ABSTRACT

FENG, YUAN. Nonparametric Methods for High-dimensional and Longitudinal Data. (Under the direction of Luo Xiao and Eric Chi).

Multivariate response data are frequently produced in modern scientific applications. In this dissertation we focus on developing statistical and computational methods for the following types of multivariate data:

1. growth data $\left\{(T_{ij}, Y_{ij}), \quad j = 1, \cdots, m_i, \quad i = 1, \cdots, n\right\}$

   T_{ij} is the (time, measurement) pair for subject i at the jth visit, m_i is the number of visits for subject i, and n is the number of subjects.

2. multivariate response data with covariates $\{(\mathbf{y}_i, \mathbf{x}_i), \quad i = 1, \cdots, n\}$

   $\mathbf{y_i} \in \mathbb{R}^q$ is the response vector of interest and $\mathbf{x_i} \in \mathbb{R}^p$ is the covariate vector.

In Chapter 1 we use data from the Fetal Growth Longitudinal Study of the INTERGROWTH-21$^{\text{st}}$ Project to model the longitudinal dependence of fetal growth metrics with a two-stage approach. The first stage involved finding a suitable transformation of the raw measurements and applied to the longitudinal data to provide standardized deviations (Z-scores). In the second stage, the focus is to model the temporal correlation which we model by a Gaussian process with zero mean and unit variance. We consequently provide formulae and visualization tools for obtaining the correlation for each fetal measure at any two time points between 14 and 40 weeks.

In Chapter 2 we propose a sparse multivariate single index model, where responses and predictors are linked by unspecified smooth functions and multiple matrix level penalties are employed to select predictors and induce low-rank structures across responses. An alternating direction method of multipliers (ADMM) based algorithm is proposed for optimization. We demonstrate the effectiveness of proposed methods in simulation studies and an application to conducting a genetic association study.

In Chapter 3 we propose computation strategies for the penalized matrix bi-factorization approach to address some of the computational issues for the multivariate response linear regression model when the dimension of the coefficient matrix is high. We also generalize this approach to deal with matrix regression. The proposed computation strategy has good estimation accuracy and can be easily scaled to solve large size problems.

Nonparametric Methods for High-dimensional and Longitudinal Data

by
Yuan Feng

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2019

APPROVED BY:

Rui Song	Jessie Jeng
Luo Xiao Co-chair of Advisory Committee	Eric Chi Co-chair of Advisory Committee

DEDICATION

To my daughter Claire.

BIOGRAPHY

The author was born in Wuhan, China. He earned a Bachelor of Science degree from Department of Mathematics and Statistics, Wuhan University in 2013 and a Master of Business degree from Department of Statistics, National Cheng Kung University in 2015. Later in 2015, he entered the Ph.D. program of Statistics at North Carolina State University. He will graduate with a Ph.D. in Statistics in 2019.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincerest gratitude to my co-advisors Dr. Luo Xiao and Dr. Eric Chi. Both of you are knowledgeable young professionals, and your enthusiasm on research and teaching infected me deeply. I learned a lot from our discussions, and always got patient instructions and encouragements from you when I faced difficulty in my projects. Your great support enabled me to move forward towards this dissertation. Hope that we can all be the best version of ourselves in future.

I would like to thank Dr. Rui Song, Dr. Jessie Jeng and Dr. Derek Kamper for being able to serve on my committee and provide insightful comments. I would like to thank Dr. Wenbin Lu for the help as the director of graduate program. I really appreciate the richness of the graduate level courses provided by the department and would like to extend my thank all the instructors and TAs.

I would like to thank my mentors Jerry Yang and Shuling Liu at WalmartLabs and Sabrina Wan at Merck. The two unforgettable internship experience made me more confident on my learning and myself. I would also like to thank all my friends here at NCSU. You are an important part of these four years' memory, and our friendship will long last.

Finally, I would like to thank my family: little Claire, Liuyi and our parents. You are my strength. Without you I could not have reached this far.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

CORRELATION MODELS FOR MONITORING FETAL GROWTH

1.1 Introduction

During pregnancy, fetal anthropometric measures consisting of head circumference (HC), abdominal circumference (AC) and femur length (FL) are measured using ultrasound to monitor attained fetal size at a given gestational age (GA). By comparing measurements to a defined reference or standard [TW76; Cam80], fetuses with measurements at the extreme ends of the distribution (for example below the 3^{rd}, 5^{th}, or 10^{th} centiles or above the 90^{th}, 95^{th}, or 97^{th} centiles) are identified as being at increased risk of a growth disorder, such as intra-uterine growth restriction (IUGR) that may require further investigation. Growth charts, which conventionally record only cross-sectional (attained size) information, can be extended to monitor growth rate over time (velocity) [AH89]. An

assessment of velocity enables the evaluation of an individual's growth between any two time points (rate of growth). These changes observed between two time points may be used to identify those requiring closer monitoring in the case of great differences between the observed fetal size and what is expected at a specific time. Notably, fetal growth is rapid in the first and second trimester and slows towards term. An evaluation of growth velocity ought to consider the correlation of measurements from the same individual. The correlation coefficient is not constant as it is dependent on the interval between measurements. An estimate of the correlation coefficient is straightforward for fixed time intervals, but it is clinically useless as, in normal practice, fetuses are seen and measured at arbitrary time points. For pragmatic reasons, it is impossible to see and measure everyone at fixed time intervals. Therefore, a flexible model that represents the correlation as a function of time is required. To the best of our knowledge, correlation models have previously been derived for child data [Arg08a; And18; Wri94; Col98; Col94] but not for fetal biometry data.

The main aim of this article is therefore, to model the correlation of fetal biometry i.e., HC, AC, and FL and make available formulae or a tool that can be used to obtain the correlation for each fetal measure between measurements made at any two time points between 14 and 40 weeks of GA. We model the correlations using fetal ultrasound data from the INTERGROWTH-21st Project Fetal Growth Longitudinal Study (FGLS) that were used to construct the international standards for fetal growth [Pap14; Ohu18]. Further analysis of the cohort demonstrated that the FGLS cohort remained healthy with adequate growth and motor development up to 2 years of age [Vil18].

1.2 Data

The INTERGROWTH-21st Project was a population-based longitudinal study that measured serial fetal growth scans every $5^{\pm 1}$ weeks from recruitment at 9^{+0} - 13^{+6} weeks of gestation until, but not beyond, 42^{+0} weeks of gestation. The FGLS component of the INTERGROWTH-21st Project is a unique dataset as it is the largest prospective study to collect data on fetal ultrasound measurements from optimally healthy pregnant women to date, collecting data in eight geographically diverse populations and using many quality control measures. The FGLS involved measuring serial fetal

growth scans every $5^{\pm 1}$ weeks after the initial dating scan, so that the possible ranges after the dating scan were $14-18$, $19-23$, $24-28$, $29-33$, $34-38$, and $39-42$ weeks of gestation. To ensure all centres collected high-quality data that were comparable within and between the study sites, all sonographers and anthropometrists were trained, and all ultrasound measurements were performed in a standardized manner following strict protocols [Sar13]. All centres adopted uniform methods, used identical ultrasound equipment in all of the study sites, adopted standardized methodology to take fetal measurements, and employed locally accredited ultra-sonographers who underwent standardization training and monitoring.

The analysis was based on the target sample of 4,321 women (20,313 ultrasound scans) who had pregnancies without major complications and delivered live singletons without congenital malformations that contributed data for the construction of the INTERGROWTH-21st international fetal growth standards [Pap14], international gestation-specific newborn standards [Vila], gestational weight gain standards [Lei], and preterm postnatal growth standards [Vil15]. This cohort experienced very low maternal and perinatal mortality and morbidity rates [Pap14; Vila], confirming that the participants were at low risk of adverse outcome and therefore contributed to the construction of the international fetal growth standards. The baseline characteristics of the study cohort across the eight sites were very similar, which was expected because women were selected from the underlying low-risk populations using the same clinical and demographic criteria [Pap14; Vilb]. The median number of ultrasound scans (excluding the dating scan) was 5.0 (mean = 4.9, SD = 0.8, range from 4 to 7) and 97% of women had 4 scans (mean = 5.0, SD = 0.6, range from 4 to 7), indicating that participants adhered well to the protocol. Eighty-five percent of the 20,313 ultrasound scans were performed within the expected gestational age window of the protocol [Pap14].

The INTERGROWTH-21st Project was approved by the Oxfordshire Research Ethics Committee "C" (reference: 08/H0606/139), the research ethics committees of the individual participating institutions, and the corresponding regional health authorities where the project was implemented. Participants provided written consent to be involved in the study.

1.3 Statistical Methodology

Consider the longitudinal data $\{(T_{ij}, Y_{ij}), 1 \le j \le m_i, 1 \le i \le n\}$, where T_{ij} is the gestational age in weeks for subject i at the jth visit, Y_{ij} is one of the three ultrasound growth measurements continuously valued in millimeters at T_{ij}, m_i is the number of visits for subject i, and n is the number of subjects. Note that the total number of visits can be different for each subject. The goal is to construct a correlation matrix of the ultrasound measurement at different gestational ages. Fetuses will not be dichotomized by gender for growth measurement evaluation, and it is also not the norm as not all populations desire to know fetal sex. Because the marginal distributions of ultrasound growth measurements may be non-normal, e.g., skewed, a suitable transformation of the raw growth measurements is first identified and applied to the data to construct a working marginal reference chart. The raw measurements are then transformed accordingly to provide standardized deviations (Z-scores). Next the Z-scores are modeled by a Gaussian process with zero mean and unit variance so that the temporal correlation of the process can be estimated.

1.3.1 Working Models for Marginal Reference Distribution

We consider the LMS transformation [CG92] which could transform non-normal data into normal data. Let Y be a positive random variable and its LMS transformation is given by

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left\{\left(\frac{Y}{\mu}\right)^{\nu} - 1\right\} & \text{if } \nu \neq 0, \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right) & \text{if } \nu = 0. \end{cases} \tag{1.1}$$

Here $\mu, \sigma \in \mathbb{R}^+$ and $\nu \in \mathbb{R}$ are location, scale and skewness parameters, respectively. If Z has a standard normal distribution, then Y is said to follow the three-parameter Box-Cox Cole-Green distribution [CG92] denoted by BCCG(μ, σ, ν). A fourth parameter can be added to further model kurtosis: if Z has a t distribution with degrees of freedom $\tau \in \mathbb{R}^+$, then Y is said to follow the Box-Cox t distribution [RS06] denoted by BCT(μ, σ, ν, τ); if Z has a standard power exponential distribution with parameter $\tau \in \mathbb{R}^+$, then Y is said to follow the Box-Cox power exponential distribution [RS04]

denoted by BCPE(μ,σ,ν,τ). Note that BCT($\mu,\sigma,\nu,\tau=+\infty$) and BCPE($\mu,\sigma,\nu,\tau=2$) reduce to BCCG($\mu,\sigma,\nu$).

In practice, ultrasound measurements are often taken at irregular visits, i.e., observed at gestational ages that differ between subjects. Thus, it may not be feasible to find the proper transformation at each gestational age separately. Instead, a more practical and statistically efficient approach is to model the parameters in (1.1) as a smooth function of gestational age. We shall adopt such an approach and the parameters for BCCG become $\mu(t)$, $\nu(t)$ and $\sigma(t)$, where t denotes the gestational age. The additional parameters in BCT and BCPE are similarly defined. The GAMLSS method [SR07] can be used to estimate such functions. Under the LMS framework, suppose that $\{\hat{\mu}(t),\hat{\sigma}(t),\hat{\nu}(t)\}$ are the obtained estimates, then the transformed measurements Z_{ij} can be computed as

$$Z_{ij}=\begin{cases}\frac{1}{\hat{\sigma}(T_{ij})\hat{\nu}(T_{ij})}\left\{\left(\frac{Y_{ij}}{\hat{\mu}(T_{ij})}\right)^{\hat{\nu}(T_{ij})}-1\right\} & \text{if } \hat{\nu}(T_{ij})\neq 0,\\ \frac{1}{\hat{\sigma}(T_{ij})}\log\left(\frac{Y_{ij}}{\hat{\mu}(T_{ij})}\right) & \text{if } \hat{\nu}(T_{ij})=0.\end{cases} \tag{1.2}$$

Under the BCCG model, marginally Z_{ij} has approximately a standard normal distribution. Under the BCT or the BCPE models, additional transforms of Z_{ij} are needed to make Z_{ij} normal. For simplicity, we assume all proper transformations have been taken. Then we model Z_{ij} as a zero-mean Gaussian process with a constant variance of 1. The Gaussian process is fully identified by estimating its correlation matrix.

1.3.2 Correlation Models

In this section, we focus on constructing a growth correlation matrix for the Z-scores. We shall compare several parametric and nonparametric models. The parametric models considered here are those that have been used for modeling child growth correlation. The exponential model [Dig88] (denoted by P1) is

$$\text{cor}(Z_{ij},Z_{ik})=e^{\{-b|T_{ij}-T_{ik}|^a\}},$$

where $a, b \in \mathbb{R}^+$ are two unknown parameters that can be interpreted as the order and the rate of the change in the correlation. This model is commonly used due to its simple form and stationarity, i.e., the correlation depends only on the distance between two gestational ages. The second model (denoted by P2), proposed by [Arg08a] for child growth, takes the form

$$\text{cor}(Z_{ij}, Z_{ik}) = e^{\left\{-b\left|\log \frac{T_{ij}+\tau}{T_{ik}+\tau}\right|\right\}},$$

where $\tau, b \in \mathbb{R}^+$ are two unknown parameters. The model is non-stationary, but possesses the Markovian property. Indeed, via the transformation $S_{ij} = \log(T_{ij}+\tau)$ and $S_{ik} = \log(T_{ik}+\tau)$, the correlation becomes $\text{cor}(Z_{ij}, Z_{ik}) = \rho^{|S_{ij}-S_{ik}|}$, where $\rho = e^{(-b)}$. Because growth measurements might have non-ignorable measurement errors, a nugget effect term is usually added to the above correlation models. The exponential correlation with a nugget effect model (denoted by P1+) takes the form

$$\text{cor}(Z_{ij}, Z_{ik}) = \frac{1}{1+\sigma^2}\left[e^{\{-b|T_{ij}-T_{ik}|^a\}} + \sigma^2 \mathbb{1}_{\{T_{ij}=T_{ik}\}}\right],$$

where σ^2 is the variance of the measurement error in the Z-scores and $\mathbb{1}_{\{\}}$ is the indicator function which is 1 if the statement inside the bracket is true and 0 otherwise. Similarly, the P2+ correlation model has the form

$$\text{cor}(Z_{ij}, Z_{ik}) = \frac{1}{1+\sigma^2}\left[e^{\left\{-b\left|\log \frac{T_{ij}+\tau}{T_{ik}+\tau}\right|\right\}} + \sigma^2 \mathbb{1}_{\{T_{ij}=T_{ik}\}}\right].$$

Note that with the nugget term, neither the stationary property nor the Markovian property holds.

Parametric models are simple and easy to interpret, but they can be subject to model misspecification. Thus, in addition to the above parametric correlation models, we shall also consider two nonparametric correlation models. The first one is based on functional data analysis [Yao05], which models the Z-score of a subject as the sum of a smooth random function of the gestational age and

Table 1.1 Correlation models.

Model	Abbreviation	Correlation form
Exponential	P1	$e^{\{-b\|T_{ij}-T_{ik}\|^a\}}$
Exponential with nugget effect	P1+	$\frac{1}{1+\sigma^2}\left[e^{\{-b\|T_{ij}-T_{ik}\|^a\}}+\sigma^2\mathbb{1}_{\{T_{ij}=T_{ik}\}}\right]$
Markovian	P2	$e^{\left\{-b\left\|\log\frac{T_{ij}+\tau}{T_{ik}+\tau}\right\|\right\}}$
Markovian with nugget effect	P2+	$\frac{1}{1+\sigma^2}\left[e^{\left\{-b\left\|\log\frac{T_{ij}+\tau}{T_{ik}+\tau}\right\|\right\}}+\sigma^2\mathbb{1}_{\{T_{ij}=T_{ik}\}}\right]$
1st nonparametric	NP1	$\mathscr{C}(T_{ij},T_{ik})$: fully unspecified and smooth
2nd nonparametric	NP2	$\mathscr{C}(T_{ij},T_{ik})=g(\|T_{ij}-T_{ik}\|)$: g unspecified and smooth

a random measurement error term. Specifically, the functional data model is

$$Z_{ij} = b_i(T_{ij}) + \epsilon_{ij}, \tag{1.3}$$

where $b_i(\cdot)$ is a random and smooth function modeled by a zero-mean Gaussian process with a smooth covariance function $\mathscr{C}(T_{ij},T_{ik}) = \mathrm{Cov}\{b_i(T_{ij}), b_i(T_{ik})\}$, $\{\epsilon_{i1},\ldots,\epsilon_{im_i}\}$ are independent measurement errors with variance σ_ϵ^2, and $b_i(\cdot)$ is independent from the measurement errors. Such a covariance function does not impose any parametric assumption for modeling correlations between the repeated observations and is hence very flexible. Since $\mathrm{Var}(Z_{ij})=1$, the correlation of the Z-scores at two distinctive time points T_{ij} and T_{ik} with $j\neq k$ is automatically given by $\mathscr{C}(T_{ij},T_{ik})$. Thus, we call $\mathscr{C}$ the correlation function and its estimation is described in Section 1.3.3.

The correlation function from the functional data method is in general nonstationary. We shall also consider a stationary but yet nonparametric correlation by assuming that the correlation function $\mathscr{C}$ satisfies $\mathscr{C}(T_{ij},T_{ik}) = g(|T_{ij}-T_{ik}|)$, where g is a smooth but unspecified univariate function. Note that due to the presence of measurement error in the functional data model, the overall correlation between the Z-scores is still nonstationary. The estimation of g is also given in Section 1.3.3.

The various correlation models are summarized in Table 1.1.

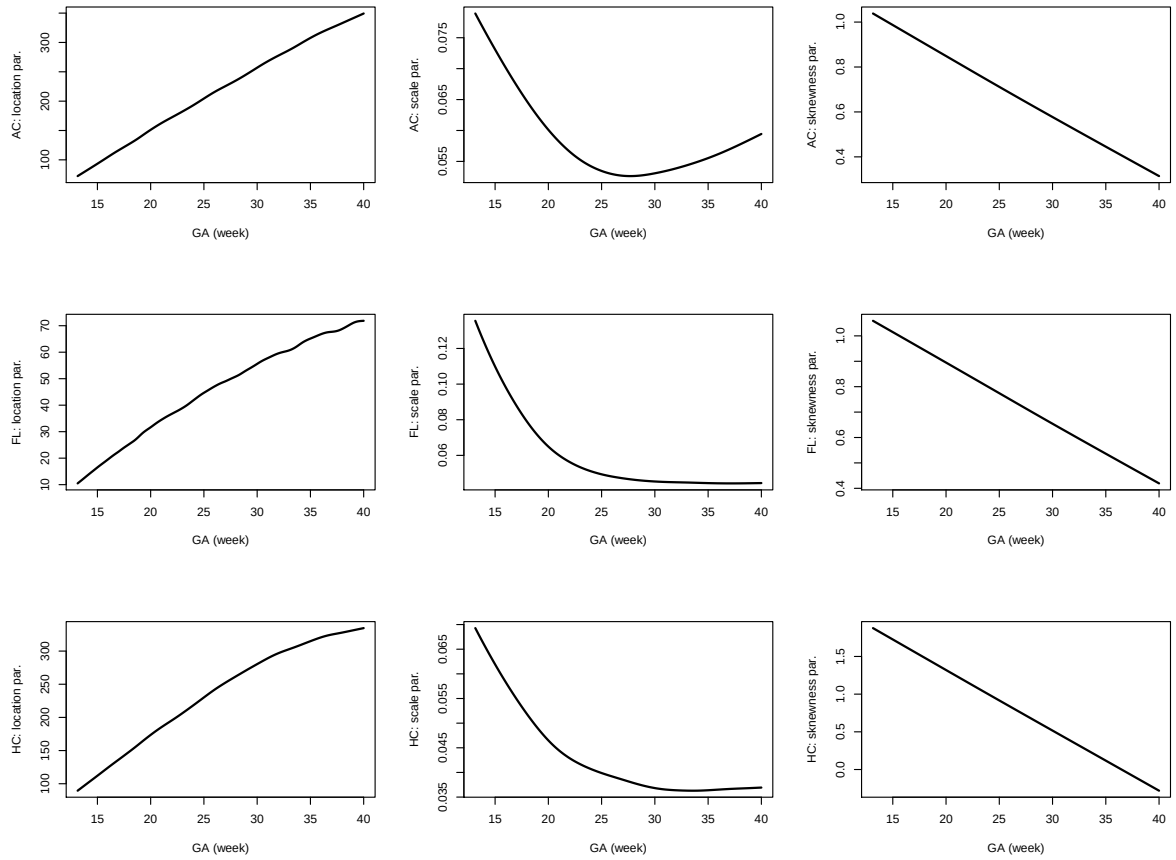

Figure 1.1 Estimated location, scale and skewness parameters as functions of gestational age for the three fetal growth measurements.

1.3.3 Estimation of the Correlation Models

The parametric correlation models can be easily estimated by maximizing likelihood of the Z-scores under normality. We now focus on the estimation of the two nonparametric models. Estimation methods for the functional data model have been well developed in the statistics literature and here we use the fast covariance estimation method for longitudinal data, developed in [Xia18]. We briefly describe the method here, which will also be useful for explaining our estimation method for the second nonparametric model. First, empirical estimates of the correlation function are constructed. Specifically, let $r_{ijk} = Z_{ij}Z_{ik}$ for $1 \le j,k \le m_i, 1 \le i \le n$. Then $\mathbb{E}(r_{ijk}) = \mathscr{C}(T_{ij},T_{ik}) + 1_{\{j=k\}}\sigma_\epsilon^2$. Thus,

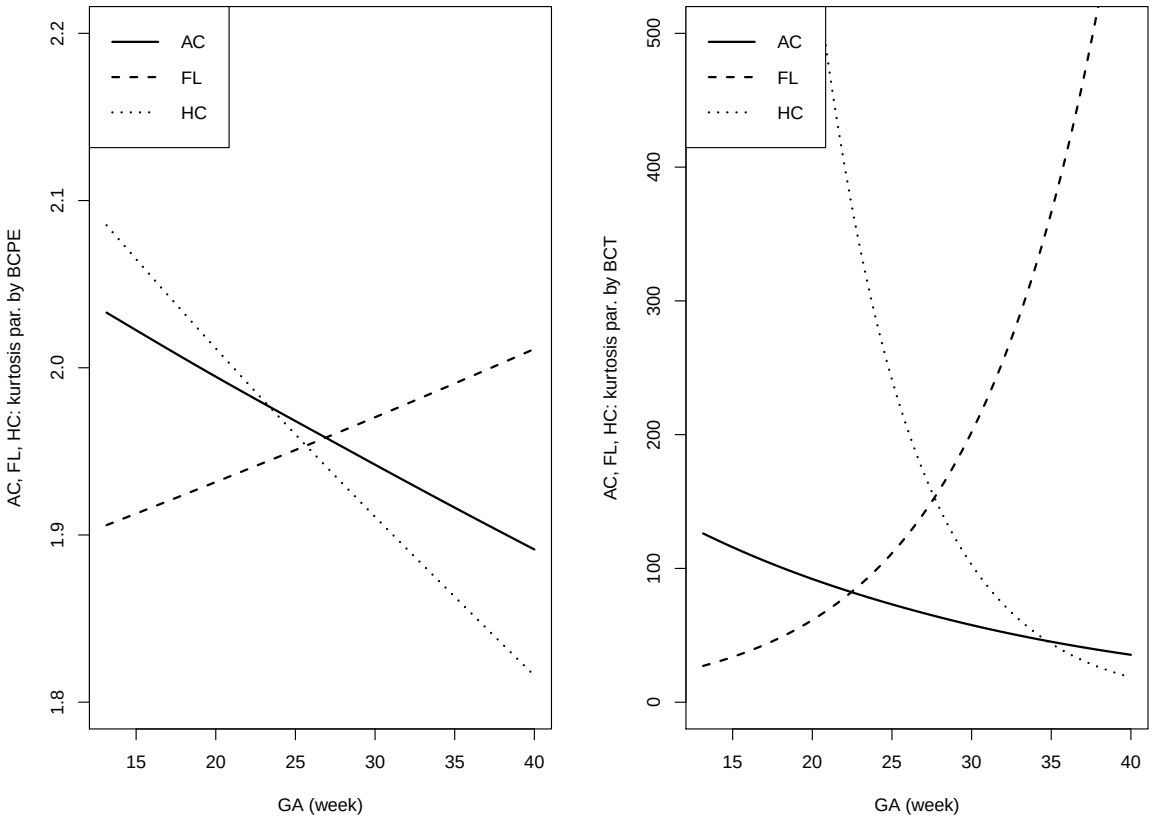

Figure 1.2 Estimated kurtosis parameters as functions of gestational age for the three fetal growth measurements using BCPE and BCT.

