL discoveries in a single tissue. For each of the nine tissues, we first fit the 1-dimensional model with just the primary tissue and then added other tissues one by one alphabetically to get a sequence of super-models. For each considered model, we applied the adaptive thresholding procedure to the marginal local FDR for the primary tissue, and recorded the number of significant discoveries in that tissue. Figure 4.7 shows the number of significant discoveries versus the dimension of a model. Each curve corresponds to a case where one of the nine tissues is set to be the primary tissue. In all cases, the number of eQTL discoveries in the primary tissue increases with the dimension of a model.

4.7 Discussion and Future Work

In this chapter, we proposed a hierarchical Bayesian model, MT-eQTL, for multi-tissue eQTL analysis. We adopted an empirical Bayes approach to estimate the model and to perform inferences. The proposed methodology greatly enhances classical single-tissue eQTL

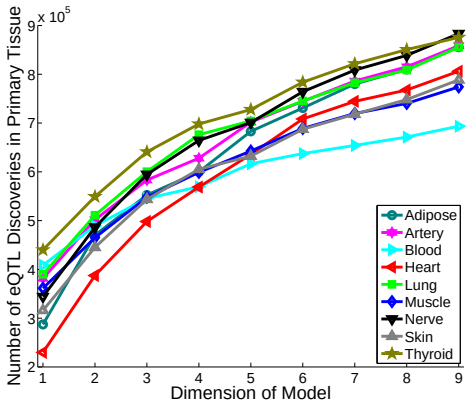


Figure 4.7: The number of significant discoveries in a primary tissue versus the dimension of a MT-eQTL model. Each curve corresponds to a case where one of the nine tissues is set to be the primary tissue. The FDR threshold is fixed to be 0.05.

analysis methods by accounting for the information shared among tissues. In particular, our method has the following desirable features.

- The MT-eQTL model can be used to analyze complex multi-tissue data where the sample sizes and degrees of overlap may vary from tissue to tissue. We directly models z-statistic vectors, i.e., Fisher-transformed covariate-adjusted correlation vectors between genes and SNPs across tissues, which facilitate interpretation and visualization.

- The model captures the presence or absence of an eQTL in each tissue simultaneously. It also explicitly reflects the heterogeneity of effect size variations across tissues. In addition, the model is able to identify correlations of effect sizes among tissues arising from experimental factors and from biological factors.

- The local FDR, which accounts for the distribution under the alternative when testing the presence of an eQTL, can be easily computed from the model. It can be generalized to analyze subsets of tissues while accounting for information in auxiliary tissues. The step-up thresholding procedure for the local FDR effectively control the overall FDR.

- The assessment of tissue specificity using the MAP rule or the other proposed methods takes into account the data across tissues simultaneously, leading to a reasonable global view of configuration assessments.

The MT-eQTL approach is applied to the new, 9-tissue data set from the GTEx initiative. To our knowledge, this is the first attempt to jointly analyze eQTLs in multiple tissues using the GTEx data. Our analysis results provide useful directions for follow-up biological research. Altogether, we anticipate that the proposed approach could have a significant contribution to the eQTL analysis of the emerging multi-tissue data. The R code for MT-eQTL is available from http://www.bios.unc.edu/research/genomic_software/Multi-Tissue-eQTL/. The Matlab code is available upon request.

The following topics pertaining to multi-tissue eQTL analysis need further investigation:

- Theoretically justify that the use of auxiliary tissues increases statistical power of the analysis of subsets of tissues. Previously, we have shown in the simulation and real data study that accounting more auxiliary tissues increases the number of significant discoveries in the primary subset of tissues. This result is intuitive since auxiliary tissues may provide additional information. However, this is only a conjecture since no rigorous statement has been proved yet.

- Study trans eQTLs. Currently, we only focus on cis gene-SNP pairs where genes and SNPs are close in distance. For trans eQTL study, the number of candidate gene-SNP pairs is tremendous while the true eQTLs are rare. Increasing computational efficiency and controlling error rate are of particular interest and difficulty.

- Study gene-level eQTLs. An important feature of our model is that each gene-SNP pair is treated independently in the multiple testing procedure. However, local SNPs are usually highly correlated (i.e., referred to as LD blocks). Analyzing gene-level eQTLs and specifying causal SNPs may be of practical interest and make more biological sense.

- Adapt the tissue-specificity assessment to the many-tissue situation. As the number of tissues increases, the number possible configurations grows exponentially. When the number of tissues is really large (say, 20, the corresponding number of configurations is $2^{20} \approx 10^6$), the data may not provide enough power to distinguish all configurations. Particularly, the "adjacent" ones (i.e., two binary indicator vectors only differ in few

entries) may not be well separable. In that case, more robust assessment methods for tissue specificity are needed. One possible solution is to cluster similar configurations.

