INFERENCE OF HIGH-DIMENSIONAL LINEAR MODELS WITH TIME-VARYING COEFFICIENTS

BY

YIFENG HE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Xiaohui Chen, Chair
Professor Yuguo Chen
Professor Annie Qu
Professor Douglas Simpson

Abstract

In part 1, we propose a pointwise inference algorithm for high-dimensional linear models with time-varying coefficients and dependent error processes. The method is based on a novel combination of the nonparametric kernel smoothing technique and a Lasso bias-corrected ridge regression estimator using a bias-variance decomposition to address non-stationarity in the model. A hypothesis testing setup with familywise error control is presented alongside synthetic data and a real application to fMRI data for Parkinson's disease.

In part 2, we propose an algorithm for covariance and precision matrix estimation high-dimensional transpose-able data. The method is based on a Kronecker product approximation of the graphical lasso and the application of the alternating directions method of multipliers minimization. A simulation example is provided.

Table of Contents

Chapter 1

Introduction

Between 2005 and 2015, humanity watched its aggregate data multiply 10-fold every 2 years, a pattern driven by automated data collection and advances in technology such as the falling cost of digital storage. It has become cheaper and easier to gather everything than to consider what specific data to target and collect. We cast increasingly larger nets in search of the golden fish to explain some phenomena of interest. Online retailers log thousands of details about their customers to provide more personalized services. Genomics and fMRI machines produce massive parallel datasets in an effort to better understand patients and disease. Financial institutions collect vast arrays of data to produce better alpha for their investors. Over time, we have seen the fields of technology, medicine, finance, and others embrace the collection and analysis of large scale data.

Such an explosion has resulted in friction with traditional statistical theory, where it is typically assumed that one is working with a large number of samples but only a handful of well chosen, relevant variables. Since modern data collection relinquishes domain expertise in favor of a systematic approach by storing all available information about customers, patients, and subjects, we are left with a large number of variables rather than the handful from classical statistics. These numbers skyrocket even higher when we introduce functions of existing variables such as interaction effects or n-tuples. Therefore, there is great interest in the statistical analysis of problems where the number of variables or dimensions (p) in data exceeds its sample size (n), a domain referred to as *high-dimensional statistics.*

1.1 The Curse of Dimensionality

The primary obstacle in high-dimensional statistics is the curse of dimensionality, a term coined by Richard Bellman to describe the rapid growth of an optimization problem's complexity with dimensionality, originally in the context of exhaustive enumeration in a product space. [Bellman, 1961]

In nonparametric statistics, for example, we may have n i.i.d. vectors of predictor variables $X_1, ..., X_n \in$

$\mathbb{R}^p$ and the corresponding responses $Y_1, ..., Y_n \in \mathbb{R}$ given by

$$Y_i = f(X_i) + e_i$$

where f is L-Lipschitz and $e_i \sim$ i.i.d. $\mathcal{N}(0,1)$. Let $\mathcal{F}$ be the functional class of all Lipschitz functions on $[0,1]^p$ and $\hat{f}$ be some estimator of f using the observed responses. Minimax decision theory [Ibragimov and Khasminskii, 1981] shows that for any estimator $\hat{f}$, we have

$$\sup_{f \in \mathcal{F}} \mathbb{E}\left(\hat{f}(X) - f(X)\right)^2 \geq Cn^{-2/(2+p)} \quad , \quad n \to \infty$$

where C is a constant which depends on the Lipschitz constant L. This very slow rate of convergence in high-dimensions is the price of the curse of dimensionality. In order to estimate f to an accuracy of ϵ in the example above, we would need $\mathcal{O}\left(\epsilon^{-(2+p)/2}\right)$ samples. Therefore, the appetite for data grows exponentially with the number of variables p.

1.2 The Blessing of Dimensionality

The blessing of dimensionality is driven by the concentration of measure phenomenon. Generally speaking, the concentration of measures states that Lipschitz functions such as the average of bounded, independent random variables are concentrated around their expectations.