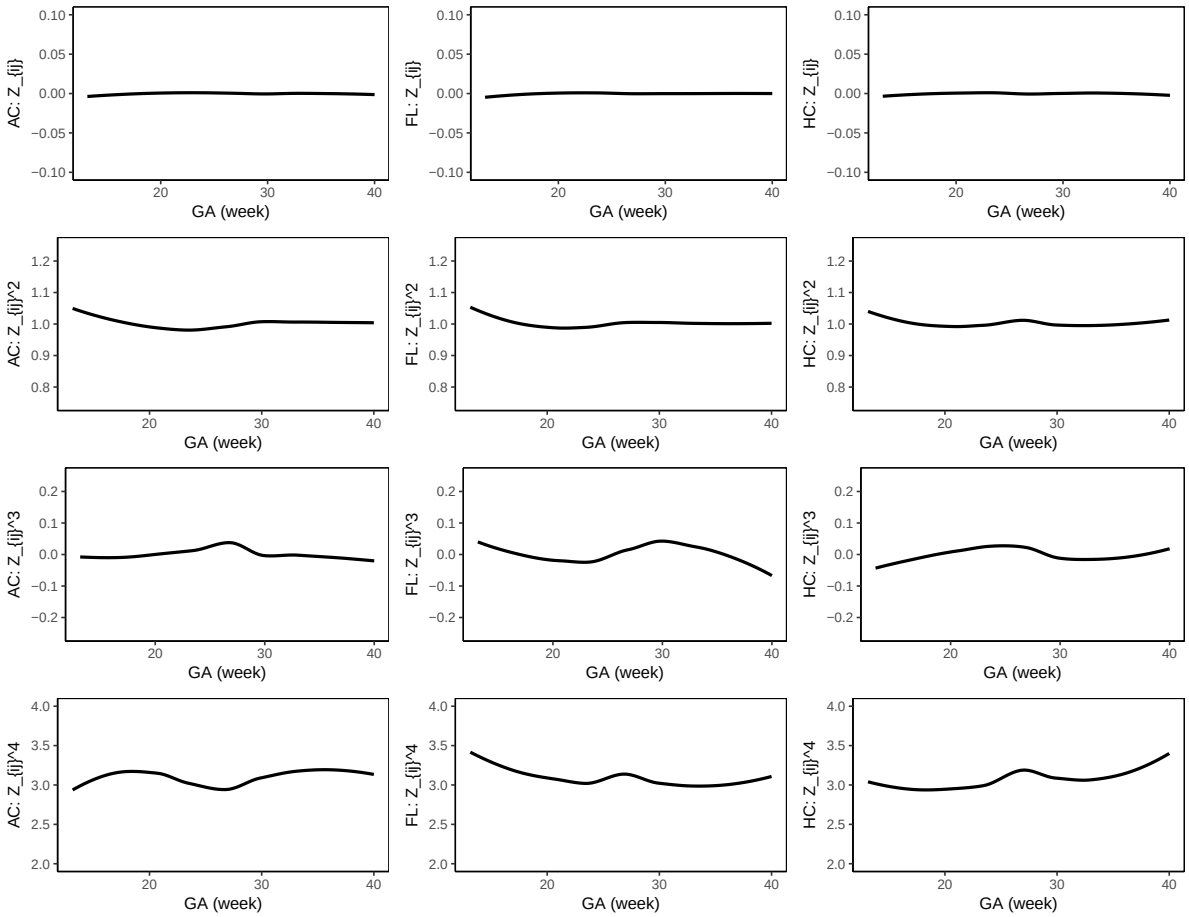

Figure 1.3 Smooth estimates of the first to fourth moments of the constructed Z-scores for AC (left panel), FL (middle panel) and HC (right panel).

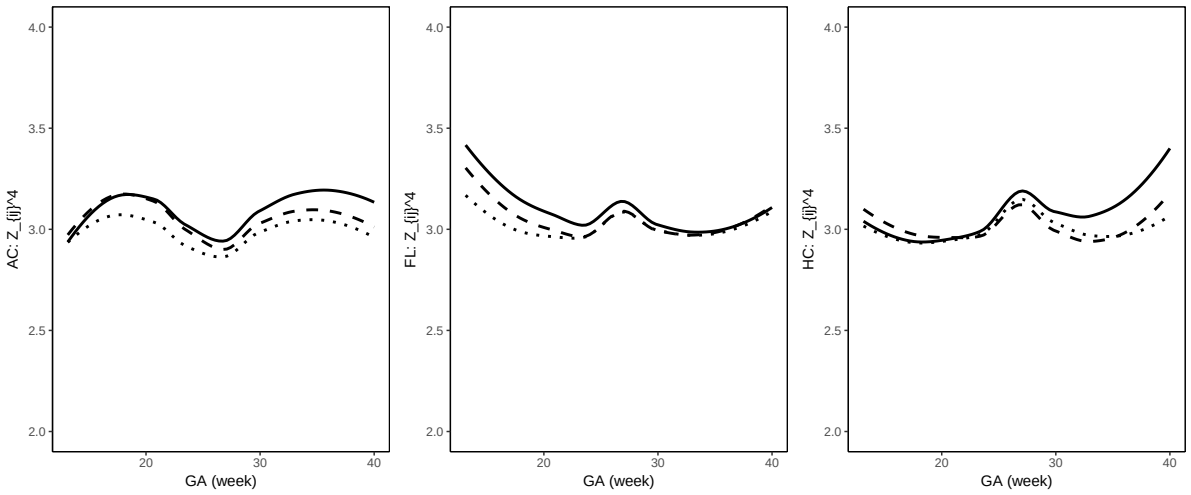

Figure 1.4 Further comparison of BCCG (solid), BCPE (dashed) and BCT (dotted) on the fourth moments of Z-scores.

r_{ijk} is an unbiased estimate of $\mathscr{C}(T_{ij}, T_{ik})$ whenever $j \neq k$. We will conduct a bivariate smoothing of the data $\{(T_{ij}, T_{ik}, r_{ijk}), 1 \leq j \neq k \leq m_i, 1 \leq i \leq n\}$ to estimate the correlation function $\mathscr{C}$. We use bivariate P-splines [EM03], which approximates the bivariate correlation function with tensor-product B-splines and employs a smoothness penalty to avoid overfit. Moreover, constraints on spline coefficients are imposed to ensure that $\mathscr{C}$ is symmetric; see [Xia18] for further details. Denote the corresponding estimate by $\hat{\mathscr{C}}(s,t)$, then we estimate the error variance σ_ϵ^2 using the equality $\mathbb{E}(r_{ijj}) = \mathscr{C}(T_{ij}, T_{ij}) + \sigma_\epsilon^2$ for $1 \leq j \leq m_i$, $1 \leq i \leq n$. For the second nonparametric model, by assumption, $\mathbb{E}(r_{ijk}) = g(|T_{ij} - T_{ik}|) + 1_{\{j=k\}}\sigma_\epsilon^2$. Thus, we smooth the data $\{(|T_{ij} - T_{ik}|, r_{ijk}), 1 \leq j \neq k \leq m_i, 1 \leq i \leq n\}$ to estimate the function g. Specifically, we use univariate P-splines [EM96], which approximates g using B-spline bases and also controls overfit using a smoothness penalty. Then the error variance σ_ϵ^2 can be estimated by the equality $\mathbb{E}(r_{ijj}) = g(0) + \sigma_\epsilon^2$ for $1 \leq j \leq m_i$, $1 \leq i \leq n$.

1.4 Results

1.4.1 Marginal Reference Charts

The estimated location, scale and skewness parameters show that a transformation model simpler than BCCG will be insufficient to fit the data marginally; see Figure 1.1. Our empirical results also indicate the sufficiency of using BCCG rather than more complicated BCPE or BCT, as Figure 1.2 suggests that the estimated parameter of kurtosis is close to 2 for BCPE model and very large for BCT model. Figure 1.3 plots the smoothed first to fourth moments of the Z-scores against the gestational age. Specifically, for $k = 1,2,3,4$, a corresponding nonparametric smooth function is estimated using the data $\{Z_{ij}^k, 1 \le j \le m_i, 1 \le i \le n\}$. If the Z-scores are indeed marginally normal, then the estimated curves should be close to the constant lines $y = 0,1,0$, and 3, respectively. Figure 1.3 suggests that the BCCG-transformed Z-scores are marginally normally distributed. A closer look at the smoothed fourth moments under different models in Figure 1.4 again suggest that BCPE and BCT are not necessary. We also note that for standardizing the raw measurements, BCCG seems to be a much better approach than the standard SDS, which presents a systematic biased estimation; see Figure Figure 1.5.

Consequently, the BCCG model will be applied to construct marginal reference charts.

1.4.2 Correlation Models

We use the BCCG model to fit the marginal distributions of the raw ultrasound measurements and then convert the transformed measurements into Z-scores. Then different parametric and nonparametric correlation models are considered and compared via model selection criteria: AIC and BIC. Both criteria require the degrees of freedom of the model. For parametric correlation models, it is the number of free parameters. For nonparametric correlation models, the effective degree of freedom, which evaluates the model complexity of nonparametric smoothers [Rup03], will be calculated.

Model comparison results for AC, FL and HC are summarized in Table 1.2. Table 1.2 shows that

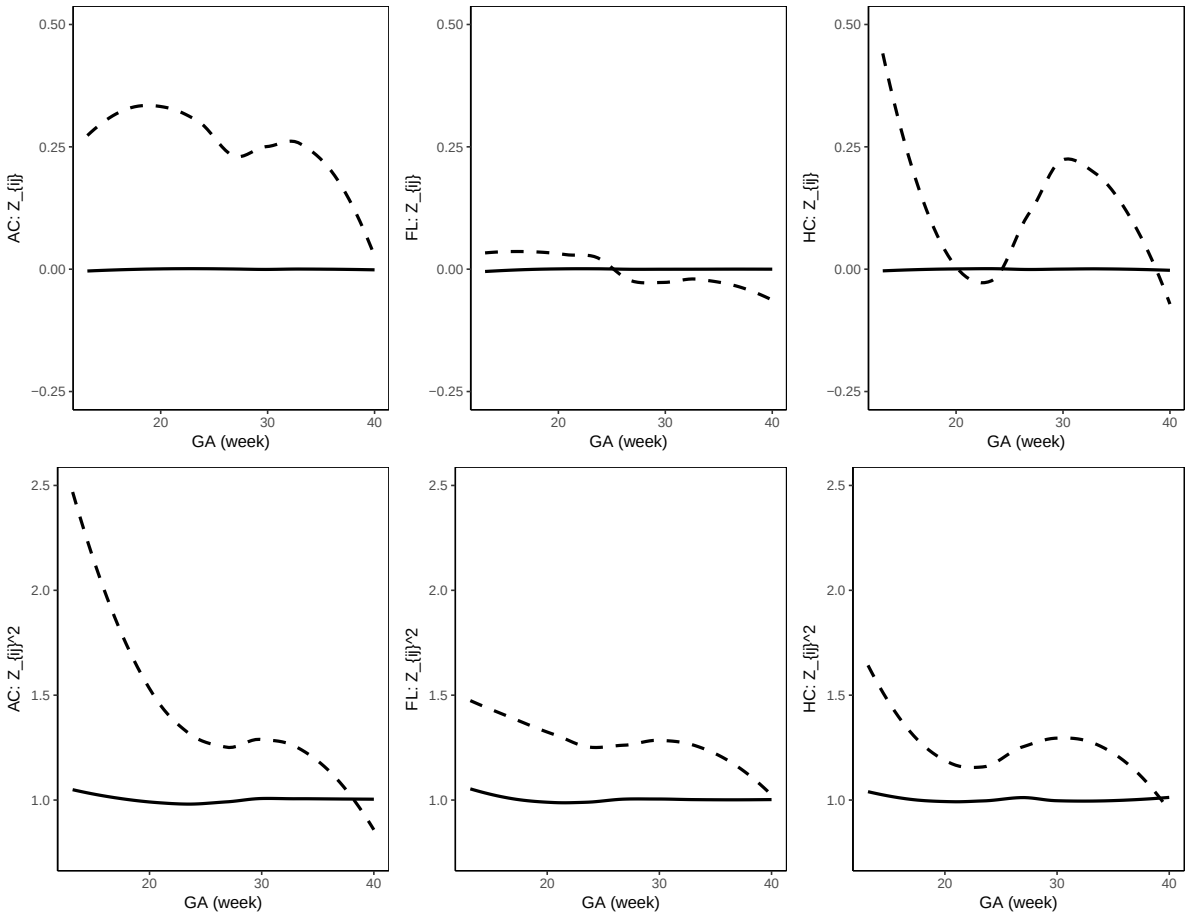

Figure 1.5 Smooth estimates of the first and second moments of Z-scores constructed via BCCG (solid) and SDS (dashed).

Table 1.2 Comparison of correlation models.

Models	**-2log-lik**	**AIC**	**BIC**	**-2log-lik**	**AIC**	**BIC**	**-2log-lik**	**AIC**	**BIC**
P1	44656.31	44660.31	44673.01	42790.43	42794.43	42807.13	41999.78	42003.78	42016.48
P1+	44505.17*	44511.17*	**44530.21***	42672.04*	**42678.04***	**42697.08***	41846.77*	41852.77*	**41871.82***
P2	45201.95	45205.95	45218.64	43471.66	43475.66	43488.36	42117.83	42121.83	42134.53
P2+	44540.30	44546.30	44565.34	42708.85	42714.85	42733.89	41943.10	41949.10	41968.15
NP1	**44489.71**	**44509.85**	44573.75	42681.81	42697.74	42748.32	**41743.67**	**41783.82**	41911.26
NP2	44524.09	44530.36	44550.24	**42670.06**	42678.54	42705.47	41863.42	41869.44	41888.52
	AC			FL			HC		

Note that * denotes the best parametric model in each column. The bold type denotes the best model in one column.

Table 1.3 MSE(·, P1+) comparison.

Model	AC	FL	HC
NP1	3.93×10^{-4}	4.46×10^{-4}	6.33×10^{-4}
NP2	2.25×10^{-4}	6.46×10^{-5}	3.75×10^{-4}
P1	2.87×10^{-3}	1.94×10^{-3}	1.98×10^{-3}
P2+	9.19×10^{-4}	3.73×10^{-4}	1.08×10^{-3}
P2	1.40×10^{-2}	1.36×10^{-2}	2.82×10^{-3}

the P1+ model is overall the best model across the three fetal growth measurements. On the one hand, P1+ fits the data best among all parametric models; on the other hand, it has a very simple form compared to nonparametric methods, and yields the smallest BIC. To quantify the differences among different correlation models, we use P1+ as the reference correlation and evaluate how the other models differ from P1+. Denote $\rho_{jk}^{\text{P1+}}$ the correlation coefficient at times (j,k) in P1+ correlation matrix, the mean squared error (MSE) of NP1, for example, to P1+ is defined as

$$\text{MSE(NP1, P1+)} = \frac{1}{(L-1)(L-2)} \sum_{1\le j<k\le L} \left(\rho_{jk}^{\text{NP1}} - \rho_{jk}^{\text{P1+}}\right)^2, \tag{1.4}$$

where L=183 is the range of gestational age in days in this study.

Table 1.3 demonstrates an ignorable difference between P1+ and NP2, which is as expected because of the stationarity nature of both models. Furthermore, difference between P1+ and NP1 is also small, suggesting that an exponential correlation model with nugget effect is sufficiently good for fetal growth measurements. Indeed, the average difference in correlation is only 0.020 for AC, 0.021 for FL and 0.025 for HC. On the other hand, the correlations from the other parametric models are relatively more divergent from those of P1+ compared to the nonparametric models NP1 and NP2, indicating the superiority of P1+ over the other parametric models.

As a result, the estimated parameters for a P1+ model are summarized in Table 1.4. For illustration, we plot the fitted correlation surface on a grid of gestational age by weeks for AC in Figure 1.6. Correlation plots for FL and HC are given in Figure A.1 and Figure A.2 in the appendix.

Table 1.4 Estimated parameters for P1+ correlation models.

Measurement	a	b	σ^2
AC	1.56	0.0060	0.29
FL	1.45	0.0065	0.24
HC	1.54	0.0080	0.17

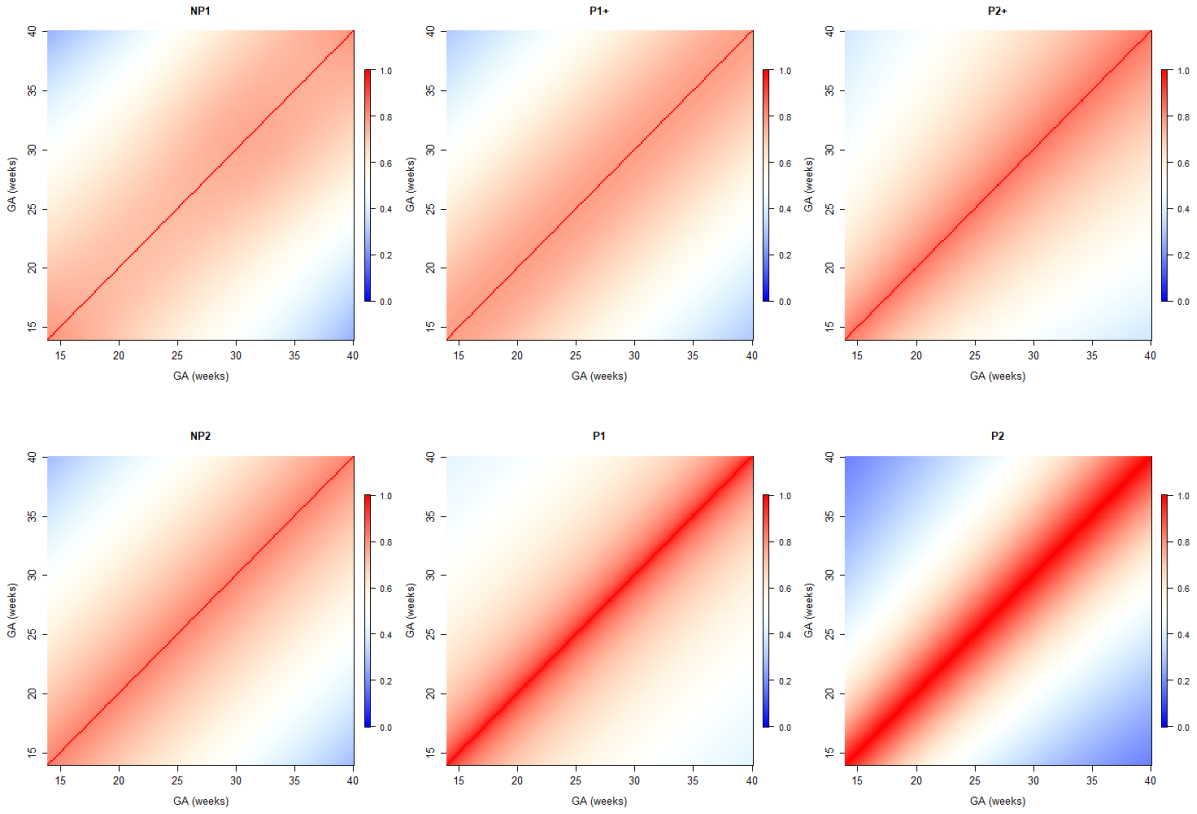

Figure 1.6 Temporal correlations of standardized AC with different correlation models.

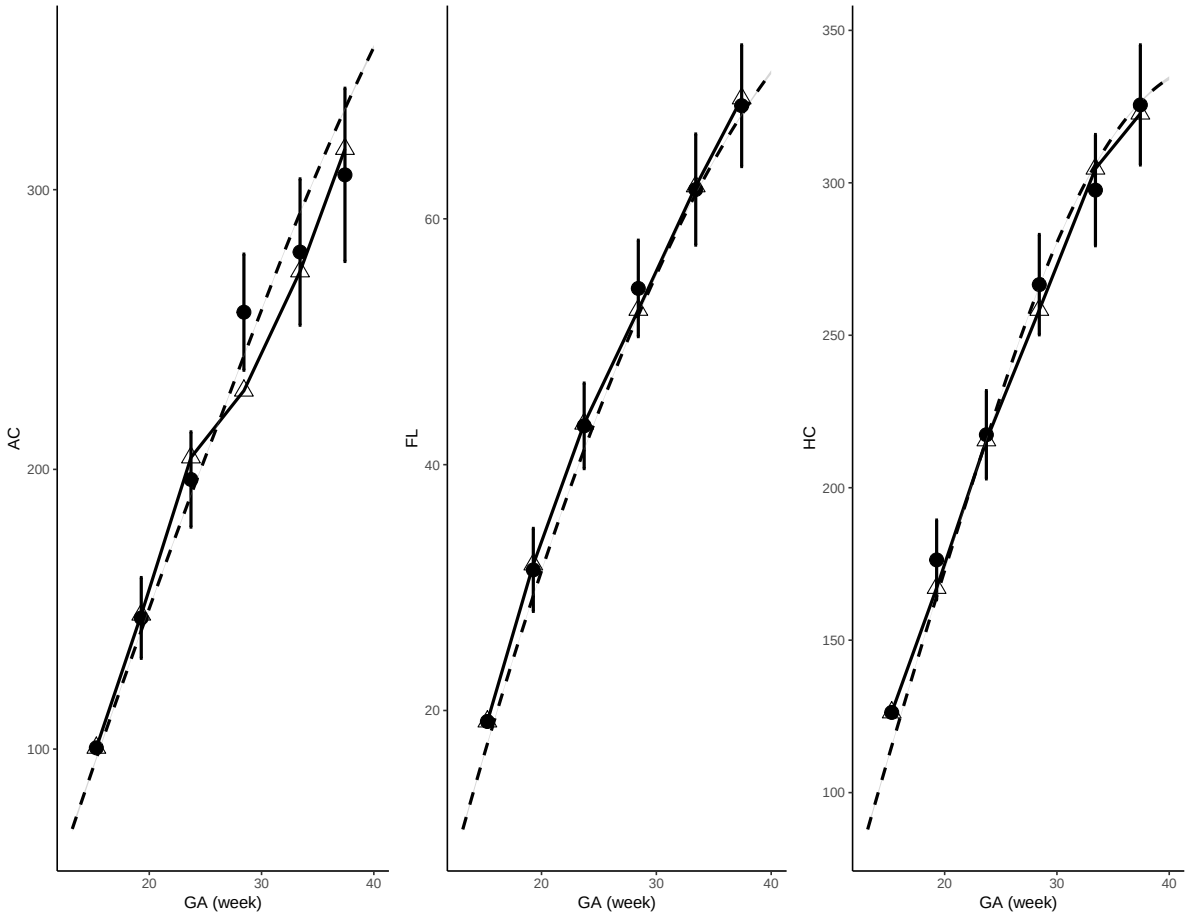

Figure 1.7 Observed growth trajectory (linked triangles) and predicted measurements (dots) given most recent observations of a randomly selected fetus. Dashed line is the population mean.