4.8 Appendix

4.8.1 Proof of Lemma 4.2.1

Proof. Let S be a subset of $\{1, \ldots, K\}$ with cardinality $|S| = r$. It follows from the defining properties of the multivariate normal distribution that if $\mathbf{U} \sim \mathcal{N}_K(\boldsymbol{\mu}, A)$ then $\mathbf{U}_S \sim \mathcal{N}_r(\boldsymbol{\mu}_S, A_S)$. It therefore follows from (4.6) that

$$\mathbf{Z}_S \;\sim\; \sum_{\boldsymbol{\gamma}\in\{0,1\}^K} p_{\boldsymbol{\gamma}}\, \mathcal{N}_r\big((\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma})_S,\, (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)_S\big) \tag{4.17}$$

Here and in the remainder of the proof we follow the convention that $\boldsymbol{\gamma}$ ranges over $\{0,1\}^K$, and $\boldsymbol{\zeta}$ ranges over $\{0,1\}^r$. Elementary arguments show that

$$(\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma})_S \;=\; \boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\gamma}_S \;\text{ and }\; (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)_S \;=\; \Delta_S + \Sigma_S \cdot \boldsymbol{\gamma}_S \boldsymbol{\gamma}_S^T$$

It then follows from (4.17) that

$$\begin{aligned}
\mathbf{Z}_S \;&\sim\; \sum_{\boldsymbol{\gamma}\in\{0,1\}^K} p_{\boldsymbol{\gamma}}\, \mathcal{N}_r\big(\boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\gamma}_S,\, \Delta_S + \Sigma_S \cdot \boldsymbol{\gamma}_S \boldsymbol{\gamma}_S^T\big) \\
&=\; \sum_{\boldsymbol{\zeta}\in\{0,1\}^r} \sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}_S=\boldsymbol{\zeta}} p_{\boldsymbol{\gamma}}\, \mathcal{N}_r\big(\boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\gamma}_S,\, \Delta_S + \Sigma_S \cdot \boldsymbol{\gamma}_S \boldsymbol{\gamma}_S^T\big) \\
&=\; \sum_{\boldsymbol{\zeta}\in\{0,1\}^r} \mathcal{N}_r\big(\boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\zeta},\, \Delta_S + \Sigma_S \cdot \boldsymbol{\zeta}\boldsymbol{\zeta}^T\big) \sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}_S=\boldsymbol{\zeta}} p_{\boldsymbol{\gamma}} \\
&=\; \sum_{\boldsymbol{\zeta}\in\{0,1\}^r} p_{\boldsymbol{\zeta},S}\, \mathcal{N}_r\big(\boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\zeta},\, \Delta_S + \Sigma_S \cdot \boldsymbol{\zeta}\boldsymbol{\zeta}^T\big),
\end{aligned}$$

which is the desired expression for distribution of $\mathbf{Z}_S$. □

4.8.2 Proof of Theorem 4.4.1

Lemma 4.8.1. *Let U be a bounded, non-negative random variable. For $t \geq 0$ define*

$$G(t) \;=\; \mathbb{E}[\,U\,|\,U \leq t\,] \;=\; \frac{\mathbb{E}[\,U\,\mathbb{I}(U \leq t)\,]}{\mathbb{P}(U \leq t)}. \tag{4.18}$$

Then the following hold:

1. *G is non-decreasing and right continuous;*
2. *If $\mathbb{P}(U = t) = 0$ then G is continuous at t;*
3. *If $\mathbb{P}(a < U < b) > 0$ for each $0 < a < b < L$ then G is strictly increasing on $(0, L)$.*

Proof. To show that G is non-decreasing it suffices to show that $G(t+\delta) - G(t) \geq 0$ for each fixed $t \geq 0$ and $\delta > 0$. If $G(t) = 0$ then the result is immediate as the function G is non-negative. If $G(t)$ is positive, then

$$\begin{aligned} G(t+\delta) - G(t) \;&=\; \frac{\mathbb{E}[\,U\,\mathbb{I}(U \leq t+\delta)\,]}{\mathbb{P}(U \leq t+\delta)} \;-\; \frac{\mathbb{E}[\,U\,\mathbb{I}(U \leq t)\,]}{\mathbb{P}(U \leq t)} \\ &=\; \frac{\mathbb{E}[\,U\,\mathbb{I}(U \leq t+\delta)\,]\,\mathbb{P}(U \leq t) \;-\; \mathbb{E}[\,U\,\mathbb{I}(U \leq t)\,]\,\mathbb{P}(U \leq t+\delta)}{\mathbb{P}(U \leq t+\delta)\,\mathbb{P}(U \leq t)}. \end{aligned}$$