Take, for example, the concentration of Lipschitz functions of Gaussian random variables [Donoho, 2000]. Let $(X_1, ..., X_n)$ be a vector of i.i.d. Gaussian random variables and $f : \mathbb{R}^p \to \mathbb{R}$ is L-Lipschitz with respect to the l_2 metric. Then the random variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian such that

$$P\Big[|f(X) - \mathbb{E}[f(X)]| \geq \epsilon\Big] \leq 2\exp\left(-\frac{\epsilon^2}{2L^2}\right)$$

Therefore, a Lipschitz function of standard Gaussian random variables is nearly constant and the tails, at worst, behave like a scalar Gaussian variable with variance L^2, regardless of dimension.

Similar tools exist for the concentrations of product measures with respect to l_1 and Hamming metrics, uniform measures over the unit sphere surface with respect to l_2, and others. These concentrations of measures are heavily used to establish theoretical results in high-dimensional statistics, most commonly for the bounding of error probabilities.

1.3 High-Dimensional Linear Regression

In the traditional linear regression model, we are given predictor measurements $X_1, ..., X_n \in \mathbb{R}^p$ and the responses $Y_1, ...Y_n \in \mathbb{R}$ to be modeled by

$$Y = \mathbf{X}\beta + e$$

where $\mathbf{X} = (X_1^\top, ..., X_n^\top)^\top \in \mathbb{R}^{n \times p}$ is the design matrix, $Y \in \mathbb{R}^n$ the response vector, $e \in \mathbb{R}^n$ the noise or error vector, and $\beta \in \mathbb{R}^p$ the coefficient vector to be estimated. Typically, e is assumed to be independent of $\mathbf{X}$ with mean zero.

To claim that the ordinary least squares (OLS) estimator has been well studied in low-dimensions would be an understatement, but in the case where $n > p$, the OLS estimator can be written as

$$\hat{\beta}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y$$

This is well defined where the columns of $\mathbf{X}$ are linearly independent but this assumption clearly cannot hold when $p > n$ such that $\mathbf{X}^\top \mathbf{X}$ is singular. Therefore the parameter estimation problem is ill-posed in high-dimensions.

Furthermore in the simple case where e_i are i.i.d. with variance σ^2, the OLS linear regression estimator has expected prediction error $\sigma^2 + p\sigma^2/n$ with zero bias. Therefore, adding an additional variable X_{p+1} contributes an extra σ^2/n to the variance, regardless if said variable is linearly relevant (i.e. $\beta_{p+1} = 0$). We can improve on the mean squared error (MSE) of our estimator by reducing the variance at the expense of introducing bias, such as though shrinkage where coefficients are coerced towards zero. This is of particular note in high-dimensions; While models may involve a large number of parameters such that $p > n$, there often exists an underlying low-dimensional structure such as sparsity or smoothness which makes inference a more realistic goal.

1.3.1 Ridge Regression

Regularization is a common tool used to address the issues of regression in high-dimensions which balances a loss function such as squared error in the context of linear regression with regularization or penalization to promote some underlying structure. Ridge regression [Hoerl and Kennard, 1970] is one of the simplest forms of regularization for linear models, and its coefficients $\hat{\beta}_{Ridge} \in \mathbb{R}^p$ are defined by standardizing the columns of $\mathbf{X}$ and optimizing the following objective function

$$\hat{\beta}_{Ridge} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2 \tag{1.1}$$

where $\lambda \geq 0$ is a tuning parameter which controls the strength of shrinkage and the bias-variance tradeoff. The case of $\lambda = 0$ corresponds to the ordinary linear regression estimator where we have no bias but large variance. The case of $\lambda = \infty$ shrinks all coefficients to $\hat{\beta}_{Ridge} = 0$, which has high bias but zero variance. In the simple case where $\mathbf{X}$ is orthonormal, the shrinkage as a function of λ is given by

$$\hat{\beta}_{Ridge} = \frac{\hat{\beta}_{OLS}}{1+\lambda} \tag{1.2}$$

Furthermore, the existence property of the Ridge estimator guarantees the existence of some $\lambda > 0$ such that the MSE of $\hat{\beta}_{Ridge}$ is less than the MSE of $\hat{\beta}_{OLS}$. However, finding or choosing such an appropriate λ is a somewhat contested topic, with cross-validation currently the most standard compromise.