1.5 Case Study: Dynamic Growth Velocity

As an example, we use the result from parametric correlation models to study the growth velocity of a sample fetus chosen for the purpose of illustration, whose AC, FL and HC are consecutively measured for 6 times between week 15 and week 38. The observed growth trajectories are shown as linked triangles in Figure 1.7. Based on each observed measurement at T_j, we also dynamically predict the measurement T_{j+1} shown as dots, each with a 95% prediction interval. Specifically, each observed measurement at T_j is transformed to Z-score using (1.2). Then, a conditional Z-score $Z_{\langle T_{j+1}|T_j,\cdots,T_1\rangle}$ at time T_{j+1} is obtained assuming joint normality with the P1+ correlation model. This

conditional Z-score is then transformed back to the original measurement given marginal references. Clinicians might use this approach to compare the observed fetal growth measurements versus its expected measurements at a certain age to determine if a fetus is growing normally. They can also calculate and compare velocity increments. We will use the correlations studied in this paper for the subsequent clinical paper on conditional fetal velocity for use by clinicians.

It is observed that for this randomly sampled fetus, the growth of FL and HC are regular and can be accurately predicted. For GA, its measurements are higher (still normal) than predicted during the third visit, but much lower than expected during the fourth visits. This suggests that a closer monitoring might be needed. The following visits indicate that the GA of the sampled fetus becomes consistently below the population mean.

To facilitate the usage of the obtained results in practice, a Shiny application is built along with this paper, where functionalities such as visualization, calculating correlation, prediction and cSDS are integrated for all the three fetal growth measurements (https://lxiao5.shinyapps.io/shinycalculator/); see Figure 1.8.

In the meanwhile, correlation tables for fetal growth measurements are provided in appendix.

1.6 Discussion

We have modelled the correlation coefficients of fetal growth for HC, AC, and FL and provided formulae that can be used to obtain the correlation coefficient for each fetal measure between measurements made at any two time points between 14 and 40 weeks based on the FGLS data. The FGLS cohort is the largest prospective study to date to collect data on fetal ultrasound measurements among optimally healthy pregnant women, used many quality control measures, encompasses eight geographically diverse populations, was population-based, involved a cohort of women at low risk of intra-uterine growth restriction and preterm, remained healthy with adequate growth and motor development up to 2 years of age, hence making it an ideal dataset for characterising the expected correlation of fetal size measurements [Pap14; Vil18; Sar13; Vil15; Vil13].

We modelled empirical correlations of fetal measurements and showed that the fitted corre-

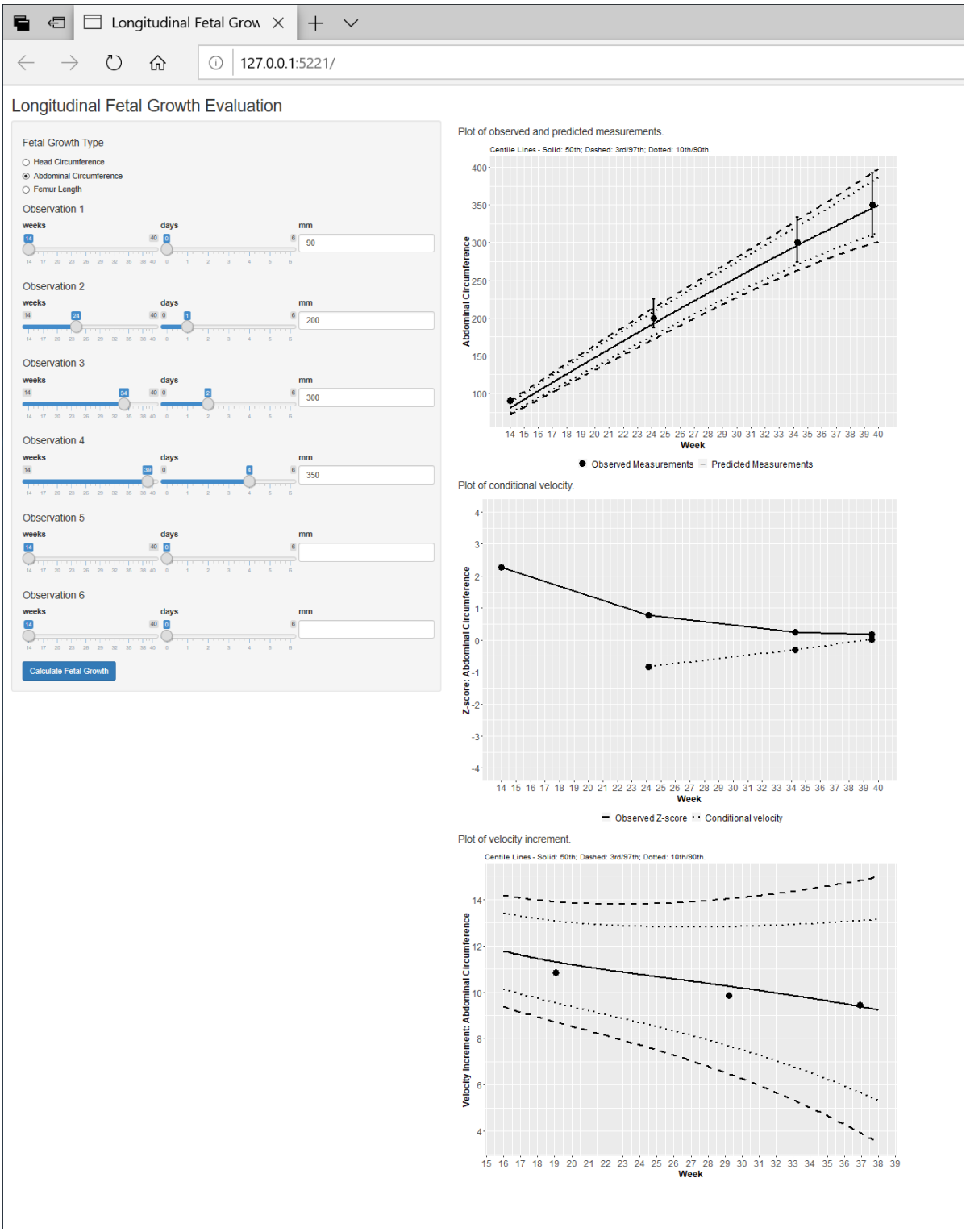

Figure 1.8 Longitudinal fetal growth calculator.

lations offered a good fit to the empirical correlation structure. We modelled the correlations in weekly intervals on the assumption that the correlations were reasonably stable within each week of gestation. The fitted correlations decreased with increasing time interval as expected. Although velocity charts could be an important complement to attained size charts [Pap14], they are not often used clinically. For example, a clinician may be interested to know whether fetal HC at 20 weeks is a good predictor of that same fetuses HC at 30 weeks. If we know the correlation between these two time points, then we are able to make inference about what fetal HC might be at 30 weeks given that we know their fetal HC at 20 weeks. Such predictions can be useful for identifying fetuses that are faltering in growth i.e., if their observed fetal HC at 30 weeks is significantly less than what was predicted. This would be indicative of poor growth within that time period that may be require further investigation or intervention. A limitation of the study was paucity of data and small sample sizes for some pairs of gestational ages especially in early gestation (first trimester) and at term (40 weeks).

In summary, we have for the first time provided formulae for obtaining correlation coefficients for fetal biometry using data that was prospectively collected, involved eight countries from diverse settings, collected using unified protocols, measurement procedures and standardisation, using high quality data based on the rigorous data quality process that was put in place during the study period. INTERGROWTH-21st Project is the largest prospective study of fetal growth involving multiple measurements per fetus that were purposely obtained for the study. These equations for obtaining corresponding correlation coefficients for any pair of data between 14 and 40 weeks and consequently the calculation of a velocity Z-score provide a potentially useful tool for clinicians who wish to monitor the fetal growth and development over time. To facilitate ease of use, a web application (Shiny application for now) that calculates the expected correlation between any two time points in the interval 14 to 40 weeks for HC, AC, and FL will be made freely available on the INTERGROWTH-21st website where other applications for fetal, preterm, and newborn size are already available (https://intergrowth21.tghn.org/).

1.7 Acknowledgments

This study was funded by the INTERGROWTH-21st grant 49038 from the Bill & Melinda Gates Foundation to the University of Oxford; we gratefully acknowledge their support. We would also like to thank the INTERGROWTH-21st Project team and participants who contributed data.

CHAPTER

2

SPARSE SINGLE INDEX MODELS FOR MULTIVARIATE RESPONSES

2.1 Introduction

Multivariate response linear regression models are commonly used to predict several quantities simultaneously given a common set of covariates. Consider data $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^q \times \mathbb{R}^p$, a coefficient matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$ and uncorrelated random errors $\{\boldsymbol{\epsilon}_i\}_{i=1}^n \in \mathbb{R}^q$ assuming $\mathbb{E}(\boldsymbol{\epsilon}_i) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2\mathbf{I}_q$, the model is given by

$$\mathbf{y}_i \;=\; \mathbf{B}^\mathsf{T}\mathbf{x}_i + \boldsymbol{\epsilon}_i. \tag{2.1}$$

The model (2.1) may be rewritten more compactly as $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, where $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)^\mathsf{T} \in \mathbb{R}^{n\times q}$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\mathsf{T} \in \mathbb{R}^{n\times p}$ and $\mathbf{E} = (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n)^\mathsf{T} \in \mathbb{R}^{n\times q}$. We assume that $\mathbf{Y}$ and $\mathbf{X}$ have been centered so that intercept is not present in the model.

While univariate regression models can be separately fitted to each response, joint regression models exploit correlations among responses and hence can be superior in prediction accuracy [BF97]. Indeed, various methods have been proposed in the literature, such as partial least-squares regression (PLR), canonical correlation regression (CCR), principal component regression (PCR) and reduced rank regression (RRR) [Hai98; VR13]. Regularization techniques have been adopted for multivariate response regression models when predictors or responses reside in high dimensional spaces. On the one hand, to deal with the challenge of having more predictors than samples, i.e. $p > n$, Obozinski et al. [Obo08; Obo10] utilized a block-structured penalty on the coefficient matrix $\mathbf{B}$ to select covariates that are predictive for all the responses. Peng et al. [Pen10] considered a combination of penalties that induces both overall sparsity and row sparsity of a matrix. Li et al. [Li15] proposed models to include arbitrary group structure for the regression coefficient matrix. On the other hand, when q is large, low-rank approaches [Yua07; Bun11; Che13; Ma14] are often adopted by thresholding the singular values of $\mathbf{B}$. These low-rank models assume the underlying regression structures have a shared representation for all the responses, and hence have substantially fewer unknown parameters to estimate, potentially leading to more efficient model estimation. More parsimonious models can be obtained by enforcing both a variable selection penalty and a low-rank penalty [Bun12; CH12] to handle high-dimensional models with both large p and large q. The problem of multi-task learning [Car98] in the machine learning community presents a similar problem, and we refer interested readers to Evgeniou & Pontil [EP04], Argyriou et al. [Arg07] and Argyriou et al. [Arg08b] for some representative work on the multi-task learning.

A key assumption of model (2.1) is the linear relationship between each response and covariates, which can be seriously violated in practice. Figure 2.1 illustrates the relationship among different responses and a common set of covariates in a genetic association study, where a joint linear regression model seems insufficient. Thus, nonparametric regression methods can be useful to accommodate

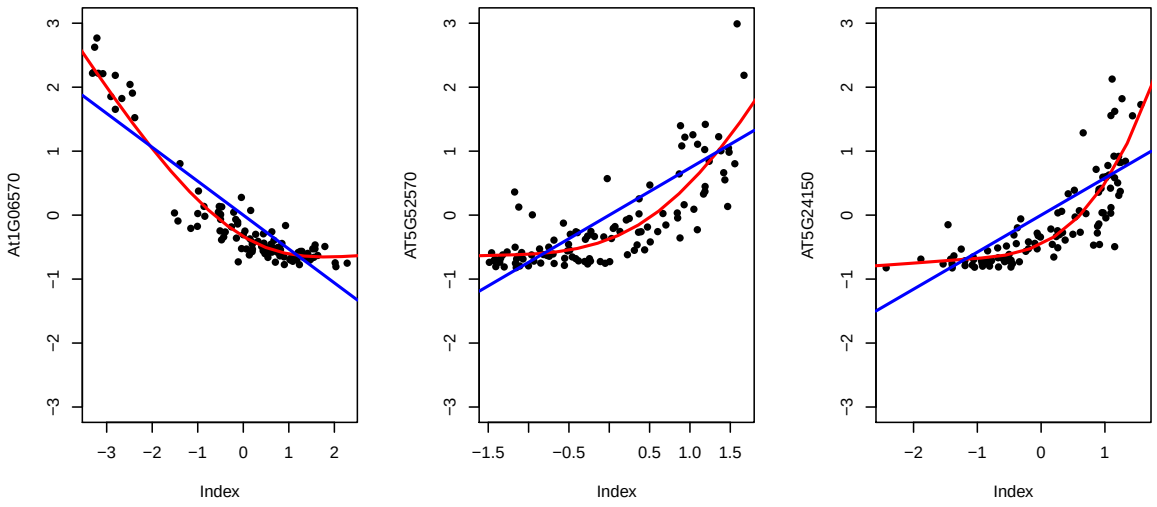

Figure 2.1 Linear and nonlinear regression models based on the same set of pre-selected covariates for various responses - red: single index model; blue: linear model. In order to compare index model with linear model in the same plot, the coefficient vector for each response is fixed for both models and only the unknown function of the index is considered.

potential nonlinear relations. Because of the curse of dimensionality, a fully nonparametric regression approach is often undesired. To address the dimension problem, Yuan et al. [Yua07] proposed additive models for each response, where each additive component is a nonparametric univariate function of a covariate. An alternative approach is the single index model (SIM), in which each response is modeled as a nonparametric function of an index that is a linear combination of the covariates. For univariate response ($q = 1$), SIM, as well as its various forms, such as partially linear SIM and generalized SIM, were extensively studied by Härdle & Stoker [HS89], Ichimura [Ich93], Carroll et al. [Car97], Hristache et al. [Hri01], Liang et al. [Lia10] and Wang et al. [Wan10] using kernel methods, and by Yu & Ruppert [YR02] and Wang & Yang [WY09] using splines. For spline-based methods, recent studies of Kong & Xia [KX07] and Foster et al. [Fos13] considered the problem of variable selection for SIM.

To tackle both potential nonlinearity in multivariate response regression and challenges in high dimensional data, we propose the multivariate response single index model (MSIM), where

the response vector $\mathbf{y}_i$ is linked to covariates $\mathbf{x}_i$ and coefficient matrix $\mathbf{B}$ through q unknown and unspecified link functions $\mathscr{F} = \{f_1, \ldots, f_q\}$ such that

$$\mathbf{y}_i \;=\; \left\{f_1(\mathbf{x}_i^\top \mathbf{B}_{\cdot 1}), \cdots, f_q(\mathbf{x}_i^\top \mathbf{B}_{\cdot q})\right\}^\top + \boldsymbol{\epsilon}_i. \tag{2.2}$$

In (2.2), $\mathbf{B}_{\cdot j} \in \mathbb{R}^p$ is the jth column of $\mathbf{B}$ and is the coefficient vector corresponding to the jth response y_{ij} in $\mathbf{y}_i = (y_{i1}, \cdots, y_{iq})^\top$ for all $i = 1, \cdots, n$. For model identifiability we assume that $\|\mathbf{B}_{\cdot j}\|_2^2 = 1$, and without loss of generality let b_{1j}, the first element of $\mathbf{B}_{\cdot j}$, be nonnegative. Such an assumption is commonly used in single index models [Ich93]. Moreover, we shall assume that $\mathbf{B}$ has sparse structures that induces information sharing among multivariate responses. For example, we assume that the non-zero rows of $\mathbf{B}$ are sparse, or the vector of singular values of $\mathbf{B}$ are sparse, or both are sparse. We will impose these sparsity structures via corresponding penalty functions, as shall be shown later in this paper. The recovery of the sparsity structures in $\mathbf{B}$ and the form of the link functions could thus provide insights on how the covariates are related to multivariate responses. However, to the best of our knowledge, the problem of MSIM with multiple penalties has not been addressed in the literature. Thus, the main contribution of this paper is a computational framework for jointly estimating the unknown smooth link functions $\mathscr{F}$ and coefficient matrix $\mathbf{B}$ under a general collection of structure-inducing assumptions. At a high level we employ a classic alternating strategy for estimating $\mathscr{F}$ conditioned on $\mathbf{B}$ and vice versa. The former problem can be solved using existing nonparametric methods. The latter, however, requires a new approach based on the alternating direction method of multipliers (ADMM).

The rest of this paper introduces our framework in the following stages. In Section 2.2, we describe the main model framework and discuss various sparsity-inducing penalties for MSIM. In Section 2.3, we detail a new alternating two-step iterative algorithm for regularized MSIM. In Section 2.4, we discuss how to choose tuning parameters for the model. In Section 2.5 and Section 2.6, we demonstrate the effectiveness of our model through simulation studies and an application in a genetic association study. In Section 3.5, we make conclusions and discuss future work.

2.2 Regularized MSIM

We first introduce notation. Bold upper-case letters (e.g. $\mathbf{B}$) denote matrices, bold lower-case letters or upper-case letters with dot notation (e.g. $\mathbf{b}$, $\mathbf{B}_{\cdot j}$ or $\mathbf{B}_{k\cdot}$) denote vectors and lower-case letters denote scalars (e.g. b). In particular, for any matrix $\mathbf{B}$, we use $\mathbf{B}_{\cdot j}$ and $\mathbf{B}_{k\cdot}$ to denote the jth column and kth row of $\mathbf{B}$, respectively. We use $\|\cdot\|_2$ to denote the Euclidean norm of a vector and $\|\cdot\|_{\mathrm{F}}$ to denote the Frobenius norm of a matrix.

To fit our proposed regularized MSIM (2.2), we consider the following optimization problem:

$$\underset{\mathscr{F},\mathbf{B}}{\text{minimize}} \left\{ \mathscr{L}_{\mathscr{F}}(\mathbf{B}) + \sum_{l=1}^{h} \mathscr{R}_{l,\lambda_l}(\mathbf{B}) \right\} \tag{2.3}$$

subject to identifiability constraints $\|\mathbf{B}_{\cdot j}\|_2^2 = 1$ and $b_{1j} \geq 0$ for $j = 1, \ldots, q$. In (2.3), $\mathscr{L}_{\mathscr{F}}$ is a loss function depending on the q link functions and $\mathscr{R}_{l,\lambda_l}$ is a regularizer each depending on a tuning parameter λ_l. Treating f_j as an elementwise function if its argument is a vector and noticing that $\left\|\mathbf{Y}_{\cdot j} - f_j(\mathbf{X}\mathbf{B}_{\cdot j})\right\|_2^2 = \sum_{i=1}^{n} \left\{ y_{ij} - f_j(\mathbf{x}_i^{\mathsf{T}}\mathbf{B}_{\cdot j}) \right\}^2$, we shall use least-squares loss

$$\mathscr{L}_{\mathscr{F}}(\mathbf{B}) \;=\; \frac{1}{2n} \sum_{j=1}^{q} \left\|\mathbf{Y}_{\cdot j} - f_j(\mathbf{X}\mathbf{B}_{\cdot j})\right\|_2^2.$$

The jth column of $\mathbf{Y}$, i.e. $\mathbf{Y}_{\cdot j} = (y_{1j}, \cdots, y_{nj})^{\mathsf{T}}$, corresponds to the jth response.

The regularizer $\mathscr{R}_{l,\lambda_l}$ enforces sparsity on $\mathbf{B}$. We use multiple sparsity-inducing penalties to recover a parsimonious and sparse coefficient matrix $\mathbf{B}$. Below we review some popular matrix penalties.

- Column penalty

  By treating each column of $\mathbf{B}$ as a block, penalizing column vectors in $\mathbf{B}$ yields q sparse SIMs under the least-squares loss. For a fixed index j, problem (2.3) reduces to minimizing $\mathscr{L}_{f_j}(\mathbf{B}_{\cdot j}) + \mathscr{R}_{\lambda}(\mathbf{B}_{\cdot j})$, where $\mathscr{L}_{f_j}(\mathbf{B}_{\cdot j}) = \frac{1}{2n}\|\mathbf{Y}_{\cdot j} - f_j(\mathbf{X}\mathbf{B}_{\cdot j})\|_2^2$. This is a generic model to study variable selection for SIM. Popular penalty terms such as the lasso [Tib96], elastic net [ZH05], adaptive

lasso [Zou06], SCAD [FL01], MC+ [Zha10] and so on have been studied under the assumption that f_j is a linear function and consequently some have been applied to the SIM by Carroll et al. [Car97], Liang et al. [Lia10], Peng & Huang [PH11] and Foster et al. [Fos13]. Our model includes sparse SIM as a special case and the algorithm proposed to solve (2.3) is directly applicable for this simpler model.

- Row penalty

  The row penalty is a group lasso type penalty that penalizes each row of $\mathbf{B}$ as a group. The implication of the row penalty is that it identifies covariates that are active for predicting all q responses simultaneously in a joint model. Take the $L_{2,1}$ norm regularization for example; it is equivalent to a group lasso penalty $\mathscr{R}_\lambda(\mathbf{B}) = \lambda \sum_{k=1}^{p} \|\mathbf{B}_{k\cdot}\|_2$ [YL06]. The tuning parameter λ controls the amount of shrinkage to each row. When $q = 1$, this penalty reduces to the standard lasso penalty for SIM.