By elementary arguments the numerator of the last fraction can be expressed as

$$\begin{aligned} &\mathbb{E}[\,U\,\mathbb{I}(t < U \leq t+\delta)\,]\,\mathbb{P}(U \leq t) \;-\; \mathbb{E}[\,U\,\mathbb{I}(U \leq t)\,]\,\mathbb{P}(t < U \leq t+\delta) \\ &\quad\geq\; t\,\mathbb{P}(t < U \leq t+\delta)\,\mathbb{P}(U \leq t) \;-\; t\,\mathbb{P}(U \leq t)\,\mathbb{P}(t < U \leq t+\delta) \\ &\quad=\; 0. \end{aligned} \tag{4.19}$$

Thus G is non-decreasing. Right continuity of G follows by applying the monotone convergence theorem to the numerator and denominator in (4.18). If $\mathbb{P}(U = t) = 0$ then continuity of G at t follows from the dominated convergence theorem in a similar fashion. Finally, if $\mathbb{P}(t < U < t+\delta) > 0$ then the inequality in (4.19) is strict, and the final claim follows by considering $t \in [0, L)$ and $\delta > 0$ such that $t + \delta < L$. □

Lemma 4.8.2. *For $i = 0, \ldots, m$ let f_i be the density of the d-variate normal distribution $\mathcal{N}_d(\mu_i, \Sigma_i)$ and let $c_1, \ldots, c_m$ be positive constants. If at least one of $f_1, \ldots, f_m$ is not equal to f_0, then*

$$m_d\Big(\Big\{x : f_0(x) = \textstyle\sum_{j=1}^m c_j\, f_j(x)\Big\}\Big) = 0$$

where $m_d(\cdot)$ denotes Lebesgue measure on $\mathbb{R}^d$.

Proof. Define $h(x) = f_0(x) - \sum_{j=1}^m c_j\, f_j(x)$ and let $A = \{x : h(x) = 0\}$. As h is continuous, A is a closed subset of $\mathbb{R}^d$. We establish the result by way of contradiction. Consider first the case in which $d = 1$ and $h(x) = 0$ for each $x \in \mathbb{R}$. By an easy argument, we can assume that the densities f_i, $i = 0, 1, \ldots, m$ are distinct and that $m \geq 1$. Let μ_i and σ_i be, respectively, the mean and variance of the distribution specified by the density f_i. Let (σ_j, μ_j) be the largest element, under the usual lexicographic order, of the set $\{(\sigma_i, \mu_i) : 0 \leq i \leq m\}$. Considering the limit of $h(x)/f_j(x)$ as x tends to infinity, we conclude that $c_j = 0$ if $j \neq 0$ or $1 = 0$ if $j = 0$. In either case we obtain a contradiction, and therefore $h(x)$ cannot be identically equal to zero.

The remainder of the proof proceeds by induction on d. Consider first the case $d = 1$. Note that $h(x)$ is an analytic function of the real variable x. If $m_1(A) > 0$ then there exists $M < \infty$ such that $m_1(A \cap [-M, M]) > 0$. In particular, there are infinitely many points of A in the compact set $[-M, M]$. Thus A has a limit point x_0, and $h(x_0) = 0$ as A is closed. As the zeros of a non-zero analytic function are necessarily isolated, it follows that $h(x)$ is identically zero. This contradicts the argument given above, and we conclude that $m_1(A) = 0$.

Assume now that the lemma holds for dimensions $1, \ldots, d-1$, and consider the general case of dimension d. Suppose that $m_d(A) > 0$. By Fubini's theorem, there exist a Borel measurable set $B \subset \mathbb{R}$ such that (i) $m_1(B) > 0$ and (ii) for every $x_d \in B$ the section

$$A(x_d) = \{x_1^{d-1} : (x_1^{d-1}, x_d) \in A\} \subseteq \mathbb{R}^{d-1}$$

has $(d-1)$-dimensional Lebesgue measure greater than zero. (Here x_1^{d-1} denotes the ordered

sequence $x_1, \ldots, x_{d-1}$.) Note that $h(x) = 0$ can be written in the equivalent form

$$0 \;=\; f_0(x_1^{d-1} \,|\, x_d)\, f_0(x_d) \;-\; \sum_{j=1}^{m} c_j\, f_j(x_1^{d-1} \,|\, x_d)\, f_j(x_d) \qquad x \in A \tag{4.20}$$