A convenient property of the Ridge estimator is the differentiability of its objective function. As such, it's easy to obtain a closed form solution within the rowspace $\mathcal{R}(\mathbf{X})$.

$$\hat{\beta}_{Ridge} = (\mathbf{X}^\top\mathbf{X}) + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top Y \tag{1.3}$$

$$\text{Bias}(\hat{\beta}_{Ridge}) = -\lambda(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\beta \tag{1.4}$$

$$\text{Var}(\hat{\beta}_{Ridge}) = \sigma^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1} \tag{1.5}$$

Therefore, the Ridge estimator $\hat{\beta}_{Ridge}$ is asymptotically biased and not a consistent estimator of β. While its distribution can be somewhat characterized, its bias still poses a challenge to performing statistical inference.

1.3.2 Lasso and Variable Selection

The Lasso [Tibshirani, 1996] also solves a similar penalized least squares regression problem using an l_1 penalty, commonly with some assumption of sparsity.

$$\hat{\beta}_{Lasso} = \arg\min_{\beta\in\mathbb{R}[} \frac{1}{2}\|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^1 \tag{1.6}$$

Unlike the Ridge regression, the Lasso's objective function is not differentiable and its solution does not necessarily have a closed form expression. However, there are many fast algorithms for obtaining the Lasso coefficients, and the effect of the regularization is essentially a soft-thresholding function.

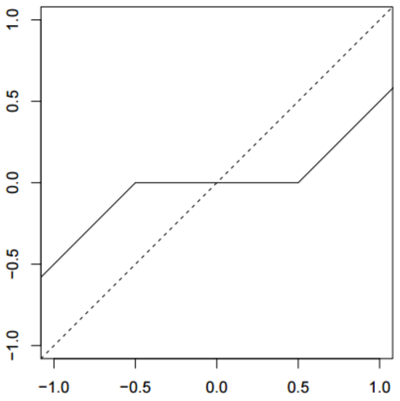


Figure 1.1: A soft-thresholding function

Due to the nonlinear nature of the Lasso, its solution does not exist in the rowspace of $\mathbf{X}$ and its distribution is difficult to characterize. Since the soft-thresholding effect coerces some coefficients to zero, the Lasso effectively performs variable selection.

Let $S_0 = \{j : \beta_j \neq 0, j = 1, ..., p\}$ denote the active set and $\hat{S}_0$ the estimated active set using a variable selection technique such as the Lasso. The said method demonstrates model selection consistency if

$$P(\hat{S}_0 = S_0) \to 1 \quad p > n \to \infty \tag{1.7}$$

In the case of Lasso, model selection consistency requires a very strong set of assumptions on the design matrix referred to as the irrepresentable condition and the beta-min condition. Roughly speaking, the irrepresentable condition requires the columns of the design matrix $\mathbf{X}$ corresponding to the active set S_0 to be nearly orthogonal to columns of $\mathbf{X}$ corresponding to the inactive set $S_0^{\complement}$. As the name implies, the beta-min condition imposes a nonzero lower bound to $|\beta_j|, j \in S_0$. We refer to [Meinshausen and Bühlman, 2006] and [Zhao and Yu, 2006] for a more detailed study of these assumptions. The irrepresentable condition can be relaxed to a restricted eigenvalue condition to obtain model screening consistency such that

$$P(\hat{S}_0 \supseteq S_0) \to 1 \quad p > n \to \infty \tag{1.8}$$

although the undesirable beta-min condition remains.

Furthermore, the Lasso yields an ℓ_2 consistent estimator [Meinshausen and Yu, 2009] under milder assumptions with suitable $\lambda \asymp \sqrt{\log(p)/n}$, even when the variable selection assumptions are violated and the sparsity patterns cannot be recovered.