- Rank penalty

  It is often useful to assume that the coefficient matrix $\mathbf{B}$ has low-rank. To induce the recovery of a low-rank $\mathbf{B}$, we may use $\mathscr{R}_\lambda(\mathbf{B}) = \lambda \text{rank}(\mathbf{B})$ or its tightest convex relaxation $\mathscr{R}_\lambda(\mathbf{B}) = \lambda \|\mathbf{B}\|_*$ that penalizes the nuclear norm of $\mathbf{B}$. Extensions of this line of research include, for example, Chen et al. [Che13] and Josse & Sardy [JS16]. An alternative way to achieve the recovery of a low-rank $\mathbf{B}$ is to directly fix the rank of the coefficient matrix; see, e.g. Bunea et al. [Bun12] and Chen & Huang [CH12].

In addition to these penalties, we can also penalize individual elements in $\mathbf{B}$, or any group of elements properly defined. In principle, the form of the penalty is based on the structure of the coefficient matrix that one wishes to recover. Our proposed model estimation is general and accommodates different types and combinations of penalties, but the computational challenge also increases as we increase the number of structure-inducing penalties. The next section introduces a flexible and general framework based on alternating minimization and the ADMM algorithm for handling an arbitrary collection of penalties. Proven to be useful for multivariate linear models, the

marriage of the row penalty and the rank penalty will be the extensively studied under MSIM in later sections.

2.3 Estimation

Simultaneously estimating a collection of q unknown smooth functions $\mathscr{F} = \{f_1, \ldots, f_q\}$ and the index coefficient matrix $\mathbf{B}$ is computationally challenging. Estimating the collection of functions $\mathscr{F}$ given $\mathbf{B}$, and vice versa estimating $\mathbf{B}$ given $\mathscr{F}$, in contrast is computationally more straightforward and consequently a commonly used strategy in fitting single index models [WY09; Fos13]. Thus, we adopt the following two-step iterative approach.

Step 1: Given the index coefficient matrix $\mathbf{B}$, we estimate each response $f_j \in \mathscr{F}$ via univariate smoothing; see Section 2.3.1.

Step 2: Given $\mathscr{F}$, we estimate $\mathbf{B}$ via the proposed optimization algorithm in Section 2.3.2.

To obtain an initial estimate of $\mathbf{B}$, we fit q univariate ridge regressions separately.

2.3.1 Step 1: Estimating $\mathscr{F}$

Given the index coefficient matrix $\mathbf{B}$, we fit each unknown and unspecified function f_j using penalized splines [EM96]. Let $u_{ij} = \mathbf{x}_i^\top \mathbf{B}_{\cdot j}$ be the index for the jth response, then $y_{ij} = f_j(u_{ij}) + \epsilon_{ij}$. The penalized splines approximate $f_j(t)$ by a linear combination of B-splines, i.e.

$$f_j(t) \;=\; \sum_{k=1}^{m} \theta_{kj} \phi_{kj}(t), \tag{2.4}$$

where $\{\phi_{1j}(t), \ldots, \phi_{mj}(t)\}$ is a set of m cubic B-spline basis functions constructed from equally-spaced knots defined in the range of u_{ij} and θ_{kj} are the associated coefficients to be estimated. To simplify notation, we use the same number of B-spline basis functions for all responses. Note that the knots depend on the range of the indices and hence are response specific. Moreover, they could vary in the proposed algorithm, since estimates of the index parameters and the nonlinear functions

are iteratively updated. In practice we use 10 basis functions to reduce the approximation bias and enforce a smoothness penalty on the estimate to control over-fitting. Specifically, we estimate the coefficients $\Theta_{\cdot j} = (\theta_{1j}, \ldots, \theta_{mj})^{\mathsf{T}}$ by minimizing

$$\mathscr{Q}_{\gamma_j}(\Theta_{\cdot j}) \;=\; \frac{1}{2n}\sum_{i=1}^{n}\left\{ y_{ij} - \sum_{k=1}^{m}\theta_{kj}\phi_k(u_{ij})\right\}^2 + \gamma_j \Theta_{\cdot j}^{\mathsf{T}}\mathbf{D}\Theta_{\cdot j}, \tag{2.5}$$

where $\mathbf{D}$ is a second-order difference penalty matrix [EM96] and γ_j is a smoothing parameter that balances model fit and model complexity. The optimization problem is a least-squares problem and to quickly and automatically select γ_j, we use generalized cross-validation (GCV) [Rup02] for each response.

2.3.2 Step 2: Estimating B

Even with $\mathscr{F}$ fixed, estimating $\mathbf{B}$ is not straightforward when multiple non-smooth sparsity inducing penalties are involved. To address this challenge, we introduce a collection of dummy variables $\mathscr{C} = \{\mathbf{C}_0, \mathbf{C}_1, \cdots, \mathbf{C}_h\}$, one for each of the $h+1$ penalty terms. Given $\mathscr{F}$, the following equality constrained problem is equivalent to problem (2.3) after variable splitting:

$$\begin{aligned} &\text{minimize } \mathscr{L}_{\mathscr{F}}(\mathbf{B}) + \sum_{l=1}^{h}\mathscr{R}_{l,\lambda_l}(\mathbf{C}_l) + \sum_{j=1}^{q}\iota_S\left(\mathbf{C}_{0,\cdot j}\right) \\ &\text{subject to } \mathbf{B} = \mathbf{C}_0 = \mathbf{C}_1 = \cdots = \mathbf{C}_h, \end{aligned} \tag{2.6}$$

where ι_S is the indicator function of the identifiability constraint set S and is given by

$$\iota_S(\mathbf{c}) \;=\; \begin{cases} 0 & \text{if } \|\mathbf{c}\|_2 = 1 \text{ and } c_1 \geq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

With the help of dummy variables $\mathscr{C}$, the optimization problem in (2.6) can be iteratively solved using the classic alternating direction method of multipliers (ADMM) algorithm [GM76; GLT89]. In recent years, ADMM has enjoyed widespread use in constrained optimization and sparse regression

problems [Boy11]. The algorithm for (2.6) admits a simple iterative block updates for the matrices $\mathbf{B}$ and $\mathbf{C}_l$'s that involve minimizing the augmented Lagrangian function of (2.6). Let $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$ denote the inner product between the matrices $\mathbf{A}$ and $\mathbf{B}$, then the augmented Lagrangian function of (2.6) is

$$\mathscr{L}_\rho(\mathbf{B}, \mathscr{C}, \mathscr{M}) \;=\; \mathscr{L}_{\mathscr{F}}(\mathbf{B}) + \sum_{l=1}^{h} \mathscr{R}_{l,\lambda_l}(\mathbf{C}_l) + \sum_{j=1}^{q} \iota_S\left(\mathbf{C}_{0,\cdot j}\right) + \mathscr{Q}(\mathbf{B}, \mathscr{C}, \mathscr{M})$$

where

$$\mathscr{Q}(\mathbf{B}, \mathscr{C}, \mathscr{M}) \;=\; \sum_{l=0}^{h} \mathscr{Q}_l(\mathbf{B}, \mathbf{C}_l, \mathbf{M}_l) \;=\; \sum_{l=0}^{h} \left(\langle \mathbf{M}_l, \mathbf{C}_l - \mathbf{B} \rangle + \frac{\rho}{2} \|\mathbf{C}_l - \mathbf{B}\|_{\mathrm{F}}^2 \right),$$

$\mathscr{M} = \{\mathbf{M}_0, \cdots, \mathbf{M}_h\}$ are Lagrange multiplier matrices, and ρ is a positive tuning parameter. We will discuss the choice of ρ in next section.

At the kth iteration of the algorithm, minimization of the augmented Lagrangian is illustrated by the following steps:

- Update 1: $\mathbf{B}^{(k+1)} = \underset{\mathbf{B}}{\arg\min}\, \mathscr{L}_{\mathscr{F}}(\mathbf{B}) + \mathscr{Q}\left(\mathbf{B}, \mathscr{C}^{(k)}, \mathscr{M}^{(k)}\right)$
- Update 2: $\mathbf{C}_l^{(k+1)} = \begin{cases} \underset{\mathbf{C}_0}{\arg\min}\, \mathscr{Q}_0\left(\mathbf{B}^{(k+1)}, \mathbf{C}_0, \mathbf{M}_0^{(k)}\right) + \sum_{j=1}^{q} \iota_S\left(\mathbf{C}_{0,\cdot j}\right) & \text{if } l = 0 \\ \underset{\mathbf{C}_l}{\arg\min}\, \mathscr{Q}_l\left(\mathbf{B}^{(k+1)}, \mathbf{C}_l, \mathbf{M}_l^{(k)}\right) + \mathscr{R}_{l,\lambda_l}(\mathbf{C}_l) & \text{if } l > 0 \end{cases}$
- Update 3: $\mathbf{M}_l^{(k+1)} = \mathbf{M}_l^{(k)} + \rho\left(\mathbf{C}_l^{(k+1)} - \mathbf{B}^{(k+1)}\right)$

Update 1: B

Due to the sum-of-squares structure of the Frobenius norm, define $g_j^{(k)}(\mathbf{B}_{\cdot j}) = \frac{1}{2n}\|\mathbf{Y}_{\cdot j} - f_j(\mathbf{X}\mathbf{B}_{\cdot j})\|_2^2 + \sum_{l=0}^{h} \left\{ \left\langle \mathbf{M}_{l,\cdot j}^{(k)}, \mathbf{C}_{l,\cdot j}^{(k)} - \mathbf{B}_{\cdot j} \right\rangle + \frac{\rho}{2} \left\| \mathbf{C}_{l,\cdot j}^{(k)} - \mathbf{B}_{\cdot j} \right\|_2^2 \right\}$, then $\mathbf{B}$ can be updated column by column through

$$\mathbf{B}_{\cdot j}^{(k+1)} \;=\; \underset{\mathbf{B}_{\cdot j}}{\arg\min}\, g_j^{(k)}(\mathbf{B}_{\cdot j}). \tag{2.7}$$

However, there is no explicit solution to optimize a general g_j, thus we propose to inexactly minimize (2.7) using a first-order or second-order method. With the spline representation (2.4), f_j' and f_j'' can be estimated together to formulate the gradient ∇g_j and the Hessian matrix $\nabla^2 g_j$ (see the supplemental materials). Thus, for an iterative optimization algorithm nested in the $(k+1)$th iteration of the ADMM algorithm, we set $\mathbf{B}_{\cdot j}^{(k+1,1)} = \mathbf{B}_{\cdot j}^{(k)}$ as the initial value and iterate by the formula

$$\mathbf{B}_{\cdot j}^{(k+1,m+1)} \;=\; \mathbf{B}_{\cdot j}^{(k+1,m)} - t_m \nabla g_j^{(k)}\left(\mathbf{B}_{\cdot j}^{(k+1,m)}\right), \tag{2.8}$$

where t_m is a step size that can potentially depend on the current inner iteration index m. Symbolically, the desired final result is $\mathbf{B}_{\cdot j}^{(k+1)} = \lim_{m\to\infty} \mathbf{B}_{\cdot j}^{(k+1,m)}$, but in practice we stop when the relative change of two consecutive iterations falls within a pre-specified threshold. To achieve a balance between the convenience and speed of the inner-loop and the convergence rate, we recommend using a quasi-Newton type algorithm (e.g. BFGS), where an approximate Hessian matrix $\widehat{\nabla^2 g_j}$ is updated through a recursion requiring only the information of first-order derivative ∇g_j and f_j'. Then we replace t_m by $\widehat{\nabla^2 g_j}^{-1}$ in (2.8).

Update 2: $\mathbf{C}_0$

Similar to the column-wise update to $\mathbf{B}$, after some simple manipulations, we can see the subproblem in this step is to minimize $\iota_S\left(\mathbf{C}_{0,\cdot j}\right) + \frac{\rho}{2}\left\|\mathbf{C}_{0,\cdot j} - \mathbf{B}_{\cdot j}^{(k+1)} + \rho^{-1}\mathbf{M}_{0,\cdot j}^{(k)}\right\|_2^2$. The solution is

$$\mathbf{C}_{0,\cdot j}^{(k+1)} \;=\; \mathscr{P}_S\left(\mathbf{B}_{\cdot j}^{(k+1)} - \rho^{-1}\mathbf{M}_{0,\cdot j}^{(k)}\right), \tag{2.9}$$

where $\mathscr{P}_S(\cdot)$ is the Euclidean projection of a vector onto the surface of unit-ball with the first element being nonnegative.

Update 2: $\mathbf{C}_l$

Without loss of generality, we ignore the subscript l and consider one penalty at a time. Then the subproblem is to minimize $\mathscr{R}_\lambda(\mathbf{C}) + \mathscr{Q}\left(\mathbf{B}^{(k+1)}, \mathbf{C}, \mathbf{M}^{(k)}\right)$. Recall that $\mathscr{Q}$ is a quadratic function of $\mathbf{C}$, we

Table 2.1 Common matrix penalties and proximal maps.

Type	$\mathscr{R}_\lambda$	Solution	Comment
$L_{2,1}$	$\lambda\sum_{r=1}^{p}\|\mathbf{C}_{r\cdot}\|_2$	$\mathbf{C}_{r\cdot}=\left[1-\lambda/\rho\|\mathbf{Z}_{r\cdot}\|_2\right]_+\mathbf{Z}_{r\cdot}$	row-wise update
rank	$\lambda\text{rank}(\mathbf{C})$	$\mathbf{C}=\mathbf{U}\text{Hard}(\mathbf{D},\lambda/\rho)\mathbf{V}^\mathsf{T}$	hard-thresholding
nuclear norm	$\lambda\|\mathbf{C}\|_*$	$\mathbf{C}=\mathbf{U}\text{Soft}(\mathbf{D},\lambda/\rho)\mathbf{V}^\mathsf{T}$	soft-thresholding

then have

$$\mathbf{C}^{(k+1)} \;=\; \arg\min_{\mathbf{C}} \mathscr{R}_\lambda(\mathbf{C})+\frac{\rho}{2}\|\mathbf{C}-\mathbf{B}^{(k+1)}+\rho^{-1}\mathbf{M}^{(k)}\|_{\text{F}}^2. \tag{2.10}$$

The subproblem of (2.10) is known as the proximal map and for many useful regularization penalties, there exists explicit solutions [CW05; PB14; Pol15]. We illustrate this result by first defining a new response matrix $\mathbf{Z}^{(k)}=\mathbf{B}^{(k+1)}-\rho^{-1}\mathbf{M}^{(k)}$. Under $L_{2,1}$ penalty, for instance, the explicit solution is to update $\mathbf{C}$ row by row through

$$\mathbf{C}_{r\cdot}^{(k+1)} \;=\; \left[1-\lambda/\rho\|\mathbf{Z}_{r\cdot}^{(k)}\|_2\right]_+\mathbf{Z}_{r\cdot}^{(k)},$$

where $r=1,\cdots,p$ and $[u]_+=\max\{0,u\}$ denotes the positive part of the scalar u. Moreover, Table 2.1 summarizes some common matrix penalties mentioned in Section 2.2 and their corresponding proximal maps. Note that after singular value decomposition of the new response matrix $\text{SVD}(\mathbf{Z})=\mathbf{UDV}^\mathsf{T}$, $\text{Hard}(\mathbf{D},\lambda/\rho)$ is the element-wise hard thresholding of the diagonal matrix $\mathbf{D}$ with tuning parameter λ/ρ and similarly $\text{Soft}(\mathbf{D},\lambda/\rho)$. Also note that with penalties in Table 2.1, the updates are all non-iterative.

Another useful type of rank penalization is to restrict the minimizer to have exactly rank r using the indicator function

$$\iota_r(\mathbf{C}) \;=\; \begin{cases} 0 & \text{if rank}(\mathbf{C})=r, \\ +\infty & \text{otherwise} \end{cases}$$

as the penalty function. This is a fixed rank approach, which is similar to but computationally cheaper than using a tuning paramter λ to penalize the rank because different λ's may result in a same matrix having a particular rank. For the simplicity of narration, we treat r and λ equally in tuning parameter searching.

It is obvious that by using this block updating approach, we can conveniently recover a coefficient matrix that has various structures at the same time by switching among different penalties.

2.3.3 Marginal Screening and Identifiability

When applying the $L_{2,1}$ norm penalty on the coefficient matrix $\mathbf{B}$ to conduct variable selection, it is possible that a large tuning parameter λ shrinks all rows in $\mathbf{B}$ to zero. This problem also exists for SIM, and will apparently violate the identifiability constraints. A possible solution would be to leave a certain row of $\mathbf{B}$, denoted by $\mathbf{B}_{K\cdot}$, unpenalized and keep the sign to be positive for each element in this row. In this case, there is always one active covariate for all responses. The choice of the Kth row, however, is not trivial. Motivated by the nonparametric variable screening technique [Fan11], we propose finding the particular covariate having a combined largest effect for all responses. Consider univariate simple nonlinear regression for the the jth response and kth covariate $y_{ij} = \mu_{kj}(x_{ik}) + \epsilon_{ikj}$, where ϵ_{ikj} are i.i.d. errors. Thus, we can choose K by

$$K = \arg\min_{k} \sum_{j=1}^{q} \sum_{i=1}^{n} \{y_{ij} - \hat{\mu}_{kj}(x_{ik})\}^2.$$

$\hat{\mu}_{kj}$ can be obtained by any standard univariate smoothing technique. In our application, we apply the penalized spline smoother with tuning parameter selected by GCV.

2.3.4 Algorithm, Computation Complexity and Convergence

The above discussion leads to Algorithm 1 for regularized MSIM. Note that in practice we will terminate the algorithm after finitely many iterations. At termination each dummy variable will posses one structure (e.g. sparse, low-rank, or row-sparse), and the dummy variables while close to each

other will not be exactly identical. Thus, further adjustments are needed to form the final coefficient matrix $\widehat{\mathbf{B}}$ that simultaneously possesses multiple structures imposed by the set of structure-inducing penalties. Also note that for univariate model, the penalized SIM objective function can be optimized with the same procedure by simply treating vectors as matrices. Specifically, if the column penalty is the lasso penalty, then we can use a similar soft-thresholding rule as the $L_{2,1}$ penalty for MSIM.

Algorithm 1 Regularized MSIM

INPUT: $\mathbf{Y}, \mathbf{X}, \boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_h), \mathbf{B}^{(0)}, \mathbf{M}^{(0)}, \rho$
OUTPUT: $\mathbf{B}$

```
k ← 0
repeat
    for j = 1, ··· , q do                                   ▷ estimate 𝒻
        Update Θk + 1_j by (2.5)
    end for
    for j = 1, ··· , q do                                   ▷ inner-loop for B
        repeat
            Update B_j^(k+1) by (2.8)
        until convergence
    end for
    for j = 1, ··· , q do                                   ▷ identifiability
        Update C_{0,·j}^(k+1) by (2.9)
    end for
    for l = 1, ··· , h do                                   ▷ matrix level penalty
        Update C_l^(k+1) with λ_l by Table 2.1
        M_l^(k+1) ← M_l^(k+1) + ρ (C_l^(k+1) − B^(k+1))
    end for
    k ← k + 1
until convergence
return 𝒞
```

For one iteration of the two-step algorithm, the computational cost for univariate smoothing is $O(n)$ flops [HDH85] given cubic spline basis for each response, thus fitting q multivariate responses requires $O(qn)$ flops. Updating $\mathbf{B}_{\cdot j}$ is iterative, and for the inner-loop, the cost is $O(np + p^2)$ per step. Updating $\mathbf{C}_0$ and each $\mathbf{M}_l$ takes $O(pq)$ flops. The cost of updating $\mathbf{C}_l$ depends on the specific form of the lth penalty function. The low-rank inducing penalties require computing an SVD which

requires $O(pq^2)$ flops, while the row sparsity inducing penalties require $O(pq)$ flops. Thus, under the high-dimensional setting where $p > \max(q,n)$, the computational complexity is a factor of $O(p^2q)$, and the factor is also influenced by the number of iterations of the inner-loop.

Recently, Wang et al. [Wan15] studied the convergence of ADMM under a general setting, where the objective function can contain components of non-convex functions, compact manifolds and common sparsity inducing penalties. We apply their convergence analysis to the task of solving the equality constrained problem (2.6) using Algorithm 1.

Proposition 2.3.1 *If $\mathscr{L}_{\mathscr{F}}$ is Lipschitz differentiable, then for any sufficiently large ρ, the sequence $\left(\mathbf{B}^{(k)}, \mathscr{C}^{(k)}, \mathscr{M}^{(k)}\right)$ in the proposed ADMM algorithm has at least one limit point, and each limit point is a stationary point of $\mathscr{L}_\rho$.*

We note that Proposition 2.3.1 holds because 1) the row penalty $\mathscr{R}_{l,\lambda_l}$ and indicator function ι_S are prox-regular functions and 2) assuming that $\|\mathbf{x}_i\|_2$ is bounded, we have an bounded index $\mathbf{x}_i^\mathsf{T}\mathbf{B}_{\cdot j}$ and $\|\mathbf{Y}_{\cdot j} - f_j(\mathbf{X}\mathbf{B}_{\cdot j})\|_2^2$ is Lipschitz differentiable over the identifiability constraint set S, given the cubic spline construction of f_j. A closed form expression for the Lipschitz constant, however, is hard, if not impossible, to find, thus we adopt an empirical based approach introduced by Boyd et al. [Boy11] to set to set $\rho = 1$ in practice.

2.4 Tuning Parameter Selection

We now discuss how to select the tuning parameters for our proposed two-step algorithm. For the univariate smoothing step, we use GCV to choose the tuning parameters $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_q)$ for the penalized spline smoothers, which is built in the iterative algorithm. For the coefficient matrix estimation step, to choose the tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_h)$ we use the Bayesian information criterion (BIC) defined as

$$\text{BIC}(\boldsymbol{\lambda}) \;=\; \log\left\{\frac{1}{nq}\sum_{j=1}^{q}\|\mathbf{Y}_{\cdot j} - f_j(\mathbf{X}\widehat{\mathbf{B}}_{\cdot j})\|_2^2\right\} + \frac{\log(nq)}{nq}\widehat{\text{df}}_{\boldsymbol{\lambda}},$$

where $\widehat{\mathrm{df}}_{\boldsymbol{\lambda}}$ is the estimated degrees of freedom of the estimation procedure using coefficient matrix $\widehat{\mathbf{B}}$ that minimizes (2.3) for a given $\boldsymbol{\lambda}$. Note that for univariate SIM, the degrees of freedom can be estimated by the number of non-zero index coefficients [Lia10]. Hence, for MSIM, we will assume that the degrees of freedom is the same as that in a multivariate linear regression model with the same penalty and in the following we shall derive the degrees of freedom assuming the latter model. We first consider the low-rank multivariate response model. Under the low-rank assumption, the coefficient matrix $\mathbf{B}$ can be decomposed as $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}}$, where $\mathbf{U} \in \mathbb{R}^{p\times q}$, $\mathbf{D} \in \mathbb{R}^{q\times q}$ and $\mathbf{V} \in \mathbb{R}^{q\times q}$. The degrees of freedom can then be estimated by the number of free parameters $(r_{\mathbf{X}} + q - r)r$, where $r_{\mathbf{X}}$ is the rank of the design matrix and r is the number of non-zero elements in the diagonal matrix $\mathbf{D}$ [Bun11; Muk15]. To further take into account row sparsity and notice that the non-zero rows in $\mathbf{B}$ is the same as the non-zero rows in $\mathbf{U}$, we can adjust $r_{\mathbf{X}}$ by $r_0 = \min(r_{\mathbf{X}}, p_0)$ where p_0 is the number of non-zero rows in $\mathbf{U}$. Finally, with q identifiability constrains, the proposed degrees of freedom is $(r_0 + q - r)r - q$ for MSIM with low-rank and row-sparse penalties.