where $f_j(x_1^{d-1} \,|\, x_d)$ denotes the conditional density of x_1^{d-1} given x_d under f_j, and $f_j(x_d)$ denotes the marginal density of x_d under f_j. If for each $x_d \in B$ the conditional densities $f_j(x_1^{d-1} \,|\, x_d)$ are equal on $A(x_d)$ then (4.20) becomes

$$0 \;=\; f_0(x_d) \;-\; \sum_{j=1}^{m} c_j\, f_j(x_d) \qquad x_d \in B,$$

which contradicts the induction hypothesis. Suppose then that for some $x_d \in B$ the conditional densities $f_j(x_1^{d-1} \,|\, x_d)$ are not all equal on $A(x_d)$. Then equation (4.20) becomes

$$0 \;=\; f_0(x_1^{d-1} \,|\, x_d) \;-\; \sum_{j=1}^{m} c'_j\, f_j(x_1^{d-1} \,|\, x_d) \qquad x_1^{d-1} \in A(x_d)$$

where $c'_j = c_j\, f_j(x_d)/f_0(x_d)$. Our assumption regarding the conditional densities ensures that $f_j(x_1^{d-1} \,|\, x_d)$ is different from $f_0(x_1^{d-1} \,|\, x_d)$ for some $j \geq 1$, again contradicting the induction hypothesis. This completes the proof. □

Lemma 4.8.3. *Let $\eta(\mathbf{z})$ be defined as in (4.12) and assume that every diagonal entry of Σ is positive. Then the following hold.*

1. $\inf_{\mathbf{z} \in \mathbb{R}^d} \eta(\mathbf{z}) = 0$.

2. *For every $c \geq 0$ the Lebesgue measure of the set $\{\mathbf{z} : \eta(\mathbf{z}) = c\}$ in $\mathbb{R}^K$ is zero.*

Proof. **Proof of 1:** As $\eta(z)$ is always positive, it is enough to show that there exists $\mathbf{z} \in \mathbb{R}^d$ and $\boldsymbol{\gamma} \in \{0,1\}^K$ such that $f_{\mathbf{0}}(b\mathbf{z})/f_{\boldsymbol{\gamma}}(b\mathbf{z}) \to 0$ as $b \to \infty$. From the exponential form of the multivariate normal densities, it can be seen that the last relation will hold if the matrix $\Delta^{-1} - (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1}$ has an eigenvalue greater than zero.

Let $\mathbf{x}_0$ be an eigenvector of the matrix Δ corresponding to the smallest eigenvalue $\lambda_{\min}(\Delta)$ (which is positive by assumption). Assume without loss of generality that $||\mathbf{x}_0|| = 1$. Using the variational formula for eigenvalues, and the relationship between the eigenvalues of a matrix and those of its inverse, we find that

$$\begin{aligned}
\lambda_{\max}(\Delta^{-1} - (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1}) &= \max_{z:||z||=1} z^T(\Delta^{-1} - (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1})z \\
&\geq \max_{z:||z||=1} z^T\Delta^{-1}z \; - \; \max_{z:||z||=1} z^T(\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1}z \\
&= \lambda_{\max}(\Delta^{-1}) \; - \; \lambda_{\max}((\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1}) \\
&= \lambda_{\min}(\Delta) \; - \; \lambda_{\min}(\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T) \\
&\geq \mathbf{x}_0^T\Delta\mathbf{x}_0 \; - \; \mathbf{x}_0^T(\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)\mathbf{x}_0 \\
&= \mathbf{x}_0^T(\Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)\mathbf{x}_0
\end{aligned}$$

Let $1 \leq i \leq K$ be any index for which $x_{0,i} \neq 0$. If $\boldsymbol{\gamma}$ is the binary K-vector having a 1 in position i and all other entries equal to 0, then it is easy to see that the last expression above is $\sigma_{ii}\, x_{0,i}^2$, which is positive.

Proof of 2: This follows immediately from Lemma 4.8.2 □

Proposition 4.8.1. *The function $F(t)$ defined in (4.13) is continuous and strictly increasing on the interval $(0, L_\eta)$, where $L_\eta = \sup_{\mathbf{z}\in\mathbb{R}^d} \eta(\mathbf{z}) < 1$.*

Proof. Note that $F(t)$ is of the form $g(t)$ in (4.18) with $U = \eta(\mathbf{Z})$. Part 2 of Lemma 4.8.3 establishes that $\mathbb{P}(\eta(bZ) = t) = 0$, and continuity of F then follows from Lemma 4.8.1. For $0 < a < b < L_\eta$ we have

$$\mathbb{P}(a < \eta(\mathbf{Z}) < b) \; = \; \mathbb{P}(\eta(\mathbf{Z}) \in (a, b)) \; = \; \mathbb{P}(\mathbf{Z} \in \eta^{-1}(a, b)).$$