1.4 Statistical Inference in High-Dimensions

High-dimensional statistical theory has historically focused heavily on the estimation consistency, prediction consistency, oracle inequalities, and variable selection. The assessment of uncertainty such as assigning p-values in high-dimensional settings has remained an relatively poorly understood topic until only recently.

The primary measure of uncertainty discussed here will be the p-value in the context of hypothesis testing for linear models. That is, for each $j \in 1, ..., p$, we are interested in testing

$$\begin{aligned} H_{0,j}: &\quad \beta_j = 0 \\ H_{a,j}: &\quad \beta_j \neq 0 \end{aligned}$$

We present several methods of assigning significance in high-dimensional linear models from current literature.

1.4.1 Multi-Sample Splitting

Multi-sample splitting or selective inference [Meinshausen et al., 2009] is a general two-step method of performing inference in high dimensions using a dimension reduction step and an inference in low-dimensions step.

Given a high-dimensional dataset with n observations, multi sample splitting first partitions the indices $\{1, ..., n\}$ into two parts denoted I_1 and I_2 with $|I_1| = \lfloor n/2 \rfloor$, $|I_2| = \lceil n/2 \rceil$, $I_1 \cup I_2 = \{1, ..., n\}$, and $I_1 \cap I_2 = \varnothing$. Let Y_{I_i} and $\mathbf{X}_{I_i}$) denote, respectively, the elements of Y and the rows of $\mathbf{X}$ corresponding to the I_i index, $i = 1, 2$. A variable screening procedure is applied on $(Y_{I_1}, \mathbf{X}_{I_1})$ to obtain $\hat{S}_{0,I_1}$ with cardinality $|\hat{S}_{0,I_1}| \leq |I_2|$. The Lasso is a common candidate for variable screening and implicitly satisfies the cardinality condition under weak assumptions. Indeed the choice of using roughly evenly sized partitions $|I_1| \leq |I_2|$ appears to cater especially to the Lasso. While capable of screening consistency under stronger assumptions, a screening property is not always necessary for the construction of valid p-values [Bühlmann and Mandozzi, 2014].

Let $\mathbf{X}_{I_2,\hat{S}_{0,I_1}}$ denote the columns of $\mathbf{X}_{I_2}$ corresponding to the variables screened by $\hat{S}_{0,I_1}$. The second step of the procedure simply regresses Y_{I_2} onto $\mathbf{X}_{I_2,\hat{S}_{0,I_1}}$, which is readily solved by ordinary least squares regression and produces the desired p-values.

One glaring issue concerning the two-step selection and inference procedure is that the constructed p-values are very sensitive to the choice of I_1, I_2, affectionately referred to by its authors as the p-value lottery. The problem is somewhat addressed through repeated randomization in the sampling the partitions and aggregating the constructed p-values. Overall, the multi sample splitting approach is a generalized method with certain assumptions on variable screening and selection cardinality. However, the need for beta-min or zonal conditions for most screening methods runs counter to the purpose of a significance test. Additionally, its repeated sampling of partitions does not lend itself well to cases where complex dependencies exist in noise, such as spatio-temporal data.

1.4.2 Lasso Projection

The de-sparsified Lasso or low-dimension projection estimator (LPDE), proposed by [Zhang and Zhang, 2014] and [Dezeure et al., 2015] corrects the bias from Lasso using a projection onto a low-dimensional orthogonal space constructed using nodewise regression.