2.5 Simulation Studies

In this simulation study, we evaluate the performance of our proposed algorithm on both variable selection, rank determination and index estimation. The sample size is fixed at $n = 200$ in the simulation. To construct multivariate responses, we consider 6 different types of link functions and duplicate them 3 times to create $q = 18$ responses with the data generating model

$$\mathbf{y}_i \quad = \quad \left(\mathbf{y}_{i,1}^{\mathsf{T}}, \mathbf{y}_{i,2}^{\mathsf{T}}, \mathbf{y}_{i,3}^{\mathsf{T}}\right)^{\mathsf{T}} + \boldsymbol{\epsilon}_i,$$

where $\mathbf{y}_{i,h} \in \mathbb{R}^6$, a coefficient matrix $\mathbf{B} \in \mathbb{R}^{p\times 18}$ and

$$\mathbf{y}_{i,h} = \left\{\mathbf{x}_i^{\mathsf{T}}\mathbf{B}_{\cdot 6h-5}, (\mathbf{x}_i^{\mathsf{T}}\mathbf{B}_{\cdot 6h-4})^2, (\mathbf{x}_i^{\mathsf{T}}\mathbf{B}_{\cdot 6h-3})^3, \sin(\mathbf{x}_i^{\mathsf{T}}\mathbf{B}_{\cdot 6h-2}), \arctan(\mathbf{x}_i^{\mathsf{T}}\mathbf{B}_{\cdot 6h-1}), \exp(\mathbf{x}_i^{\mathsf{T}}\mathbf{B}_{\cdot 6h})\right\}^{\mathsf{T}}$$

for $h = 1, 2, 3$. The number of covariates p varies from 100 to 400 to accommodate different situations. Among all p covariates, the number of signal is set to $s = 0.06p$, i.e. only the first $0.06p$ rows of the coefficient matrix $\mathbf{B} \in \mathbb{R}^{0.06p \times 18}$ are non-zero. We construct a row-sparse and low-rank index coefficient matrix $\mathbf{B}$ using a block structure to reflect weak, moderate and strong signals such that

$$\mathbf{B} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{W}_2 & \mathbf{W}_3 \\ \mathbf{W}_2 & \mathbf{W}_3 & \mathbf{W}_1 \\ \mathbf{W}_3 & \mathbf{W}_2 & \mathbf{W}_1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where $\mathbf{W}_h \in \mathbb{R}^{0.02p \times 6}$ is a constant matrix of value h. The coefficient matrix $\mathbf{B}$ is further normalized column-wise to satisfy identifiability. In this case, the true rank of $\mathbf{B}$ is 3. The covariates $\mathbf{x}_i$ are generated from $N_p(\mathbf{0}, \Sigma_x)$ and errors $\boldsymbol{\epsilon}_i$ from independent normal distributions with different variances. In the simulation Σ_x is auto-regressive with correlation 0 and 0.5. The magnitude of variation in the error term depends on the responses to take into account the effect of different functional forms. We adjust $(\sigma_1^2, \cdots, \sigma_{18}^2)$ to maintain a fixed signal to noise ratio (SNR) defined as $\text{SNR}_j = \text{Var}\left\{f_j(\mathbf{x}^\top \mathbf{B}_{\cdot j})\right\}/\sigma_j^2$ for all $j = 1, \cdots, 18$. We set different SNR levels to test the performance under different settings. The BIC criterion with proposed degrees of freedom is used for tuning parameter selection. A total number of 100 simulation is conducted for each setting.

For variable selection, the $L_{2,1}$ penalty is used and true positive rate (TPR) as well as false positive rate (FPR) are calculated for each model. Denote $\mathbf{1}$ as the indicator function, we define $\text{TPR}(\hat{\mathbf{B}}, \mathbf{B}) = 100\sum_{k,j} \mathbf{1}\{\hat{b}_{kj} \neq 0\}\mathbf{1}\{b_{kj} \neq 0\}/\sum_{k,j} \mathbf{1}\{b_{kj} \neq 0\}$ and $\text{FPR}(\hat{\mathbf{B}}, \mathbf{B}) = 100\sum_{k,j} \mathbf{1}\{\hat{b}_{kj} \neq 0\}\mathbf{1}\{b_{kj} = 0\}/\sum_{k,j} \mathbf{1}\{b_{kj} = 0\}$, respectively. For rank selection, we use the fixed rank approach to expedite calculation and select the rank that leads to smallest BIC. For the accuracy of index coefficient estimation, we consider mean squared error $\text{MSE}(\hat{\mathbf{B}}, \mathbf{B}) = 100\|\hat{\mathbf{B}} - \mathbf{B}\|_{\text{F}}^2$ under MSIM. The results are summarized in Table 2.2 and Table 2.3 with the mean (standard error) for each metric. Note that 'N' stands for nonlinear model, 'S' stands for row-sparse model, and 'R' stands for low-rank model.

The simulation study demonstrates the effectiveness of our proposed method in selection and

Table 2.2 True positive rate, false positive rate and rank selection result.

(p, SNR)	N+S+R, cor=0			N+S+R, cor=0.5		
(100, 1)	100(0)	0.24(0.83)	3.00(0.35)	100(0)	0.07(0.31)	2.27(0.44)
(100, 0.5)	100(0)	0.06(0.33)	2.42(0.62)	100(0)	0.59(3.96)	2.15(0.36)
(100, 0.25)	99.83(1.67)	2.47(14.18)	2.94(1.00)	98.67(7.37)	0.82(2.59)	3.04(0.62)
(400, 8)	100(0)	0.43(1.07)	3.58(1.02)	99.96(0.42)	0.58(1.10)	3.10(0.36)
(400, 4)	100(0)	1.81(4.41)	3.71(1.08)	99.96(0.42)	0.73(1.23)	3.15(0.41)
(400, 2)	99.92(0.59)	4.67(5.43)	3.19(0.90)	98.58(9.35)	4.17(4.96)	3.01(0.64)
	TPR (%)	FPR (%)	rank	TPR (%)	FPR (%)	rank

Table 2.3 MSE for signal part.

(p, SNR)	N+S+R, cor=0	N+S+R, cor=0.5
(100, 1)	0.66(0.43)	1.17(0.23)
(100, 0.5)	1.18(0.41)	1.56(0.28)
(100, 0.25)	1.87(0.60)	2.31(0.60)
(400, 8)	0.13(0.11)	0.23(0.08)
(400, 4)	0.21(0.14)	0.31(0.11)
(400, 2)	0.42(0.13)	0.41(0.16)

estimation under both $p < n$ and $p > n$. Note that 1) due to the low FPR, the MSE is only calculated for the signal part and 2) after normalization for identifiability, the effect size (1, 2, 3) in coefficient matrix is actually (0.19, 0.38, 0.57) for $p = 100$ and (0.09, 0.19, 0.28) for $p = 400$ to be compared against the standard errors in Table 2.3. The performance of selection of variables, ranks and estimation is better when SNR is higher, and is generally good when correlation is present in the design matrix.

2.6 Application to a Genetic Association Study

We focus on a genetic association study of the regulatory control mechanisms in the gene network for isoprenoid in Arabidopsis Thaliana [Wil04]. Experiments have verified the existence of connections between some downstream pathways and two isoprenoid biosynthesis pathways. Thus in our analysis, the expression levels of $q = 62$ genes from four downstream pathways are formalized as multivariate responses, the expression levels of $p = 39$ genes from the two isoprenoid biosynthesis

pathways serve as predictors and a total of $n = 118$ GeneChip microarray experiments were performed. Due to the relatively large number of correlated responses and potential outliers in the data, She & Chen [SC17] proposed a robust reduced rank regression model to analyze this data. We would similarly use a reduced rank model, but on the other hand assume that the potential outliers can be incorporated by a more flexble nonlinear regression model and study this dataset through regularized MSIM. To satisfy the homogeneous variance assumption for multivariate responses, we center and scale both the response matrix and the design matrix.

Table 2.4 Gene pathway data 10-fold cross-validation result, $(n, p, q) = (118, 39, 62)$.

Model	R	S	S+R	N+S+R
MSPE	0.58(0.15)	0.46(0.13)	0.57(0.14)	0.38(0.16)
selected (%)	-	80.51(7.57)	85.90(3.68)	100(0)

Table 2.5 Contaminated gene pathway data (simulated) result, $(n, p, q) = (118, 200, 62)$.

	R	S	S+R	N+S+R
TPR (%)	-	58.79(5.91)	75.69(5.53)	69.82(14.07)
FPR (%)	-	0.04(0.16)	71.99(2.68)	7.59(7.36)
rank	1(0)	-	2(0)	4.03(0.73)

First, we compare various multivariate response models on a 10-fold split of the data and summarize the result in Table 2.4. The MSPE and the percentage of variables selected suggests that there are potential nonlinearity in the data and all the 39 predictors might be important. Indeed, when the row-sparse multivariate models (either linear or nonlinear) are fit to the whole data, no covariate is excluded. Based on these clues, we fit a 'N+S+R' model on the whole data and obtained estimated $\hat{\mathscr{F}} = \{\hat{f}_1, \ldots, \hat{f}_{62}\}$, index coefficient matrix $\hat{\mathbf{B}}$ of rank 5 and variance $\hat{\sigma}^2 = 0.30$. The 62 estimated mean functions are plotted in the supplemental materials, where nonlinearity is observed in many of the responses; see Figure 2.2 for a few examples. The singular values of the

estimated coefficient matrix are {4.83, 3.90, 3.52, 2.87, 1.68}. The estimated rank 5 is consistent with the suggestion of optimal rank used by the robust reduced rank model in [SC17].

Second, we design a simulation study to further evaluate the performance of variable selection of the proposed MSIM. Conditioning on the design matrix, we generate new multivariate responses under the fitted MSIM model $\{\hat{\mathscr{F}}, \hat{\mathbf{B}}, \hat{\sigma}^2\}$. We then contaminate the data by adding 161 independent noise genes from standard normal distribution so that now $p = 200$. Similar to the simulation study, the performance is evaluated by variable selection (TPR and FPR) and rank determination assuming the original 39 covariates are true signals. We run 100 simulations and compare various models; Table 2.5 summarizes the results. Note that the reduced rank model cannot select rows in the coefficient matrix and row-sparse model cannot select low-rank coefficient matrices. Our method demonstrates a strong ability to identify signal covariates and to control false positive rate compared to other simpler models. Both linear reduced rank and row-sparse reduced rank models fail to recover the optimal rank when noise variables are added. Our method, on the other hand, selects the rank close to the optimal rank with the presence of extra noise variables. The relatively small magnitude of the fifth singular value (10% of the sum of all singular values) might explain why the selected rank is less than 5 on average.

2.7 Conclusion and Future Work

We proposed a MSIM model to deal with potential nonlinearity in multivariate regression problems. To exploit the sparsity structures on the index coefficient matrix, our method allowed for multiple penalties to coexist. A simple block update approach concerning each penalty at a time was developed using the ADMM algorithm. Our simulation and real data analysis demonstrated the ability of the algorithm to recover structured coefficient matrices and the superiority to relax the common linear assumption in multivariate regression.

In our model we make an independent and homogeneous error assumption, i.e. $\text{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2 \mathbf{I}_q$. Recent work by Rothman et al. [Rot10], Lee & Liu [LL12] and Yin & Li [YL13], however, considered estimating both a sparse coefficient matrix and a sparse inverse covariance matrix using penalized

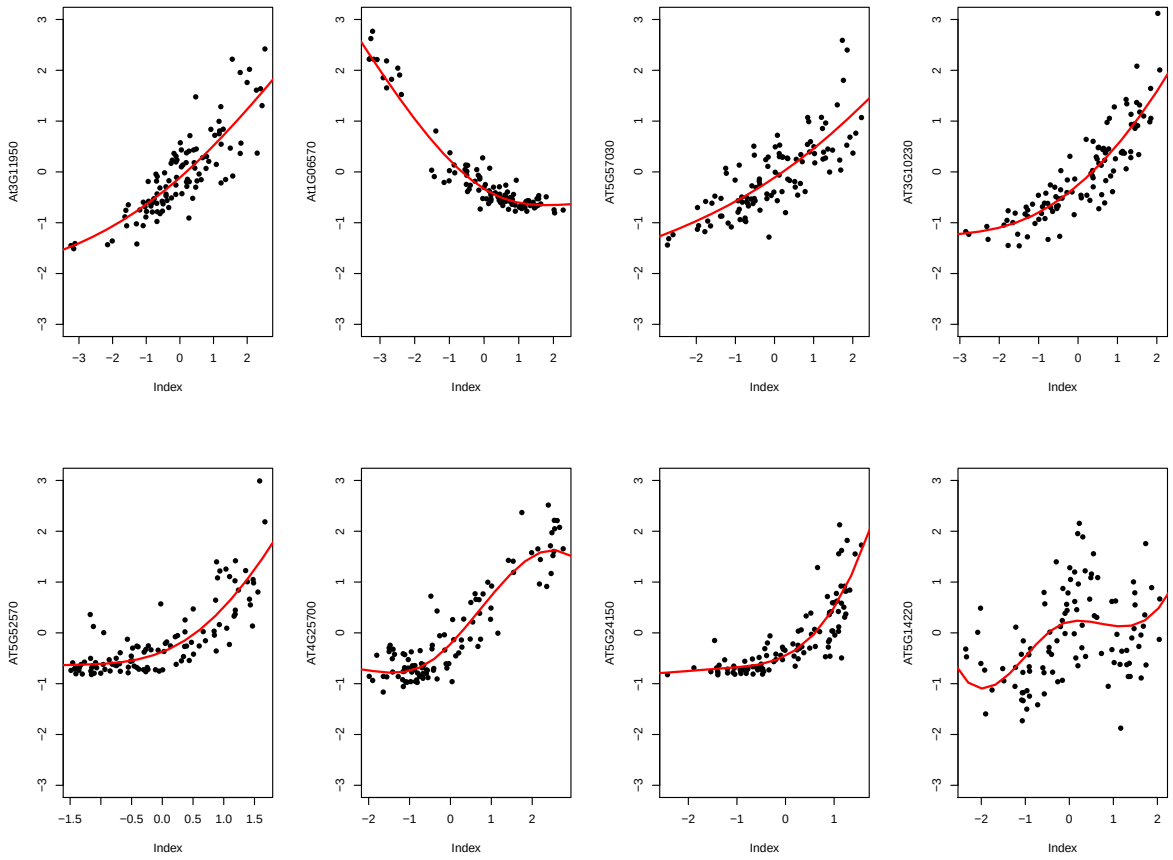

Figure 2.2 Mean functions for some responses using MSIM model for the gene pathway data.

likelihood to further improve prediction in the multivariate linear regression framework. A future research direction is to also incorporate inverse covariance matrix penalization into our framework.

CHAPTER

3

PENALIZED MATRIX FACTORIZATION

3.1 Introduction

Data in matrix form are naturally and frequently produced in modern scientific applications. For example in the matrix completion problem, we seek to recover a complete matrix from a sparse sampling of its entries [CR09]; in the multivariate response regression problem, a joint model is specified for the response vector and the regression coefficients for each response can be put compactly as a coefficient matrix [BF97]; in the matrix regression problem, the connection between the matrix type covariates and the response variable is established by learning a matrix of the same dimension as the covariate matrix [ZL14]; etc. Applications of these models can be seen in large scale recommender system [Eks11], gene-gene association study [Vou10] and brian image regression [Zho13], respectively.

For all these models, one common feature is that they all involve the estimation of a parameter

matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$ under a properly defined but problem specific loss function $\mathscr{L}(\mathbf{B})$. For matrix completion, let $\mathbf{Y} \in \mathbb{R}^{p\times q}$ be the underlying complete matrix and $\mathbf{W} \in \{0,1\}^{p\times q}$ be the indicator matrix with elements (in the jth row and kth column)

$$w_{jk} = \begin{cases} 1 & \text{if } y_{jk} \text{ is observed,} \\ 0 & \text{if } y_{jk} \text{ is missing,} \end{cases}$$

we can recover $\mathbf{Y}$ by minimizing a least squares loss

$$\mathscr{L}^{\mathrm{MC}}(\mathbf{B}) = \frac{1}{2}\|\mathbf{W} \circ (\mathbf{Y} - \mathbf{B})\|_{\mathrm{F}}^2,$$

where $\circ$ is the Hadamard product and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. For multivariate response (linear) regression, the model for data $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^q \times \mathbb{R}^p$ can be written in matrix form $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$, where $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times q}$ is the response matrix, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times p}$ is the design matrix and $\mathbf{E} = (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times q}$ is the error matrix. It is also straightforward to use a loss function

$$\mathscr{L}^{\mathrm{MRR}}(\mathbf{B}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_{\mathrm{F}}^2.$$

For matrix regression, define $\langle \cdot, \cdot \rangle$ as the (Frobenius) inner product of two matrices, the model is $y_i = \langle \mathbf{B}, \mathbf{X}_i \rangle + \epsilon_i$ with no further covariates included in the model for simplicity. Similarly we can use

$$\mathscr{L}^{\mathrm{MR}}(\mathbf{B}) = \frac{1}{2}\sum_{i=1}^{n} \big(y_i - \langle \mathbf{B}, \mathbf{X}_i \rangle\big)^2$$

as the loss function, where $y_i \in \mathbb{R}^1$ and $\mathbf{X}_i \in \mathbb{R}^{p\times q}$.

Challenges naturally arise in these applications when the scale of the problem is large (large p and large q). Therefore, dimension reduction is necessary for these models in high dimensional settings. As a common general scheme, regularization techniques have been extensively studied for matrix in the statistical literature [Obo08; Li15] and such a problem can be formulated as a penalization problem on the complexity of $\mathbf{B}$. A popular statistical framework that incorporates

structured matrix penalization is to optimize

$$\underset{\mathbf{B}}{\operatorname{argmin}} \quad \mathscr{L}(\mathbf{B}) + \mathscr{P}_\lambda(\mathbf{B}), \tag{3.1}$$

where λ is a tuning parameter in the penalty function $\mathscr{P}$ that controls the overall 'dimension' of $\mathbf{B}$. The most common measure of complexity of a matrix is its rank. Consequently, a direct rank penalty in the form of

$$\mathscr{P}_\lambda^{\mathrm{RK}}(\mathbf{B}) = \lambda \operatorname{rank}(\mathbf{B})$$

might be desired. However, the exercise of rank penalty is largely prohibited due to its discontinuous and non-convex nature. Recently, the nuclear norm (also known as the trace norm, denoted by $\|\cdot\|_*$), as a convex relaxation of the rank, is suggested as an alternative measure of complexity for matrix in many applications. In this case,

$$\mathscr{P}_\lambda^{\mathrm{NNP}}(\mathbf{B}) = \lambda \|\mathbf{B}\|_*.$$

Extensions of the nuclear norm penalty can be found in [Che13] using an adaptive procedure, but it is not the focus of this paper.

Despite the fact that regularization techniques can solve problems under a high dimensional setting, computational challenges still exist for very large scaled problems. For example, the modern recommender system in e-commerce usually involves completing matrices of millions of users as rows and millions of merchandise as columns, making algorithms that deal with matrix penalization fail immediately. This issue will be discussed in detail in the following chapter. To address the challenge, further dimension reduction methods based on matrix factorization have been proposed for large scale collaborative filtering and matrix completion in various literature [Lee10; Cil17]. Motivated by this line of research, we will generalize this strategy to both multivariate response regression models and matrix regression models. In Section 3.2 we review current computational

methods and discuss the necessity of bi-linear factorization with various penalties. In Section 3.3 and Section 3.4 we implement our proposed algorithms through simulation studies and justify the strength of these algorithms. In Section 3.5 we make conclusions and point out future working directions.

3.2 Penalized Matrix Factorization

We begin this section by looking into the current computational strategies for matrix penalization models. Simple but effective algorithms have been developed for some special models: for example, the non-iterative hard-thresholding algorithm for multivariate response regression with rank penalty [Bun11]. However, due to the potentially different forms of the loss function, a more general algorithm is usually iterative. A generalized proximal point algorithm [FM81] can be applied to solve optimization problem (3.1) when the loss function is smooth. The algorithm is sketched in Algorithm 2. The idea is to, at each iteration, replace the loss function $\mathscr{L}(\mathbf{B})$ by a quadratic approximation localized at previous $\mathbf{B}_k$, where k is the current iteration number. Then, for many penalty functions, step-3 in Algorithm 2 can be solved in closed form at each iteration. We can stop the algorithm when the consecutive $\mathbf{B}$'s are close, but we will literally use notation $\mathbf{B}_\infty$ to denote the final estimate. Note that we use a line search technique in step-4 for the update.

Algorithm 2 Proximal point algorithm.

INPUT: $\mathbf{Y}, \mathbf{X}, \mathbf{B}_0$, other algorithm parameters
OUTPUT: $\mathbf{B}$

1: $k \leftarrow 0$
2: **repeat**
3: $\quad \widehat{\mathbf{B}}_k = \underset{\mathbf{B}}{\operatorname{argmin}} \langle \nabla f(\mathbf{B}), \mathbf{B} \rangle + \tau_k^{-1} \|\mathbf{B} - \mathbf{B}_k\|_{\mathrm{F}}^2 + \mathscr{P}_\lambda(\mathbf{B})$
4: $\quad$ search for smallest l s.t. $f\left(\mathbf{B}_k + \gamma^l(\widehat{\mathbf{B}}_k - \mathbf{B}_k)\right) \leq f(\mathbf{B}_k) - \alpha\gamma^l \|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_{\mathrm{F}}^2$
5: $\quad \mathbf{B}_{k+1} \leftarrow (1-\gamma^l)\mathbf{B}_k + \gamma^l \widehat{\mathbf{B}}_k$
6: $\quad k \leftarrow k+1$
7: **until** convergence
8: **return** $\mathbf{B}_\infty$

Despite the generosity of the proximal point algorithm, this iterative procedure can have a high computation cost when the closed form solution for step-3 is expensive to compute. For instance, when using $\mathscr{P}_\lambda^{\mathrm{NNP}}$ as the penalty function, the closed form solution is obtained through a soft-threshoding of the singular values of $c_k\mathbf{B}_k$ where c_k is some constant. This requires a singular value decomposition (SVD) of a $p\times q$ matrix at each iteration, and the cost is $O(q^3)$ per iteration if p and q are of the same order. The decomposition of a matrix seriously prohibits the computation in large scale applications.