As $\eta(\mathbf{z})$ is continuous $\eta^{-1}(a, b)$ is an open subset of $\mathbb{R}^d$. Moreover, $\eta^{-1}(a, b)$ is non-empty by Part 1 of Lemma 4.8.3. Thus $\mathbb{P}(a < \eta(\mathbf{Z}) < b) > 0$ as the density f of $\mathbf{Z}$ is positive on $\mathbb{R}^d$. Continuity of $F(t)$ then follows from Lemma 4.8.1. □

Lemma 4.8.4. *Let $G_1, G_2, \ldots : [0,1] \to \mathbb{R}$ be non-decreasing functions. For fixed $\alpha \in (0, L_\eta)$ define $\theta_n = \sup\{t : G_n(t) \leq \alpha\}$ and let $\theta \in (0,1)$ be the unique number such that $F(\theta) = \alpha$. If $G_n(t) \to F(t)$ for each t in a dense subset T of $[0,1]$ then $\theta_n \to \theta$.*

Proof. Suppose by way of contradiction that $|\theta_n - \theta| \not\to 0$. Then there exists $\delta_1, \delta_2 > 0$ such that $\{\theta - \delta_1, \theta + \delta_2\} \subseteq T$ and an infinite subsequence n_k of $1, 2, \ldots$ such that either $\theta_{n_k} \leq \theta - 2\delta_1$ for each $k \geq 1$ or $\theta_{n_k} \geq \theta + 2\delta_2$ for each $k \geq 1$. In the first case, the definition of θ_n and the monotonicity of G_n imply

$$\alpha \;\leq\; G_{n_k}(\theta_{n_k} + \delta_1) \;\leq\; G_{n_k}(\theta - \delta_1)$$

Taking limits as $k \to \infty$ we find $\alpha \;\leq\; F(\theta - \delta_1) < \alpha$ as F is strictly increasing, which is a contradiction. In the second case, a similar argument shows that

$$\alpha \;\geq\; G_{n_k}(\theta_{n_k} - \delta_2) \;\geq\; G_{n_k}(\theta + \delta_2).$$

Taking limits as $k \to \infty$ yields $\alpha \;\geq\; F(\theta + \delta_2) > \alpha$, which is again a contradiction. This concludes the proof. □

Proof of Theorem 4.4.1:

Proof. Let $\hat{\theta}_n = \sup\{t : \hat{F}_n(t) \leq \alpha\}$ and let θ be the unique number such that $F(\theta) = \alpha$. We claim that $\hat{\theta}_n \to \theta$ in probability. To show this, assume to the contrary that there exists $\delta > 0$ and a subsequence n_k such that

$$\mathbb{P}\big(|\hat{\theta}_{n_k} - \theta| > \delta\big) > \delta \quad \text{for each} \quad k \geq 1. \tag{4.21}$$

Let T be any countable, dense subset of $[0,1]$. Our assumptions imply that $\hat{F}_n(t) \to F(t)$ in probability for each $t \in T$. By a standard diagonalization argument, there exists a subsequence m_k of n_k such that $\hat{F}_{m_k}(t) \to F(t)$ with probability one for each $t \in T$. It then follows from Lemma 4.8.4 that $\hat{\theta}_{m_k} \to \theta$ with probability one, which contradicts (4.21).

In order to establish the theorem, it will be convenient to work with version of M_n and N_n in which the data-dependent threshold $\hat{\theta}_n$ is replaced by the limiting value θ. Define

$$\tilde{M}_n \;=\; \sum_{\lambda\in\Lambda_n} \mathbb{I}(\mathbf{\Gamma}_\lambda = 0)\,\mathbb{I}(\eta(\mathbf{Z}_\lambda)\le\theta) \quad \text{and} \quad \tilde{N}_n \;=\; \sum_{\lambda\in\Lambda_n} \mathbb{I}(\eta(\mathbf{Z}_\lambda)\le\theta)$$

Note that $\mathbb{E}\tilde{N}_n = |\Lambda_n|\cdot\mathbb{P}(\eta(\mathbf{Z})\le\theta)$. By an elementary conditioning argument,

$$\begin{aligned}
\mathbb{E}\tilde{M}_n &= \sum_{\lambda\in\Lambda_n} \mathbb{E}\Big\{\mathbb{P}(\mathbf{\Gamma}_\lambda = 0\,|\,\mathbf{Z}_\lambda)\,\mathbb{I}(\eta(\mathbf{Z}_\lambda)\le t_n(\alpha))\Big\} \\
&= \sum_{\lambda\in\Lambda_n} \mathbb{E}\Big\{\eta(\mathbf{Z}_\lambda)\,\mathbb{I}(\eta(\mathbf{Z}_\lambda)\le t_n(\alpha))\Big\} \\
&= |\Lambda_n|\cdot\mathbb{E}[\eta(\mathbf{Z})\,\mathbb{I}(\eta(\mathbf{Z})\le t)].
\end{aligned}$$