Let $\mathbf{X}_j \in \mathbb{R}^n$ denote the column of $\mathbf{X}$ corresponding to variable $j = 1, ..., p$ and $\mathbf{X}_{-j} \in \mathbb{R}^{n\times(p-1)}$ denote the other columns. Let $Z_j \in \mathbb{R}^n$ denote the residuals of the Lasso regression of $\mathbf{X}_j$ onto $\mathbf{X}_{-j}$. Then, for any Z_j,

$$\frac{Y^\top Z_j}{\mathbf{X}_j^\top z_j} = \beta_j + \sum_{k\neq j} \frac{\mathbf{X}_k^\top Z_j}{\mathbf{X}_j Z_j} \beta_k + \frac{e^\top Z_j}{\mathbf{X}_j^\top Z_j} \tag{1.9}$$

Therefore, the bias corrected plugin estimator using an initial Lasso estimate of β is

$$\hat{\beta}_{j,LPDE} = \frac{Y^\top Z_j}{\mathbf{X}_j^\top z_j} - \sum_{k\neq j} \frac{\mathbf{X}_k^\top Z_j}{\mathbf{X}_j Z_j} \hat{\beta}_{k,Lasso} \tag{1.10}$$

The LPDE estimator admits a Gaussian distribution under compatibility conditions on $\mathbf{X}$, when the sparsity is $s_0 = o\left(\sqrt{n}/\log(p)\right)$, the rows of $\mathbf{X}$ are i.i.d. $\mathcal{N}(0, \Sigma)$ with a positive lower bound on the eigenvalues of Σ, and Σ^{-1} is row-sparse. Aside from some somewhat undesirable assumptions on the fixed design, the LPDE estimator is very computationally demanding, roughly 2 orders of magnitude over multi-sample splitting and Ridge projection. On the other hand, it does not make any assumptions on the underlying β coefficients besides sparsity and achieves the Cramer-Rao efficiency bound, which the Ridge projection does not.

Chapter 2

Time-Varying High-Dimensional Linear Models

2.1 Introduction

We consider the following time-varying coefficient models (TVCM)

$$y(t) = \mathbf{x}(t)^\top \boldsymbol{\beta}(t) + e(t) \tag{2.1}$$

where $t \in [0, 1]$ is the time index, $y(\cdot)$ the response process, $\mathbf{x}(\cdot)$ the $p \times 1$ deterministic predictor process (i.e. fixed design), $\boldsymbol{\beta}(\cdot)$ the $p \times 1$ time varying coefficient vector, and $e(\cdot)$ the mean zero stationary error process. The response and predictors are observed at evenly spaced time points $t_i = i/n, i = 1, ..., n$, i.e. $y_i = y(t_i), \mathbf{x}_i = \mathbf{x}(t_i)$ and $e_i = e(t_i)$ with known covariance matrix $\Sigma_e = \text{Cov}(\mathbf{e})$ where $\mathbf{e} = (e_1, \cdots, e_n)^\top$. TVCM is useful for capturing the dynamic associations in the regression models and longitudinal data analysis [Hoover et al., 1998] and it has broad applications in biomedical engineering, environmental science and econometrics. In this work, we shall consider the fixed design case and focus on the *pointwise* inference for the time-varying coefficient vector $\boldsymbol{\beta}(t)$ in the high-dimensional double asymptotics framework $\min(p, n) \to \infty$. Moreover, different from longitudinal setting, we consider only observations from one subject $(\mathbf{x}_i, y_i)$.

Nonparametric estimation and inference of the TVCM in the fixed dimension has been extensively studied in literature, e.g. see [Robinson, 1989, Hoover et al., 1998, Fan and Wenyang, 1999, Zhang et al., 2002, Orbe et al., 2005, Zhang and Wu, 2012, Zhou and Wu, 2010]. In the high-dimensional setting, variable selection and estimation of varying-coefficient models using basis expansions have been studied in [Wei and Huang, 2010] and [Wang et al., 2014]. Nevertheless, our primary objective is not to estimate $\boldsymbol{\beta}(t)$, but rather to perform the statistical inference on the coefficients. In particular, for any $t \in (0, 1)$, we wish to test the following local hypothesis

$$H_{0,j,t} : \beta_j(t) = 0 \quad \text{VS} \quad H_{1,j,t} : \beta_j(t) \neq 0, \qquad \forall j = 1, \cdots, p. \tag{2.2}$$

By assigning p-value at each time point, our goal is to construct a sequence of coefficient vectors that allows us to assess the uncertainty of the dynamic patterns such as modeling the brain connectivity networks. Confidence intervals and hypothesis testing problems of lower-dimensional functionals of the high-dimensional constant coefficient vector $\boldsymbol{\beta}(t) \equiv \boldsymbol{\beta}, \forall t \in [0,1]$, have been studied in [Bühlmann, 2013], [Zhang and Zhang, 2014], and [Javanmard and Montanari, 2014]. To our best knowledge, little has been done for inference of high-dimensional TVCM. Therefore, our goal is to fill the inference gap between the classical TVCM and high-dimensional linear model.