To address this problem in practice, the target matrix can be specified to have an explicit low rank bi-factorization structure such that $\mathbf{B}=\mathbf{L}\mathbf{R}^{\mathsf{T}}$, where $\mathbf{L}\in\mathbb{R}^{p\times r}$ and $\mathbf{R}\in\mathbb{R}^{q\times r}$. This formulation yields a much smaller number of $r(p+q)$ parameters to be estimated instead of $p\times q$ parameters when $r\ll\min(p,q)$. Note that r is fixed beforehand and the rank of $\mathbf{B}$ is bounded by r. Through the bi-linear factorization, it is easy to see that bounding the rank corresponds to constrain the dimension of each row of $\mathbf{L}$ and $\mathbf{R}$. Similarly, we can constrain the norms of $\mathbf{L}$ and $\mathbf{R}$ and the optimization problem (3.1) becomes

$$\underset{\mathbf{L},\mathbf{R}}{\operatorname{argmin}}\quad \mathscr{L}(\mathbf{L},\mathbf{R})+\mathscr{P}_\lambda(\mathbf{L},\mathbf{R}). \tag{3.2}$$

To exploit the bi-factorization structure of the matrix, on the one hand, a variational form of the nuclear norm is obtained through $\|\mathbf{B}\|_*=\min_{\mathbf{B}=\mathbf{L}\mathbf{R}^{\mathsf{T}}}\frac{1}{2}(\|\mathbf{L}\|_{\mathrm{F}}^2+\|\mathbf{R}\|_{\mathrm{F}}^2)$. The benefit is that now the product of the two matrices are separated as the sum of their Frobenius norms, which is generally easy to deal with. We put this surrogate function into framework (3.2) so that the nuclear norm surrogate penalty is

$$\mathscr{P}_\lambda^{\mathrm{SUR}}(\mathbf{L},\mathbf{R})=\frac{\lambda}{2}(\|\mathbf{L}\|_{\mathrm{F}}^2+\|\mathbf{R}\|_{\mathrm{F}}^2).$$

One the other hand, we can penalize the maximum length of the row vectors (denoted by $\|\cdot\|_{2,\infty}$ for matrix and $\|\cdot\|_\infty$ for vector) instead of penalizing the average length. To achieve this, we first

observe that $\|\mathbf{B}\|_{\max} = \min_{\mathbf{B}=\mathbf{L}\mathbf{R}^\mathsf{T}} \max(\|\mathbf{L}\|_{2,\infty}^2, \|\mathbf{R}\|_{2,\infty}^2)$, then

$$\mathscr{P}_\lambda^{\mathrm{MAX}}(\mathbf{L},\mathbf{R}) = \lambda \max(\|\mathbf{L}\|_{2,\infty}^2, \|\mathbf{R}\|_{2,\infty}^2).$$

The connection between these two constraints can be established through the fact that $\|\mathbf{B}\|_* = \sum_j \sigma_j$, where $\mathbf{B} = \sum_j \sigma_j \mathbf{u}_j \mathbf{v}_j^\mathsf{T}$, $\|\mathbf{u}_j\|_2 = 1$, $\|\mathbf{v}_j\|_2 = 1$, and $\|\mathbf{B}\|_{\max} \approx \sum_j \sigma_j$, where $\mathbf{B} = \sum_j \sigma_j \mathbf{u}_j \mathbf{v}_j^\mathsf{T}$, $\|\mathbf{u}_j\|_\infty = 1$, $\|\mathbf{v}_j\|_\infty = 1$. More detailed discussion of the relationship can be found in [Jam87; SS05]. In the next section we will introduce the algorithms for both cases.

3.2.1 Nuclear Norm Surrogate Penalization

Without loss of generality we discuss the nuclear norm surrogate penalization under a multivariate response regression model and study the objective function

$$f(\mathbf{L},\mathbf{R}) = \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{L}\mathbf{R}^\mathsf{T}\|_{\mathrm{F}}^2 + \frac{1}{2}(\|\mathbf{L}\|_{\mathrm{F}}^2 + \|\mathbf{R}\|_{\mathrm{F}}^2).$$

Note that f is not a convex function jointly on $(\mathbf{L}, \mathbf{R})$, but a bi-convex function in the sense that $f(\mathbf{L} \mid \mathbf{R})$ and $f(\mathbf{R} \mid \mathbf{L})$ are convex. However, an intuitive alternative updating approach is not feasible: although to optimize $f(\mathbf{R} \mid \mathbf{L})$ is direct via $\mathbf{R} = (\mathbf{L}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{L} + \lambda\mathbf{I})^{-1}\mathbf{L}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{Y}$, to optimize $f(\mathbf{L} \mid \mathbf{R})$ requires solving a Sylvester equation $\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{L}\mathbf{R}^\mathsf{T}\mathbf{R} + \lambda\mathbf{L} - \mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{R} = 0$, which is $O(q^3)$ per iteration if p and q are of the same order. Moreover, the loss function can be different from the one in our discussion. Thus a more flexible algorithm is required.

We consequently consider a direct gradient descent approach

$$(\mathbf{L}_{k+1}, \mathbf{R}_{k+1})^\mathsf{T} = (\mathbf{L}_k, \mathbf{R}_k)^\mathsf{T} - \alpha_k \left(\nabla f(\mathbf{L}_k \mid \mathbf{R}_k), \nabla f(\mathbf{R}_k \mid \mathbf{L}_k)\right)^\mathsf{T},$$

with backtracking line search to determine the step size α_k at the kth iteration. The computation cost is $O(q^2)$ per iteration, which is much faster than the conditional updating approach. Also in practice, the number of total gradient descent steps is usually moderate, making this approach

more appealing in dealing with high dimensional matrices.

3.2.2 Max Norm Penalization

A squash algorithm is derived in [Lee10] from the KKT conditions for the proximal mapping of max norm penalization. The algorithm can optimize $\text{squash}(\mathbf{V},\lambda) = \underset{\mathbf{V}}{\text{argmin}} \frac{1}{2}\|\mathbf{Z}-(\mathbf{L}^\mathsf{T},\mathbf{R}^\mathsf{T})^\mathsf{T}\|_\text{F}^2 + \lambda\|\mathbf{Z}\|_{2,\infty}^2$ in $O(q^2)$ time when p and q are of the same order, where $\mathbf{Z}\in\mathbb{R}^{(p+q)\times r}$. Thus, Algorithm 2 is directly applicable to solve max norm penalization, with the squash function serving as step-3.

In the following two sections, we will use simulation studies to demonstrate the algorithms for different matrix penalized models.

3.3 Application to Multivariate Response Regression

Recall the multivariate response regression model $\mathbf{Y}=\mathbf{XB}+\mathbf{E}$ and assume that $\mathbf{Y}$ and $\mathbf{X}$ have been centered so that intercept is not present in the model. The focus is to compare the performance of coefficient estimation and computation speed under different penalizing approaches using $\mathscr{P}_\lambda^\text{NNP}$, $\mathscr{P}_\lambda^\text{SUR}$ and $\mathscr{P}_\lambda^\text{MAX}$. We conduct simulation studies under a similar setting to the existing literature [Bun11; Che13] on penalized multivariate response regression models. Let $b\in\mathbb{R}^+$, $\mathbf{B}_1\in\mathbb{R}^{p\times r}$ and $\mathbf{B}_2\in\mathbb{R}^{r\times q}$, the true coefficient matrix $\mathbf{B}$ is constructed as $\mathbf{B}=b\mathbf{B}_1\mathbf{B}_2$ so that it has rank r. Note that b is used to control the strength of signal and elements in $\mathbf{B}_1$ and $\mathbf{B}_2$ are i.i.d from $N(0,1)$. Further, the errors are also i.i.d from $N(0,1)$. Two dimension combinations, $p=q<n$ and $p=q>n$, are considered:

- Scenario I ($n=100,\quad p=q=25,\quad r=10$)

  The design matrix $\mathbf{X}\in\mathbb{R}^{n\times p}$ is generated by stacking $\mathbf{x}_i$ from i.i.d $MVN_p(\mathbf{0},\Gamma)$, where the elements $\gamma_{ij}=\rho^{|i-j|}$ in Γ for $i=1\cdots n$, $j=1\cdots p$. We will set different values for the correlation coefficient $\rho\in(0,1)$ to test performance.

- Scenario II ($n=20,\quad p=q=25,\quad r=5,\quad r_\mathbf{X}=10$)

  The design matrix $\mathbf{X}$ is generated by $\mathbf{X}=\mathbf{X}_0\Gamma^{1/2}$ with the same Γ as in Scenario I. However, $\mathbf{X}_0=$

$\mathbf{X}_1\mathbf{X}_2$ is constructed through a production of two matrices, where $\mathbf{X}_1 \in \mathbb{R}^{n\times r_\mathbf{X}}$ and $\mathbf{X}_2 \in \mathbb{R}^{r_\mathbf{X}\times p}$. Moreover, elements in both $\mathbf{X}_1$ and $\mathbf{X}_2$ are i.i.d from $N(0,1)$.

We run a total number of 50 simulations for each setting with randomly generated initial matrices $(\mathbf{L}_0, \mathbf{R}_0)$ from standard normal distribution. To evaluate estimation accuracy, we use mean squared error (MSE) criterion, where $\text{MSE}(\mathbf{B}, \widehat{\mathbf{B}}) = 100\|\mathbf{B}-\widehat{\mathbf{B}}\|_\text{F}^2/pq$, $\text{MSE}(\mathbf{XB}, \mathbf{X}\widehat{\mathbf{B}}) = 100\|\mathbf{XB}-\mathbf{X}\widehat{\mathbf{B}}\|_\text{F}^2/nq$ and $\widehat{\mathbf{B}} = \widehat{\mathbf{L}}\widehat{\mathbf{R}}^\mathsf{T}$. In order to make fair comparison with the results in [Bun11] and in [Che13], we also independently generate test dataset $(\mathbf{Y}^*, \mathbf{X}^*)$ in each simulation with sample size $n^* = 5n$ and calculate $\text{MSE}(\mathbf{Y}^*, \mathbf{X}^*\widehat{\mathbf{B}}) = \|\mathbf{Y}^*-\mathbf{X}^*\widehat{\mathbf{B}}\|_\text{F}^2/n^*q$ as the criteria to choose the tuning parameter λ. Simulation results are summarized in Table 3.1 and Table 3.2 for $\rho = 0.5$ and for other ρ's the results can be found in the appendix. In the table, 'RID' refers to the baseline performance using ridge regression; 'RSC' refers to the rank selection criterion for direct rank penalization; 'SUR' and 'MAX' refers to the nuclear norm surrogate penalization and max norm penalization, respectively. The number behind the column name indicates the pre-fixed rank in a bi-factorization based approach. We report the 20% trimmed mean of each metric. We also compare the computation time for each method. Note that for 'RID' and 'RSC', the algorithms are non-iterative thus we do not record the time.

From the results we are able to observe that when the pre-fixed rank is close to but greater than the true rank, the penalized bi-factorization approach can be applied successfully, and the estimation accuracy using the nuclear norm surrogate is better than using the max norm in the penalty in our simulation. Moreover, the estimation accuracy of using the surrogate is almost identical to nuclear norm in the penalty when the rank is properly selected.

We then study how the computation time is scaled with the problem size by increasing (n, p, q) m times to (mn, mp, mq). In order to achieve even faster computation, a stochastic gradient descent approach is applied to speed up calculation. We set the batch size to 1/8 of sample size in the experiment. We choose $b = 0.1$ and one specific rank for each scenario. Because 'SUR' has better estimation and prediction performances from previous studies, we only consider this penalty combined with bi-factorization. From Table 3.3 and Table 3.4 we can see that there is a great benefit in using a bi-factorization compared to penalize the nuclear norm of the original matrix, and the

Table 3.1 Scenario I, $\rho = 0.5$, true rank = 10.

b		RID	NNP	RSC5	SUR5	MAX5	RSC10	SUR10	MAX10	RSC15	SUR15	MAX15	RSC20	SUR20	MAX20
0.3	Est	2.3	1.7	17.9	18.0	18.0	1.3	1.2	1.7	2.0	1.6	2.4	2.2	1.7	2.8
	Pred	25.3	25.6	267.0	266.9	284.6	16.3	16.1	45.7	23.1	19.2	59.1	25.1	20.3	61.4
	Time	-	58.8	-	22.2	209.4	-	89.8	267.8	-	404.2	291.6	-	447.7	337.4
	Rank	-	18.0	-	-	-	-	-	-	-	-	-	-	-	-
0.1	Est	2.2	1.2	2.4	2.4	2.5	1.4	1.1	1.3	2.0	1.2	1.5	2.2	1.2	1.5
	Pred	24.5	16.8	36.7	36.0	38.0	17.6	15.3	19.5	22.6	16.2	21.3	24.3	16.5	21.1
	Time	-	65.0	-	37.1	187.5	-	184.6	221.3	-	381.0	263.3	-	513.6	312.2
	Rank	-	17.8	-	-	-	-	-	-	-	-	-	-	-	-
0.05	Est	2.2	0.9	1.2	1.0	1.1	1.6	0.9	1.0	2.0	0.9	1.0	2.1	0.9	1.0
	Pred	25.0	13.8	18.2	15.8	17.4	19.8	13.5	14.9	23.5	13.9	15.1	24.7	13.7	14.9
	Time	-	67.0	-	64.5	172.5	-	234.7	198.1	-	427.7	240.6	-	506.3	258.6
	Rank	-	15.7	-	-	-	-	-	-	-	-	-	-	-	-

Table 3.2 Scenario II, $\rho = 0.5$, true rank = 5.

b		RID	NNP	RSC3	SUR3	MAX3	RSC5	SUR5	MAX5	RSC8	SUR8	MAX8
0.3	Est	27.5	27.5	1537.1	32.0	33.2	1816.2	28.7	30.0	1831.9	27.9	30.0
	Pred	54.4	76.9	576.5	765.4	792.2	29.4	74.7	193.1	45.2	54.4	227.6
	Time	-	176.5	-	28.1	239.9	-	107.5	286.4	-	121.0	308.6
	Rank	-	8.1	-	-	-	-	-	-	-	-	-
0.1	Est	3.2	3.1	272.3	3.5	3.8	314.4	3.2	3.5	356.2	3.2	3.5
	Pred	47.5	39.7	95.2	105.3	121.7	31.3	33.1	51.4	46.1	37.4	60.4
	Time	-	181.5	-	72.4	267.3	-	191.2	307.6	-	229.2	325.6
	Rank	-	8.5	-	-	-	-	-	-	-	-	-
0.05	Est	0.9	0.9	41.4	0.9	1.0	53.4	0.9	0.9	68.6	0.9	0.9
	Pred	37.5	31.6	38.5	38.7	42.8	35.4	29.7	36.1	47.5	33.9	38.3
	Time	-	184.6	-	142.3	275.5	-	232.3	302.8	-	298.6	314.6
	Rank	-	7.5	-	-	-	-	-	-	-	-	-

Table 3.3 Scenario I, $\rho = 0.5$, true rank = 10. The three numbers are (Est, Pred, Time).

b	m	NNP	SUR15	SUR15-STOC
0.1	1	(1.3, 16.1, 17.7)	(1.3, 16.6, 21.1)	(1.8, 31.3, 10.1)
	2	(0.5, 13.2, 88.2)	(0.5, 11.5, 63.9)	(0.7, 23.8, 19.9)
	4	(0.2, 9.4, 596.9)	(0.2, 7.6, 164.4)	(0.2, 13.4, 82.0)
	8	(0.07, 6.4, 5423.3)	(0.07, 5.3, 479.3)	(0.05, 7.2, 285.9)

Table 3.4 Scenario II, $\rho = 0.5$, true rank = 5. The three numbers are (Est, Pred, Time).

b	m	NNP	SUR5	SUR5-STOC
0.1	1	(3.0, 38.9, 51.2)	(3.2, 32.8, 50.4)	(3.5, 67.6, 20.5)
	2	(1.6, 30.2, 205.3)	(1.7, 26.6, 188.8)	(2.1, 40.5, 41.7)
	4	(0.5, 19.7, 1362.2)	(0.5, 16.2, 405.3)	(0.7, 20.3, 118.0)
	8	(0.2, 10.5, -)	(0.2, 9.2, 1006.9)	(0.2, 11.2, 550.3)

stochastic gradient approach is particularly useful when the dimension is high.

3.4 Application to Matrix Regression

Recall the matrix regression model $y_i = \langle \mathbf{B}, \mathbf{X}_i \rangle + \epsilon_i, i = 1 \cdots n$ with no extra covariates to adjust effect for. The most common application of this model is for the two-dimensional digital image data, where at each row-column combination is a pixel recording the quantized brightness value of a color. The goal is usually to identify influential regions in such an image $\mathbf{X}$ that associated with the response variable. Here we only consider continuous response. The influential region can be indicated by a binary matrix $\mathbf{B}$ of the same dimension of $\mathbf{X}$, where the region is denoted by all the "1"s in $\mathbf{B}$; see the top left plot in Figure 3.1 as an example. In practice, the region construction can be well approximated by some low-rank matrix as the product of two matrices. Note that this reconstructed matrix is not necessarily binary, but has values that are close to the indicator matrix.

In the simulation study, the indicator matrix $\mathbf{B} \in \{0,1\}^{64\times 64}$ is a cross-shaped image and elements in $\mathbf{X}_i$ are generated iid from standard normal distribution. We vary the sample size n and signal-to-noise ration (SNR = var($\mathbf{y}$)/var($\boldsymbol{\epsilon}$)) to generate data and test the performance of the algorithm. For

Table 3.5 Simulation results of MSE for penalized matrix regression.

n	SNR = 1	8	64	512
200	$7.5(1.8)\times 10^{-2}$	$4.3(0.6)\times 10^{-2}$	$3.6(0.6)\times 10^{-2}$	$3.6(0.6)\times 10^{-2}$
400	$4.7(1.6)\times 10^{-2}$	$2.0(0.3)\times 10^{-2}$	$8.4(1.9)\times 10^{-3}$	$6.7(2.0)\times 10^{-3}$
600	$3.4(0.2)\times 10^{-2}$	$1.0(0.1)\times 10^{-2}$	$2.0(0.5)\times 10^{-3}$	$9.3(3.2)\times 10^{-4}$

this problem, we only consider nuclear norm surrogate penalization because a low-rank assumption is standard in image regression.

Algorithm 2 is still applicable for this problem. However, from our previous experience, randomly generating the initial matrix usually finds a local minimum; see the top middle plot in Figure 3.1 as an example. To deal with this, we use a warm-up initial value calculated from the Frank-Wolfe algorithm [Jag13] that minimizes $h(\mathbf{B}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \langle \mathbf{B}, \mathbf{X}_i\rangle)^2$ subject to $\|\mathbf{B}\|_* \leq \delta$. Thus, at iteration k, Frank-Wolfe computes $\widehat{\mathbf{B}}_k = \underset{\|\mathbf{B}\|_*\leq\delta}{\operatorname{argmin}} \langle \nabla h(\mathbf{B}_k), \mathbf{B}\rangle$ and updates $\mathbf{B}_{k+1} = \frac{k}{k+2}\mathbf{B}_k + \frac{2}{k+2}\widehat{\mathbf{B}}_k$. For nuclear norm penalty, $\widehat{\mathbf{B}}_k = -\delta \mathbf{u_1}\mathbf{v_1}^\top$ is a rank one matrix where $(\mathbf{u_1}, \mathbf{v_1})$ are the singular vectors associated with the largest singular value of $\nabla h(\mathbf{B}_k)$. Because at each iteration we perform a rank one update, a partial SVD with low computation cost (compared to full SVD) can satisfy our needs. In the simulation we use 10 iterations of Frank-Wolfe to warm-up; see the top right plot in Figure 3.1 as an example. With the initial value calculated from Frank-Wolfe, we then solve the penalized matrix regression problem and choose the tuning parameter through 5-fold cross-validation; see the bottom three plots in Figure 3.1 as examples of the final estimate with different values of λ.

Table 3.5 summarizes the mean MSE and its standard deviation for 50 simulations under each setting. When the sample size is small, the estimation accuracy reaches a lower bound. This is because now we use the information from n observations to estimate $r(p+q)$ parameters; in multivariate response regression we actually have nq observations. It is also observed that the estimation is more accurate with the increase of sample size and SNR.

3.5 Conclusion and Future Work

In this paper we applied matrix bi-factorization and penalization to solve some problems with matrix type coefficients. The proposed computation strategy has good estimation accuracy and can be easily scaled to solve large size problems.

A few key problems that requires further attention in our research include: 1) For the bi-factorization approach, the pre-fixed rank r must be greater than the true rank. This means that we need to bound the rank of the coefficient matrix of the original unpenalized problem. However, there is few literature on this topic. 2) In matrix regression, the recovered indicator matrix is noisy and we may want to filter the matrix so that the signal can be more clearly presented in an image. Denote the recovered noisy matrix as $\mathbf{B}_{(1)}$, a proper model is $\mathbf{B}_{(1)} = \mathbf{B}_{(2)} + \mathbf{E}$, or in its vectorized form $\mathbf{y} = \boldsymbol{\beta} + \text{vec}(\mathbf{E})$, where $\mathbf{y} = \text{vec}(\mathbf{B}_{(1)})$ and $\boldsymbol{\beta} = \text{vec}(\mathbf{B}_{(2)})$. Based on our experience, the classical lasso method [Tib96] can serve as a good filter, but it also requires tuning parameter selection. This leads to a two-step procedure that different tuning parameters are selected in each step. Recently a tuning-free approach called TREX is proposed in [LM15] as an alternative to the lasso. An algorithm for TREX that guarantees to find the global minimum of $\underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2}{\|\mathbf{X}^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_\infty} + \frac{\|\boldsymbol{\beta}\|_1}{2} \right\}$ in polynomial time is proposed in [Bie18]. We may utilize the tuning-free feature of this procedure, but currently the algorithm is designed to deal with an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. Since in our case the $\mathbf{X}$ is an identity matrix, directly using the algorithm will not benefit us from faster computation.