For each $\delta > 0$,

$$\begin{aligned}
\mathbb{E}|\tilde{N}_n - N_n| &\le \sum_{\lambda\in\Lambda_n} \mathbb{P}(\eta(\mathbf{Z}_\lambda)\in[\hat{\theta}_n,\theta]\cup[\theta,\hat{\theta}_n]) \\
&\le |\Lambda_n|\,\big[\mathbb{P}\big(\eta(\mathbf{Z})\in(\theta-\delta,\theta+\delta)\big) + \mathbb{P}\big(|\hat{\theta}_n-\theta|\ge\delta\big)\big].
\end{aligned}$$

As $\hat{\theta}_n\to\theta$ in probability and the distribution of $\eta(\mathbf{Z})$ has no point masses, the last inequality implies that $\mathbb{E}|\tilde{N}_n - N_n| = |\Lambda_n|\cdot o(1)$. A similar argument shows that $\mathbb{E}|\tilde{M}_n - M_n| = |\Lambda_n|\cdot o(1)$. Thus as n tends to infinity,

$$\begin{aligned}
\frac{\mathbb{E}M_n}{\mathbb{E}N_n} &= \frac{\mathbb{E}\tilde{M}_n + |\Lambda_n|\cdot o(1)}{\mathbb{E}\tilde{N}_n + |\Lambda_n|\cdot o(1)} \\
&= \frac{\mathbb{E}[\eta(\mathbf{Z})\,\mathbb{I}(\eta(\mathbf{Z})\le\theta)] + o(1)}{\mathbb{P}(\eta(\mathbf{Z})\le\theta) + o(1)} \\
&\to \frac{\mathbb{E}[\eta(\mathbf{Z})\,\mathbb{I}(\eta(\mathbf{Z})\le\theta)]}{\mathbb{P}(\eta(\mathbf{Z})\le\theta)} \;=\; F(\theta) \;=\; \alpha.
\end{aligned}$$

This completes the proof of the theorem. □

4.8.3 GTEx Estimations

The estimated model parameters Δ and Σ for the GTEx data are given below. The tissues are ordered alphabetically. The parameter $\boldsymbol{\mu}_0$ was set to zero. The estimated mass function $\mathbf{p}$ is provided in a separate file due to space limitations.

$$\Delta = \begin{pmatrix} 1.0000 & 0.1704 & 0.0923 & 0.1010 & 0.1390 & 0.1409 & 0.1687 & 0.1415 & 0.1441 \\ 0.1704 & 1.0000 & 0.0960 & 0.1179 & 0.1518 & 0.1460 & 0.1942 & 0.1336 & 0.1491 \\ 0.0923 & 0.0960 & 1.0000 & 0.0779 & 0.1312 & 0.0780 & 0.1007 & 0.0890 & 0.1032 \\ 0.1010 & 0.1179 & 0.0779 & 1.0000 & 0.1268 & 0.1192 & 0.1093 & 0.0893 & 0.1247 \\ 0.1390 & 0.1518 & 0.1312 & 0.1268 & 1.0000 & 0.1188 & 0.1543 & 0.1220 & 0.1767 \\ 0.1409 & 0.1460 & 0.0780 & 0.1192 & 0.1188 & 1.0000 & 0.1366 & 0.1095 & 0.1258 \\ 0.1687 & 0.1942 & 0.1007 & 0.1093 & 0.1543 & 0.1366 & 1.0000 & 0.1372 & 0.1477 \\ 0.1415 & 0.1336 & 0.0890 & 0.0893 & 0.1220 & 0.1095 & 0.1372 & 1.0000 & 0.1097 \\ 0.1441 & 0.1491 & 0.1032 & 0.1247 & 0.1767 & 0.1258 & 0.1477 & 0.1097 & 1.0000 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 4.2692 & 4.5320 & 4.1062 & 3.2993 & 4.6078 & 4.0864 & 4.2076 & 3.9694 & 4.4595 \\ 4.5320 & 5.4178 & 4.4545 & 3.6526 & 5.0411 & 4.5731 & 4.6975 & 4.3167 & 5.0072 \\ 4.1062 & 4.4545 & 6.1588 & 3.3196 & 5.0385 & 4.2452 & 4.0646 & 4.0090 & 4.5213 \\ 3.2993 & 3.6526 & 3.3196 & 3.2123 & 3.7223 & 3.6852 & 3.3418 & 3.1225 & 3.7332 \\ 4.6078 & 5.0411 & 5.0385 & 3.7223 & 5.5488 & 4.5088 & 4.6816 & 4.5263 & 5.2369 \\ 4.0864 & 4.5731 & 4.2452 & 3.6852 & 4.5088 & 5.1569 & 4.0399 & 3.9304 & 4.3674 \\ 4.2076 & 4.6975 & 4.0646 & 3.3418 & 4.6816 & 4.0399 & 4.5993 & 4.0265 & 4.6699 \\ 3.9694 & 4.3167 & 4.0090 & 3.1225 & 4.5263 & 3.9304 & 4.0265 & 4.3420 & 4.4163 \\ 4.4595 & 5.0072 & 4.5213 & 3.7332 & 5.2369 & 4.3674 & 4.6699 & 4.4163 & 5.6492 \end{pmatrix}.$$