As another key difference of this thesis from the existing inference literature on high-dimensional linear models based on the fundamental i.i.d. error assumption [Bühlmann, 2013], [Zhang and Zhang, 2014], [Javanmard and Montanari, 2014], the second main contribution is to provide an asymptotic theory for answering the following question: To what extent can the statistical validity of a proposed inference procedure for i.i.d. errors hold for dependent error processes? Allowing temporal dependence is of practical interest since many datasets such as fMRI data are spatio-temporal in nature and the errors are correlated in the time domain. On the other hand, theoretical analyses reveal that the temporal dependence has delicate impact on the asymptotic rates for estimating the covariance structures [Chen et al., 2013]. Therefore, it is more plausible to build an inference procedure that is also robust in the time series context. The error process e_i is modelled as a stationary linear process

$$e_i = \sum_{m=0}^{\infty} a_m \xi_{i-m}, \tag{2.3}$$

where $a_0 = 1$ and ξ_i are i.i.d. mean-zero random variables (a.k.a. innovations) with variance σ^2. When ξ_i have normal distributions, linear processes of form (2.3) is a Gaussian processes that covers the auto-regressive and moving-average (ARMA) models with i.i.d. Gaussian innovations as special cases. For the linear process, we shall deal with both weak and strong temporal dependencies. In particular, if $a_m = O(m^{-\varrho}), \varrho > 1/2$, then e_i is well-defined and it has: (i) short-range dependence (SRD) if $\varrho > 1$; (ii) long-range dependence (LRD) or long-memory if $1/2 < \varrho < 1$. For the SRD processes, it is clear that $\sum_{m=0}^{\infty} |a_m| < \infty$ and therefore the long-run variance is finite.

The chapter is organized as follows. In Section 2.2, we describe our method in details. Asymptotic theory is presented in Section 2.3. In Section 2.4, we mention some practical implementation issues and extension of the noise model to more general non-Gaussian and heavy-tailed distributions. Section 2.5 presents some simulation results and Section 2.6 demonstrates a real application to an fMRI dataset. Proofs are available in the Appendix.

2.2 Method

2.2.1 Notations and Preliminary Intuition

Let K be a non-negative symmetric function with bounded support in $[-1,1], \int_{-1}^{1} K(x)dx = 1$ and b_n be a bandwidth parameter satisfies the natural condition $b_n = o(1)$ and $n^{-1} = o(b_n)$. For each time point $t \in \varpi = [b_n, 1-b_n]$, the Nadaraya-Watson smoothing weight is defined as

$$w(i,t) = \begin{cases} \frac{K_{b_n}(t_i - t)}{\sum_{m=1}^{n} K_{b_n}(t_m - t)} & \text{if } |t_i - t| \leq b_n \\ 0 & \text{otherwise} \end{cases}, \tag{2.4}$$

where $K_b(\cdot) = K(\cdot/b)$. Let $N_t = \{i : |t_i - t| \leq b_n\}$ be the b_n-neighborhood of time t, $|N_t|$ be the cardinality of the discrete set N_t, $W_t = \text{diag}(w(i,t)_{i\in N_t})$ be the $|N_t| \times |N_t|$ diagonal matrix with $w(i,t), i \in N_t$ on the diagonal, and $\mathcal{R}_t = \text{span}(\mathbf{x}_i : i \in N_t)$ be the subspace in $\mathbb{R}^p$ spanned by $\mathbf{x}_i$, the rows of design matrix X in the N_t neighborhood. Let $\mathcal{X}_t = (w(t,i)^{1/2}\mathbf{x}_i)_{i\in N_t}^\top$, $\mathcal{Y}_t = (w(i,t)^{1/2}y_i)_{i\in N_t(i)}^\top$ and $\mathcal{E}_t = (w(i,t)^{1/2}e_i)_{i\in N_t(i)}^\top$. Then, the singular value decomposition (SVD) of $\mathcal{X}_t$ is