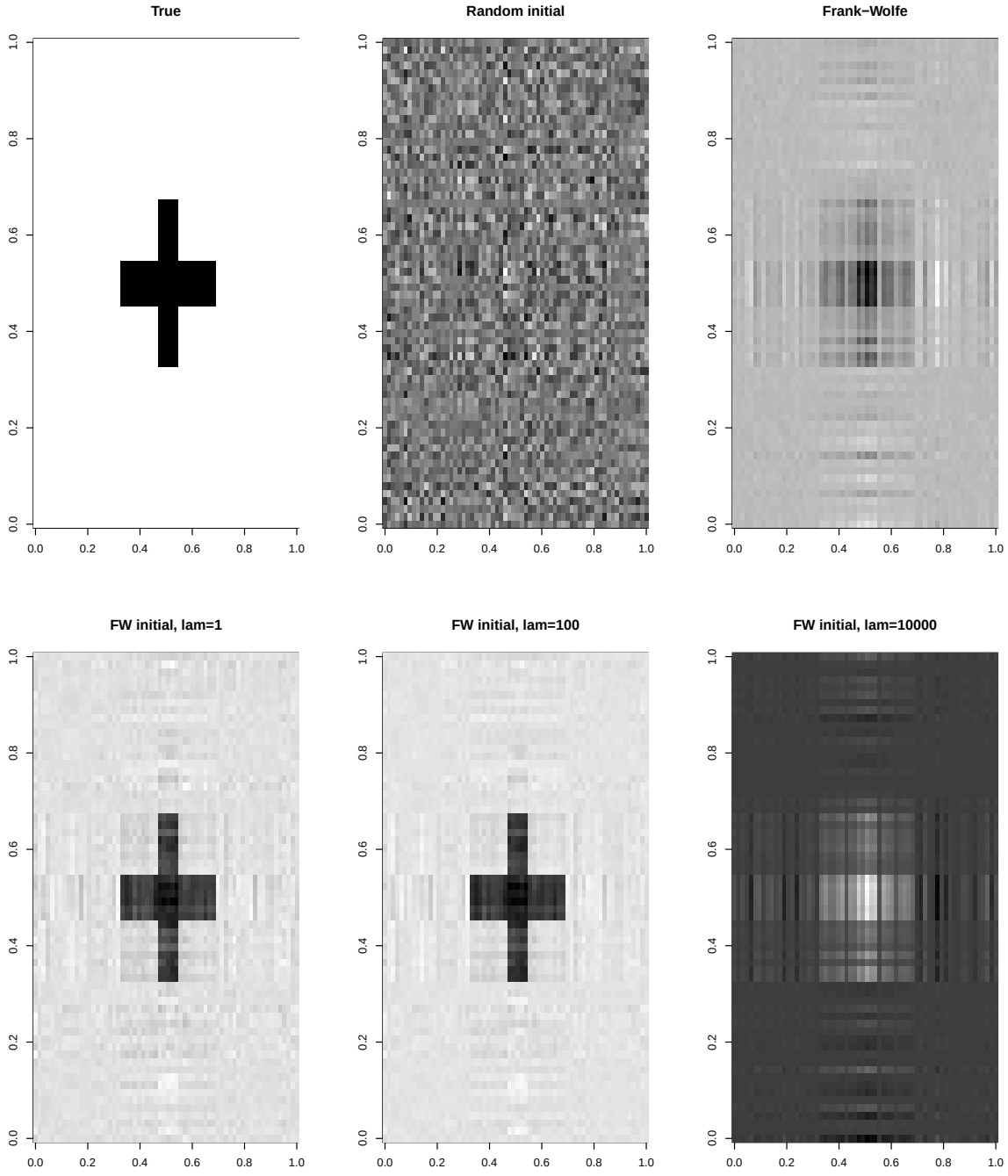

Figure 3.1 Illustration plots for penalized matrix regression. Top left is the true signal region; top middle is the recovered matrix using a randomly generated initial matrix; top right is the warm-up initial matrix using Frank-Wolfe; Bottom panel contains recovered matrices using warm-up initial matrix with different tuning parameters for penalization.

BIBLIOGRAPHY

[AH89] Altman, D. G. & Hytten, F. E. “Intrauterine growth retardation: let’s be clear about it”. *BJOG: An International Journal of Obstetrics & Gynaecology* **96**.10 (1989), pp. 1127–1128.

[And18] Anderson, C. et al. “Using data from multiple studies to develop a child growth correlation matrix”. *Statistics in Medicine* (2018).

[Arg08a] Argyle, J. et al. “Correlation models for monitoring child growth”. *Statistics in medicine* **27**.6 (2008), pp. 888–904.

[Arg07] Argyriou, A. et al. “Multi-task feature learning”. *Advances in neural information processing systems*. 2007, pp. 41–48.

[Arg08b] Argyriou, A. et al. “Convex multi-task feature learning”. *Machine Learning* **73**.3 (2008), pp. 243–272.

[Bie18] Bien, J. et al. “Non-convex global minimization and false discovery rate control for the TREX”. *Journal of Computational and Graphical Statistics* **27**.1 (2018), pp. 23–33.

[Boy11] Boyd, S. et al. “Distributed optimization and statistical learning via the alternating direction method of multipliers”. *Foundations and Trends® in Machine Learning* **3**.1 (2011), pp. 1–122.

[BF97] Breiman, L. & Friedman, J. H. “Predicting multivariate responses in multiple linear regression”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**.1 (1997), pp. 3–54.

[Bun11] Bunea, F. et al. “Optimal selection of reduced rank estimators of high-dimensional matrices”. *The Annals of Statistics* (2011), pp. 1282–1309.

[Bun12] Bunea, F. et al. “Joint variable and rank selection for parsimonious estimation of high-dimensional matrices”. *The Annals of Statistics* **40**.5 (2012), pp. 2359–2388.

[Cam80] Cameron, N. “Conditional standards for growth in height of British children from 5· 0 to 15· 99 years of age”. *Annals of human biology* **7**.4 (1980), pp. 331–337.

[CR09] Candès, E. J. & Recht, B. “Exact matrix completion via convex optimization”. *Foundations of Computational mathematics* **9**.6 (2009), p. 717.

[Car97] Carroll, R. J. et al. “Generalized partially linear single-index models”. *Journal of the American Statistical Association* **92**.438 (1997), pp. 477–489.

[Car98] Caruana, R. “Multitask learning”. *Learning to learn*. Springer, 1998, pp. 95–133.

[Che13] Chen, K. et al. “Reduced rank regression via adaptive nuclear norm penalization”. *Biometrika* **100**.4 (2013), pp. 901–920.

[CH12] Chen, L. & Huang, J. Z. “Sparse reduced-rank regression for simultaneous dimension reduction and variable selection”. *Journal of the American Statistical Association* **107**.500 (2012), pp. 1533–1545.

[Cil17] Ciliberto, C. et al. “Reexamining low rank matrix factorization for trace norm regularization”. *arXiv preprint arXiv:1706.08934* (2017).

[CG92] Cole, T. J. & Green, P. J. “Smoothing reference centile curves: the LMS method and penalized likelihood”. *Statistics in medicine* **11**.10 (1992), pp. 1305–1319.

[Col94] Cole, T. “Growth charts for both cross-sectional and longitudinal data”. *Statistics in medicine* **13**.23-24 (1994), pp. 2477–2492.

[Col98] Cole, T. “Presenting information on growth distance and conditional velocity in one chart: practical issues of chart design”. *Statistics in medicine* **17**.23 (1998), pp. 2697–2707.

[CW05] Combettes, P. L. & Wajs, V. R. “Signal Recovery by Proximal Forward-Backward Splitting”. *Multiscale Modeling & Simulation* **4**.4 (2005), pp. 1168–1200.

[Dig88] Diggle, P. J. “An approach to the analysis of repeated measurements”. *Biometrics* (1988), pp. 959–971.

[EM96] Eilers, P. H. & Marx, B. D. “Flexible smoothing with B-splines and penalties”. *Statistical science* (1996), pp. 89–102.

[EM03] Eilers, P. H. & Marx, B. D. “Multivariate calibration with temperature interaction using two-dimensional penalized signal regression”. *Chemometrics and intelligent laboratory systems* **66**.2 (2003), pp. 159–174.

[Eks11] Ekstrand, M. D. et al. “Collaborative filtering recommender systems”. *Foundations and Trends® in Human–Computer Interaction* **4**.2 (2011), pp. 81–173.

[EP04] Evgeniou, T. & Pontil, M. “Regularized multi–task learning”. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2004, pp. 109–117.

[FL01] Fan, J. & Li, R. “Variable selection via nonconcave penalized likelihood and its oracle properties”. *Journal of the American statistical Association* **96**.456 (2001), pp. 1348–1360.

[Fan11] Fan, J. et al. “Nonparametric independence screening in sparse ultra-high-dimensional additive models”. *Journal of the American Statistical Association* **106**.494 (2011), pp. 544–557.

[Fos13] Foster, J. C. et al. “Variable selection in monotone single-index models via the adaptive LASSO”. *Statistics in medicine* **32**.22 (2013), pp. 3944–3954.

[FM81] Fukushima, M. & Mine, H. “A generalized proximal point algorithm for certain non-convex minimization problems”. *International Journal of Systems Science* **12**.8 (1981), pp. 989–1000.

[GM76] Gabay, D. & Mercier, B. “A dual algorithm for the solution of nonlinear variational problems via finite element approximation”. *Computers & Mathematics with Applications* **2**.1 (1976), pp. 17–40.

[Lei] “Gestational weight gain standards based on women enrolled in the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project: a prospective longitudinal cohort study”. *BMJ* **352** (2016).

[GLT89] Glowinski, R. & Le Tallec, P. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*. Vol. 9. SIAM, 1989.

[Hai98] Hair, J. F. et al. *Multivariate data analysis*. Vol. 5. 3. Prentice hall Upper Saddle River, NJ, 1998.

[HS89] Härdle, W. & Stoker, T. M. “Investigating smooth multiple regression by the method of average derivatives”. *Journal of the American Statistical Association* **84**.408 (1989), pp. 986–995.

[Hri01] Hristache, M. et al. “Direct estimation of the index coefficient in a single-index model”. *Annals of Statistics* (2001), pp. 595–623.

[HDH85] Hutchinson, M. F. & De Hoog, F. “Smoothing noisy data with spline functions”. *Numerische Mathematik* **47**.1 (1985), pp. 99–106.

[Ich93] Ichimura, H. “Semiparametric least squares (SLS) and weighted SLS estimation of single-index models”. *Journal of Econometrics* **58**.1-2 (1993), pp. 71–120.

[Vila] “International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project”. *The Lancet* **384**.9946 (2014), pp. 857 –868.

[Jag13] Jaggi, M. “Revisiting Frank-Wolfe: Projection-free sparse convex optimization.” *ICML (1)*. 2013, pp. 427–435.

[Jam87] Jameson, G. J. O. *Summing and nuclear norms in Banach space theory*. Vol. 8. Cambridge University Press, 1987.

[JS16] Josse, J. & Sardy, S. “Adaptive shrinkage of singular values”. *Statistics and Computing* **26**.3 (2016), pp. 715–724.

[KX07] Kong, E. & Xia, Y. “Variable selection for the single-index model”. *Biometrika* **94**.1 (2007), pp. 217–229.

[LM15] Lederer, J. & Müller, C. “Don’t fall for tuning parameters: tuning-free variable selection in high dimensions with the TREX”. *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[Lee10] Lee, J. D. et al. “Practical large-scale optimization for max-norm regularization”. *Advances in neural information processing systems*. 2010, pp. 1297–1305.

[LL12] Lee, W. & Liu, Y. “Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood”. *Journal of multivariate analysis* **111** (2012), pp. 241–255.

[Li15] Li, Y. et al. “Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure”. *Biometrics* **71**.2 (2015), pp. 354–363.

[Lia10] Liang, H. et al. “Estimation and testing for partially linear single-index models”. *Annals of statistics* **38**.6 (2010), p. 3811.

[Ma14] Ma, X. et al. “Learning regulatory programs by threshold SVD regression”. *Proceedings of the National Academy of Sciences* **111**.44 (2014), pp. 15675–15680.

[Muk15] Mukherjee, A et al. “On the degrees of freedom of reduced-rank estimators in multivariate regression”. *Biometrika* **102**.2 (2015), pp. 457–477.

[Obo08] Obozinski, G. et al. “Union support recovery in high-dimensional multivariate regression”. *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*. IEEE. 2008, pp. 21–26.

[Obo10] Obozinski, G. et al. “Joint covariate selection and joint subspace selection for multiple classification problems”. *Statistics and Computing* **20**.2 (2010), pp. 231–252.

[Ohu18] Ohuma, E. O. et al. “Statistical methodology for constructing gestational age-related charts using cross-sectional and longitudinal data: The INTERGROWTH-21st project as a case study”. *Statistics in medicine* (2018).

[Pap14] Papageorghiou, A. T. et al. “International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project”. *The Lancet* **384**.9946 (2014), pp. 869–879.

[PB14] Parikh, N., Boyd, S., et al. “Proximal algorithms”. *Foundations and Trends® in Optimization* **1**.3 (2014), pp. 127–239.

[PH11] Peng, H. & Huang, T. “Penalized least squares for single index models”. *Journal of Statistical Planning and Inference* **141**.4 (2011), pp. 1362–1379.

[Pen10] Peng, J. et al. “Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer”. *The annals of applied statistics* **4**.1 (2010), p. 53.

[Pol15] Polson, N. G. et al. “Proximal algorithms in statistics and machine learning”. *Statistical Science* **30**.4 (2015), pp. 559–581.

[RS04] Rigby, R. A. & Stasinopoulos, D. M. “Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution”. *Statistics in medicine* **23**.19 (2004), pp. 3053–3076.

[RS06] Rigby, R. A. & Stasinopoulos, D. M. “Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis”. *Statistical Modelling* **6**.3 (2006), pp. 209–229.

[Rot10] Rothman, A. J. et al. “Sparse multivariate regression with covariance estimation”. *Journal of Computational and Graphical Statistics* **19**.4 (2010), pp. 947–962.

[Rup02] Ruppert, D. “Selecting the number of knots for penalized splines”. *Journal of computational and graphical statistics* **11**.4 (2002), pp. 735–757.

[Rup03] Ruppert, D. et al. “Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics 12”. *Cambridge: Cambridge Univ. Press. Mathematical Reviews (MathSciNet): MR1998720* (2003).

[Sar13] Sarris, I et al. “Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project”. *BJOG: An International Journal of Obstetrics & Gynaecology* **120** (2013), pp. 33–37.

[SC17] She, Y. & Chen, K. “Robust reduced-rank regression”. *Biometrika* **104**.3 (2017), pp. 633–647.

[SS05] Srebro, N. & Shraibman, A. “Rank, trace-norm and max-norm”. *International Conference on Computational Learning Theory*. Springer. 2005, pp. 545–560.

[SR07] Stasinopoulos, D. M., Rigby, R. A., et al. “Generalized additive models for location scale and shape (GAMLSS) in R”. *Journal of Statistical Software* **23**.7 (2007), pp. 1–46.

[TW76] Tanner, J. M. & Whitehouse, R. H. “Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty.” *Archives of disease in childhood* **51**.3 (1976), pp. 170–179.

[Vilb] “The likeness of fetal growth and newborn size across non-isolated populations in the INTERGROWTH-21st Project: the Fetal Growth Longitudinal Study and Newborn Cross-Sectional Study”. *The Lancet Diabetes & Endocrinology* **2**.10 (2014), pp. 781 –792.

[Tib96] Tibshirani, R. “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[VR13] Velu, R. & Reinsel, G. C. *Multivariate reduced-rank regression: theory and applications*. Vol. 136. Springer Science & Business Media, 2013.

[Vil13] Villar, J et al. “The objectives, design and implementation of the INTERGROWTH-21st Project”. *BJOG: An International Journal of Obstetrics & Gynaecology* **120** (2013), pp. 9–26.

[Vil15] Villar, J. et al. “Postnatal growth standards for preterm infants: the Preterm Postnatal Follow-up Study of the INTERGROWTH-21st Project”. *The Lancet Global Health* **3**.11 (2015), e681–e691.

[Vil18] Villar, J. et al. “The satisfactory growth and development at 2 years of age of the INTERGROWTH-21st Fetal Growth Standards cohort support its appropriateness for constructing international standards”. *American journal of obstetrics and gynecology* **218**.2 (2018), S841–S854.

[Vou10] Vounou, M. et al. “Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach”. *Neuroimage* **53**.3 (2010), pp. 1147–1159.

[Wan10] Wang, J.-L. et al. “Estimation for a partial-linear single-index model”. *The Annals of statistics* **38**.1 (2010), pp. 246–274.

[WY09] Wang, L. & Yang, L. “Spline estimation of single-index models”. *Statistica Sinica* (2009), pp. 765–783.

[Wan15] Wang, Y. et al. “Global convergence of ADMM in nonconvex nonsmooth optimization”. *arXiv preprint arXiv:1511.06324* (2015).

[Wil04] Wille, A. et al. “Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana”. *Genome biology* **5**.11 (2004), R92.

[Wri94] Wright, C. et al. “What is a normal rate of weight gain in infancy?” *Acta Paediatrica* **83**.4 (1994), pp. 351–356.

[Xia18] Xiao, L. et al. “Fast covariance estimation for sparse functional data”. *Statistics and computing* **28**.3 (2018), pp. 511–522.

[Yao05] Yao, F. et al. “Functional data analysis for sparse longitudinal data”. *Journal of the American Statistical Association* **100**.470 (2005), pp. 577–590.

[YL13] Yin, J. & Li, H. “Adjusting for high-dimensional covariates in sparse precision matrix estimation by âĎŞ1-penalization”. *Journal of multivariate analysis* **116** (2013), pp. 365–381.

[YR02] Yu, Y. & Ruppert, D. “Penalized spline estimation for partially linear single-index models”. *Journal of the American Statistical Association* **97**.460 (2002), pp. 1042–1054.

[YL06] Yuan, M. & Lin, Y. “Model selection and estimation in regression with grouped variables”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**.1 (2006), pp. 49–67.

[Yua07] Yuan, M. et al. “Dimension reduction and coefficient estimation in multivariate linear regression”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**.3 (2007), pp. 329–346.

[Zha10] Zhang, C.-H. et al. “Nearly unbiased variable selection under minimax concave penalty”. *The Annals of statistics* **38**.2 (2010), pp. 894–942.

[ZL14] Zhou, H. & Li, L. “Regularized matrix regression”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**.2 (2014), pp. 463–483.

[Zho13] Zhou, H. et al. “Tensor regression with applications in neuroimaging data analysis”. *Journal of the American Statistical Association* **108**.502 (2013), pp. 540–552.

[Zou06] Zou, H. “The adaptive lasso and its oracle properties”. *Journal of the American statistical association* **101**.476 (2006), pp. 1418–1429.

[ZH05] Zou, H. & Hastie, T. “Regularization and variable selection via the elastic net”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**.2 (2005), pp. 301–320.

APPENDIX

APPENDIX

A

SUPPLEMENTAL MATERIAL

A.1 Correlation Models for Monitoring Fetal Growth

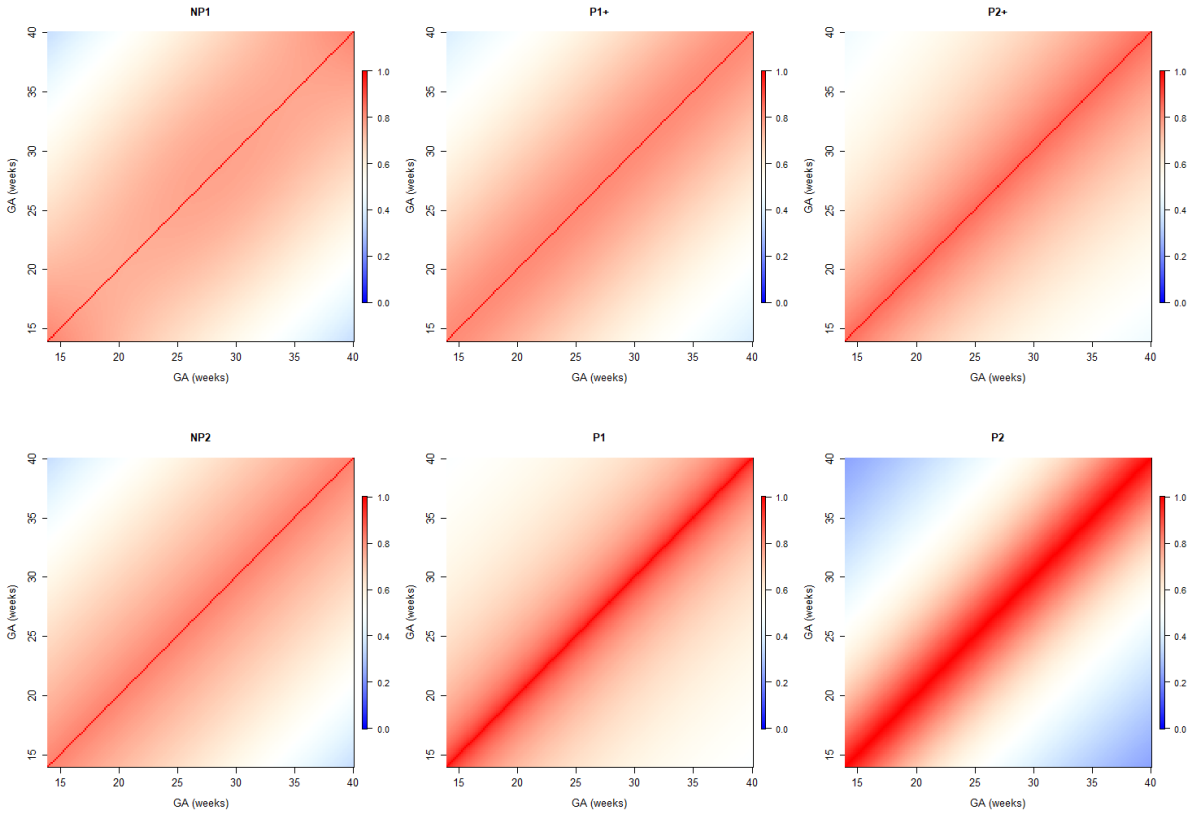

Figure A.1 Temporal correlations of standardized FL with different correlation models.

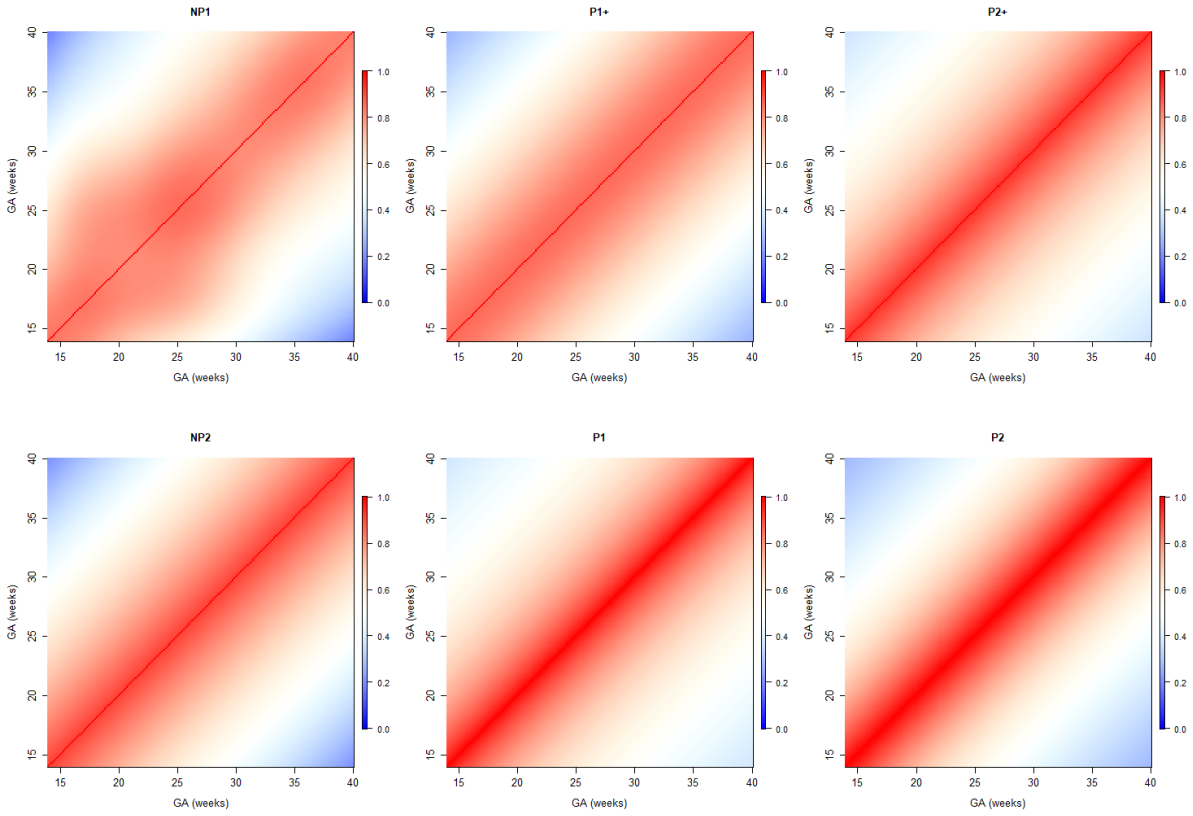

Figure A.2 Temporal correlations of standardized HC with different correlation models.