REFERENCES

Allen, G. I. (2013). Sparse and functional principal components analysis. *arXiv preprint arXiv:1309.2895*.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148.

Aneiros, G. and Vieu, P. (2014). Variable selection in infinite-dimensional problems. *Statistics & Probability Letters*, 94:12–20.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., and Yom-Tov, G. B. (2011). Patient flow in hospitals: A data-based queueing-science perspective. *New York University Technical Report*.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bongiorno, E. G., Salinelli, E., Goia, A., and Vieu, P. (2014). *Contributions in Infinite-dimensional Statistics and Related Topics*. Società Editrice Esculapio.

Brown, C. D., Mangravite, L. M., and Engelhardt, B. E. (2013). Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genetics*, 9(8):e1003649.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50.

Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4):509–540.

Bullaughey, K., Chavarria, C. I., Coop, G., and Gilad, Y. (2009). Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Human Molecular Genetics*, 18(22):4296–4303.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481.

Chand, S. (2012). On tuning parameter selection of lasso-type methods-a monte carlo study. In *Applied Sciences and Technology (IBCAST), 2012 9th International Bhurban Conference on*, pages 120–129. IEEE.

Chen, K., Chan, K.-S., and Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):203–221.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.

Cook, R., Helland, I., and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):851–877.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992.

Cook, R. D. (2007). Fisher Lecture: dimension reduction in regression. *Statistical Science*, 22(1):1–26.

Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20:927–1010.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, 100(470):410–428.

Cook, R. D. and Su, Z. (2013). Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression. *Biometrika*, 100(4):939–954.

Cook, R. D. and Zhang, X. (2015). Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(Accepted):11–25.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194.

d'Aspremont, A., Bach, F., and Ghaoui, L. E. (2008). Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294.

Dawson, J. A. and Kendziorski, C. (2012). An empirical bayesian approach for identifying differential coexpression in high-throughput experiments. *Biometrics*, 68(2):455–465.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Arcelus, M. G., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science*, 325(5945):1246–1250.

Ding, J., Gudjonsson, J. E., Liang, L., Stuart, P. E., Li, Y., Chen, W., Weichenthal, M., Ellinghaus, E., Franke, A., Cookson, W., et al. (2010). Gene expression in skin and lymphoblastoid cells: refined statistical method reveals extensive overlap in cis-eQTL signals. *The American Journal of Human Genetics*, 87(6):779–789.

Dozier, B. P. and Silverstein, J. W. (2007). On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices. *Journal of Multivariate Analysis*, 98(4):678–694.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23(1):1–22.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J., Liao, Y., and Wang, W. (2014). Projected principal component analysis in factor models. *arXiv:1406.3836*.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics*, 9(5):e1003486.

Franke, L. and Jansen, R. C. (2009). eqtl analysis in humans. In *Cardiovascular Genomics*, volume 537, pages 311–328. Springer.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–20.

Fu, J., Wolfs, M. G., Deelen, P., Westra, H.-J., Fehrmann, R. S., te Meerman, G. J., Buurman, W. A., Rensen, S. S., Groen, H. J., Weersma, R. K., et al. (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genetics*, 8(1):e1002431.

Gerrits, A., Li, Y., Tesson, B. M., Bystrykh, L. V., Weersing, E., Ausema, A., Dontje, B., Wang, X., Breitling, R., Jansen, R. C., et al. (2009). Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genetics*, 5(10):e1000692.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Hays, S., Shen, H., Huang, J. Z., et al. (2012). Functional dynamic factor models with application to yield curve forecasting. *The Annals of Applied Statistics*, 6(3):870–894.

Heinzen, E. L., Ge, D., Cronin, K. D., Maia, J. M., Shianna, K. V., Gabriel, W. N., Welsh-Bohmer, K. A., Hulette, C. M., Denny, T. N., and Goldstein, D. B. (2008). Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biology*, 6(12):e1000001.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, volume 200. Springer.

Huang, J. Z., Shen, H., and Buja, A. (2008). Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics*, 2:678–695.

Huang, J. Z., Shen, H., and Buja, A. (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association*, 104(488):1609–1620.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248–264.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327.

Kendziorski, C., Newton, M., Lan, H., and Gould, M. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22(24):3899–3914.

Kendziorski, C. and Wang, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome*, 17(6):509–517.

Lee, M., Shen, H., Huang, J. Z., and Marron, J. (2010). Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804.

Lee, T. I. and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics*, 34(1):77–137.

Lee, W. and Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of Multivariate Analysis*, 111:241–255.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161.

Li, G., Shabalin, A. A., Rusyn, I., Wright, F. A., and Nobel, A. B. (2014a). An empirical Bayes approach for multiple tissue eQTL analysis. *Technical Report*.

Li, G., Shen, H., and Huang, J. Z. (2014b). Supervised sparse and functional principal component analysis. *Technical Report*.

Li, G., Yang, D., Nobel, A. B., and Shen, H. (2015). Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis*, In Press.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, pages 1–4.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7(1):523–542.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585.

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565–577.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K.-W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of computational biology*, 8(1):37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176.

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genetics*, 7(2):e1002003.

Nikolov, D. and Burley, S. (1997). Rna polymerase ii transcription initiation: a structural view. *Proceedings of the National Academy of Sciences*, 94(1):15–22.

Ortega, J. M. and Rheinboldt, W. C. (2000). *Iterative solution of nonlinear equations in several variables*, volume 30. SIAM.

Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the svd and the nonnegative matrix factorization. *Annals of Applied Statistics*, 3(2):564–594.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642.

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., Cook, S. A., et al. (2010). New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Computational Biology*, 6(4):e1000737.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer, New York.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(1):233–243.

Rockman, M. V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics*, 7(11):862–872.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717.

Shabalin, A. and Nobel, A. (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118:67–76.

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, 81(393):142–149.

Shen, D., Shen, H., and Marron, J. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *Journal of Multivariate Analysis*, 115:317–333.

Shen, H. and Huang, J. Z. (2008a). Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410.

Shen, H. and Huang, J. Z. (2008b). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24.

Silverman, B. W. and Ramsay, J. O. (2005). *Functional Data Analysis*. Springer.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–507.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206.

Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, 98(1):133–146.

Su, Z. and Cook, R. D. (2012). Inner envelopes: efficient estimation in multivariate linear regression. *Biometrika*, 99(3):687–702.

Su, Z. and Cook, R. D. (2013). Estimation of multivariate means with heteroscedastic errors using envelope models. *Statistica Sinica*, 23(1):213–230.

Sul, J. H., Han, B., Ye, C., Choi, T., and Eskin, E. (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genetics*, 9(6):e1003491.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912.

The GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494.

Wen, X., Stephens, M., et al. (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions. *The Annals of Applied Statistics*, 8(1):176–203.

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., et al. (2013). Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature genetics*, 45(10):1238–1243.

Winterbottom, A. (1979). A note on the derivation of fisher's transformation of the correlation coefficient. *The American Statistician*, 33(3):142–143.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Wright, F. A., Shabalin, A. A., and Rusyn, I. (2012). Computational tools for discovery and interpretation of expression quantitative trait loci. *Pharmacogenomics*, 13(3):343–352.

Xia, K., Shabalin, A. A., Huang, S., Madar, V., Zhou, Y.-H., Wang, W., Zou, F., Sun, W., Sullivan, P. F., and Wright, F. A. (2012). seeQTL: a searchable database for human eQTLs. *Bioinformatics*, 28(3):451–452.

Yang, D., Ma, Z., and Buja, A. (2014). A sparse svd method for high-dimensional data. *Journal of Computational and Graphical Statistics*, 23(4):923–942.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464):968–979.

Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika*, 90(1):113–125.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, L., Shen, H., and Huang, J. Z. (2013). Robust regularized singular value decomposition with application to mortality data. *The Annals of Applied Statistics*, 7(3):1540–1561.

Zhao, X., Marron, J. S., and Wells, M. T. (2004). The functional data analysis view of longitudinal data. *Statistica Sinica*, 14(3):789–808.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

Zou, H., Hastie, T., Tibshirani, R., et al. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192.