Table A.1 Correlation matrix for AC

Week	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
14	1.00																										
15	0.77	1.00																									
16	0.76	0.77	1.00																								
17	0.75	0.76	0.77	1.00																							
18	0.74	0.75	0.76	0.77	1.00																						
19	0.72	0.74	0.75	0.76	0.77	1.00																					
20	0.70	0.72	0.74	0.75	0.76	0.77	1.00																				
21	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00																			
22	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00																		
23	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00																	
24	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00																
25	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00															
26	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00														
27	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00													
28	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00												
29	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00											
30	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00										
31	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00									
32	0.45	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00								
33	0.43	0.45	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00							
34	0.41	0.43	0.45	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00						
35	0.39	0.41	0.43	0.45	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00					
36	0.37	0.39	0.41	0.43	0.45	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00				
37	0.35	0.37	0.39	0.41	0.43	0.45	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00			
38	0.33	0.35	0.37	0.39	0.41	0.43	0.45	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00		
39	0.32	0.33	0.35	0.37	0.39	0.41	0.43	0.45	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00	
40	0.30	0.32	0.33	0.35	0.37	0.39	0.41	0.43	0.45	0.47	0.50	0.52	0.54	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.70	0.72	0.74	0.75	0.76	0.77	1.00

Table A.2 Correlation matrix for FL

Week	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
14	1.00																										
15	0.80	1.00																									
16	0.79	0.80	1.00																								
17	0.78	0.79	0.80	1.00																							
18	0.77	0.78	0.79	0.80	1.00																						
19	0.76	0.77	0.78	0.79	0.80	1.00																					
20	0.74	0.76	0.77	0.78	0.79	0.80	1.00																				
21	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00																			
22	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00																		
23	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00																	
24	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00																
25	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00															
26	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00														
27	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00													
28	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00												
29	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00											
30	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00										
31	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00									
32	0.53	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00								
33	0.51	0.53	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00							
34	0.49	0.51	0.53	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00						
35	0.47	0.49	0.51	0.53	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00					
36	0.46	0.47	0.49	0.51	0.53	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00				
37	0.44	0.46	0.47	0.49	0.51	0.53	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00			
38	0.42	0.44	0.46	0.47	0.49	0.51	0.53	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00		
39	0.41	0.42	0.44	0.46	0.47	0.49	0.51	0.53	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00	
40	0.39	0.41	0.42	0.44	0.46	0.47	0.49	0.51	0.53	0.55	0.56	0.58	0.60	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.74	0.76	0.77	0.78	0.79	0.80	1.00

Table A.3 Correlation matrix for HC

Week	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
14	1.00																										
15	0.85	1.00																									
16	0.84	0.85	1.00																								
17	0.82	0.84	0.85	1.00																							
18	0.80	0.82	0.84	0.85	1.00																						
19	0.78	0.80	0.82	0.84	0.85	1.00																					
20	0.76	0.78	0.80	0.82	0.84	0.85	1.00																				
21	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00																			
22	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00																		
23	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00																	
24	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00																
25	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00															
26	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00														
27	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00													
28	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00												
29	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00											
30	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00										
31	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00									
32	0.44	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00								
33	0.41	0.44	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00							
34	0.39	0.41	0.44	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00						
35	0.36	0.39	0.41	0.44	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00					
36	0.34	0.36	0.39	0.41	0.44	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00				
37	0.32	0.34	0.36	0.39	0.41	0.44	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00			
38	0.30	0.32	0.34	0.36	0.39	0.41	0.44	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00		
39	0.28	0.30	0.32	0.34	0.36	0.39	0.41	0.44	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00	
40	0.26	0.28	0.30	0.32	0.34	0.36	0.39	0.41	0.44	0.46	0.49	0.51	0.54	0.57	0.60	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	1.00

A.2 Sparse Single Index Models for Multivariate Responses

A.2.1 Derivation of Gradients

For each $\mathbf{B}_{\cdot j}$ and the cubic spline representation of f_j, we can derive the gradient

$$\begin{aligned}\nabla g_j(\mathbf{B}_{\cdot j}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{f_j(\mathbf{x}_i^\top\mathbf{B}_{\cdot j})-y_{ij}\right\}f_j'(\mathbf{x}_i^\top\mathbf{B}_{\cdot j})\mathbf{x}_i-\sum_{l=0}^{h}\mathbf{M}_{l,\cdot j}+\rho\left\{(h+1)\mathbf{B}_{\cdot j}-\sum_{l=0}^{h}\mathbf{C}_{l,\cdot j}\right\}\\ &= \frac{1}{n}\mathbf{X}^\top\left\{(f_j(\mathbf{X}\mathbf{B}_{\cdot j})-\mathbf{Y}_{\cdot j})\circ f_j'(\mathbf{X}\mathbf{B}_{\cdot j})\right\}-\sum_{l=0}^{h}\mathbf{M}_{l,\cdot j}+\rho\left\{(h+1)\mathbf{B}_{\cdot j}-\sum_{l=0}^{h}\mathbf{C}_{l,\cdot j}\right\}\end{aligned}$$

where $\circ$ is the Hadamard product. Similarly, the Hessian matrix is

$$\nabla^2 g_j(\mathbf{B}_{\cdot j}) = \frac{1}{n}\sum_{i=1}^{n}\Big(f_j'^2(\mathbf{x}_i^\top\mathbf{B}_{\cdot j})+f_j(\mathbf{x}_i^\top\mathbf{B}_{\cdot j})f_j''(\mathbf{x}_i^\top\mathbf{B}_{\cdot j})-y_i f_j''(\mathbf{x}_i^\top\mathbf{B}_{\cdot j})\Big)\mathbf{x}_i\mathbf{x}_i^\top+\rho(h+1)\mathbf{I}_p.$$

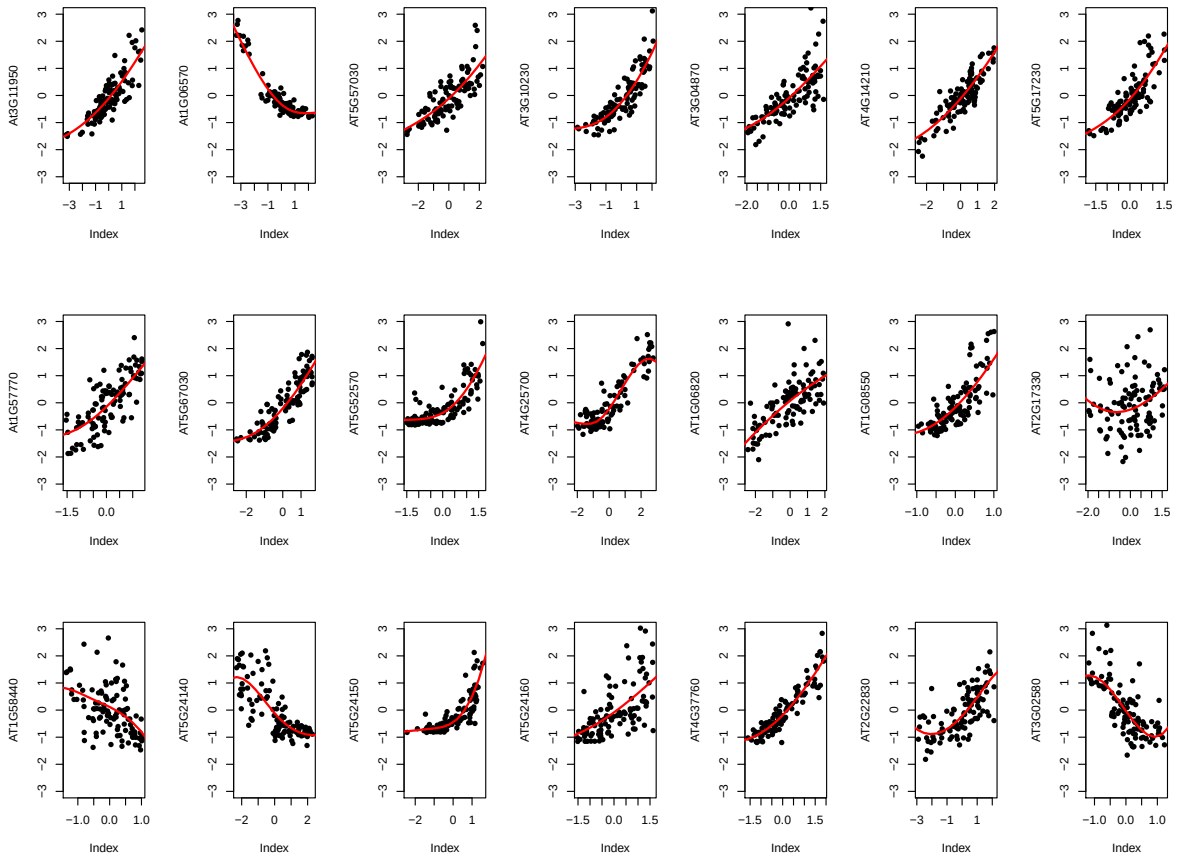

Figure A.3 Mean functions using MSIM model for the gene pathway data - I.

A.2.2 Mean functions Using MSIM Model for the Gene Pathway Data

Figures in this subsection display the estimated 62 responses from the genetic association study in section 6 of the main paper.

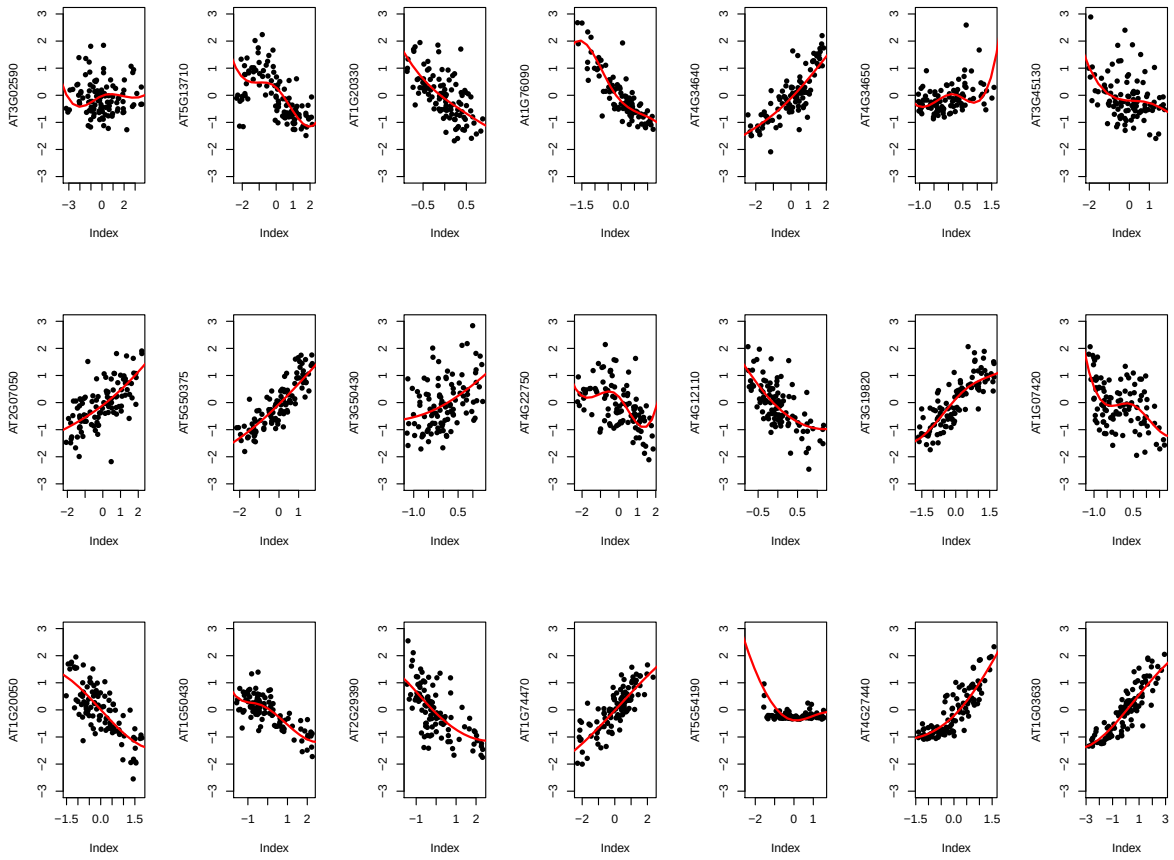

Figure A.4 Mean functions using MSIM model for the gene pathway data - II.

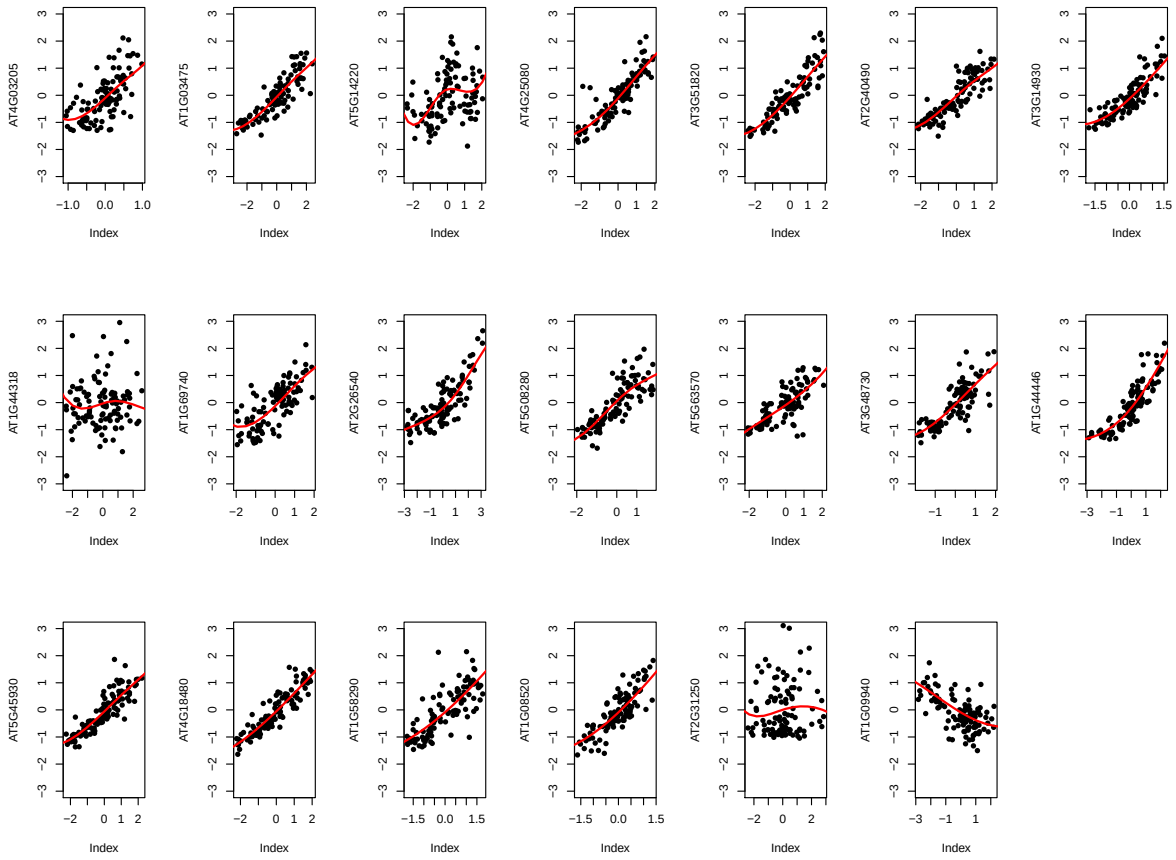

Figure A.5 Mean functions using MSIM model for the gene pathway data - III.

Table A.4 Scenario I, $\rho = 0.1$, true rank = 10.

b		RID	NNP	RSC5	SUR5	MAX5	RSC10	SUR10	MAX10	RSC15	SUR15	MAX15	RSC20	SUR20	MAX20
0.3	Est	1.4	1.1	14.6	14.6	14.8	0.8	0.8	1.2	1.3	1.0	1.6	1.4	1.1	1.8
	Pred	25.4	23.3	304.4	304.4	313.7	16.6	16.3	32.1	23.1	19.2	41.0	25.1	20.8	44.2
	Time	-	30.1	-	16.0	170.1	-	50.5	213.8	-	287.4	267.4	-	332.6	297.4
	Rank	-	18.7	-	-	-	-	-	-	-	-	-	-	-	-
0.1	Est	1.4	0.9	2.1	2.0	2.1	0.9	0.8	0.9	1.3	0.9	1.1	1.4	0.9	1.2
	Pred	25.3	17.9	42.3	41.9	43.4	17.7	15.8	19.2	23.3	17.3	23.1	25.0	17.7	23.2
	Time	-	33.9	-	28.8	141.6	-	142.6	199.5	-	287.3	232.1	-	364.1	265.5
	Rank	-	19.2	-	-	-	-	-	-	-	-	-	-	-	-
0.05	Est	1.4	0.7	0.9	0.8	0.9	1.1	0.7	0.8	1.3	0.7	0.8	1.4	0.7	0.8
	Pred	25.1	14.6	19.2	17.1	18.3	19.8	14.4	15.6	23.5	14.7	16.0	24.9	14.6	15.9
	Time	-	34.7	-	44.3	124.0	-	153.1	175.1	-	302.9	204.5	-	376.3	230.3
	Rank	-	12.3	-	-	-	-	-	-	-	-	-	-	-	-

Table A.5 Scenario I, $\rho = 0.9$, true rank = 10.

b		RID	NNP	RSC5	SUR5	MAX5	RSC10	SUR10	MAX10	RSC15	SUR15	MAX15	RSC20	SUR20	MAX20
0.3	Est	12.4	8.0	22.5	22.4	29.2	6.8	5.9	17.7	11.0	6.6	16.3	12.2	6.9	17.1
	Pred	25.1	36.5	73.3	73.0	126.9	16.9	15.6	96.1	23.0	17.5	101.3	24.9	18.1	104.4
	Time	-	209.6	-	108.6	273.4	-	524.2	322.9	-	805.0	362.0	-	883.6	421.7
	Rank	-	19.4	-	-	-	-	-	-	-	-	-	-	-	-
0.1	Est	12.2	3.8	5.4	4.3	4.8	8.2	3.8	4.4	11.1	4.0	4.2	12.1	4.0	4.2
	Pred	24.6	13.6	17.6	15.8	20.9	19.1	12.8	18.0	23.0	13.1	17.7	24.3	13.1	17.7
	Time	-	223.2	-	182.9	276.9	-	583.2	330.9	-	889.4	372.6	-	1049.5	424.7
	Rank	-	11.0	-	-	-	-	-	-	-	-	-	-	-	-
0.05	Est	12.6	1.6	4.9	1.6	1.7	9.3	1.6	1.6	11.6	1.7	1.6	12.5	1.7	1.6
	Pred	25.3	8.5	14.1	8.5	9.6	20.5	8.3	9.1	23.9	8.5	9.2	25.1	8.6	9.2
	Time	-	231.9	-	207.6	272.7	-	571.6	329.8	-	845.2	373.3	-	1020.1	426.5
	Rank	-	10.1	-	-	-	-	-	-	-	-	-	-	-	-

A.3 Correlation Models for Monitoring Fetal Growth

Table A.6 Scenario II, $\rho = 0.1$, true rank = 5.

b		RID	NNP	RSC3	SUR3	MAX3	RSC5	SUR5	MAX5	RSC8	SUR8	MAX8
0.3	Est	27.1	27.3	1786.3	31.6	32.7	2108.9	29.0	29.9	2124.1	27.7	29.8
	Pred	52.2	73.9	716.6	867.6	899.7	29.8	85.1	230.9	45.6	51.9	230.9
	Time	-	166.5	-	25.6	223.0	-	84.6	274.8	-	116.8	301.7
	Rank	-	8.4	-	-	-	-	-	-	-	-	-
0.1	Est	3.3	3.2	152.7	3.5	3.7	202.8	3.2	3.4	219.3	3.2	3.4
	Pred	48.4	39.9	97.2	113.2	118.0	32.1	35.2	47.8	47.0	38.8	57.3
	Time	-	162.1	-	60.4	269.6	-	183.6	300.3	-	208.2	321.5
	Rank	-	8.6	-	-	-	-	-	-	-	-	-
0.05	Est	0.9	0.9	40.7	0.9	1.0	56.7	0.9	0.9	65.6	0.9	0.9
	Pred	39.6	32.0	39.7	40.3	47.1	34.0	29.4	36.9	46.3	33.3	39.9
	Time	-	158.9	-	125.1	269.0	-	208.4	282.7	-	284.8	300.5
	Rank	-	8.0	-	-	-	-	-	-	-	-	-

Table A.7 Scenario II, $\rho = 0.9$, true rank = 5.

b		RID	NNP	RSC3	SUR3	MAX3	RSC5	SUR5	MAX5	RSC8	SUR8	MAX8
0.3	Est	26.4	27.4	6652.8	32.5	34.4	7396.8	28.6	31.0	7505.9	27.2	30.2
	Pred	48.3	84.9	251.7	285.8	431.3	30.1	37.6	228.1	45.4	41.1	242.6
	Time	-	256.7	-	56.8	289.9	-	169.0	327.1	-	206.3	352.4
	Rank	-	7.3	-	-	-	-	-	-	-	-	-
0.1	Est	3.4	3.4	826.8	3.6	3.8	938.1	3.3	3.6	1052.6	3.4	3.5
	Pred	38.0	34.6	42.5	45.1	61.9	33.0	29.5	49.2	45.1	32.6	53.5
	Time	-	260.9	-	163.4	314.8	-	260.2	332.2	-	338.4	361.9
	Rank	-	6.4	-	-	-	-	-	-	-	-	-
0.05	Est	1.0	1.0	111.5	1.0	1.0	146.6	1.0	1.0	186.4	1.0	1.0
	Pred	27.8	24.9	28.0	24.5	30.5	36.4	25.0	31.4	47.2	24.9	32.0
	Time	-	266.2	-	209.5	313.4	-	275.5	331.5	-	336.1	358.5
	Rank	-	5.1	-	-	-	-	-	-	-	-	-