$$\mathcal{X}_t = PDQ^\top \tag{2.5}$$

where P and Q are $|N_t| \times r$, and $p \times r$ matrices such that $P^\top P = Q^\top Q = I_r$. $D = \text{diag}(d_1, \cdots, d_r)$ is a diagonal matrix containing the r nonzero singular values of $\mathcal{X}_t$. Now let $P_{\mathcal{R}_t}$ be the projection matrix onto $\mathcal{R}_t$. Then,

$$P_{\mathcal{R}_t} = \mathcal{X}_t^\top(\mathcal{X}_t\mathcal{X}_t^\top)^- \mathcal{X}_t = QQ^\top \tag{2.6}$$

where $(\mathcal{X}_t\mathcal{X}_t^\top)^- = PD^{-2}P^\top$ is the pseudo-inverse matrix of $\mathcal{X}_t\mathcal{X}_t^\top$. Let $\boldsymbol{\theta}(t) = P_{\mathcal{R}_t}\boldsymbol{\beta}(t)$ be the projection of $\boldsymbol{\beta}(t)$ onto $\mathcal{R}_t$, the row subspace of $\mathcal{X}_t$ such that $B(t) = \boldsymbol{\theta}(t) - \boldsymbol{\beta}(t)$ is the projection bias. Let

$$\Omega(\lambda) = (\mathcal{X}_t^\top\mathcal{X}_t + \lambda I_p)^{-1}\mathcal{X}_t^\top W_t^{1/2}\Sigma_{e,t}W_t^{1/2}\mathcal{X}_t(\mathcal{X}_t^\top\mathcal{X}_t + \lambda I_p)^{-1}, \tag{2.7}$$

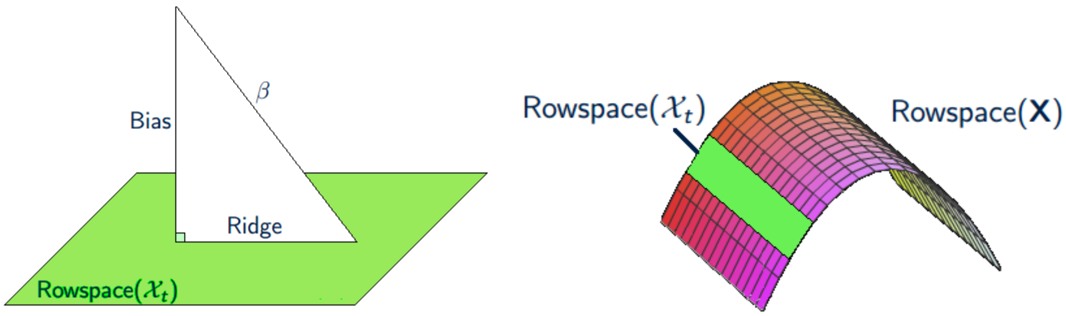
where $\Sigma_{e,t} = \text{Cov}((e_i)_{i\in N_t(i)})$, and $\Omega_{\min}(\lambda) = \min_{j\leq p}\Omega_{jj}(\lambda)$ be the smallest diagonal entry of $\Omega(\lambda)$. For a generic vector $\mathbf{b} \in \mathbb{R}^p$, we denote $|\mathbf{b}|_q = (\sum_{j=1}^{p}|b_j|^q)^{1/q}$ if $q > 0$, and $|\mathbf{b}|_0 = \sum_{j=1}^{p}\mathbf{1}(b_j \neq 0)$ if $q = 0$. Denote $\underline{w}_t = \inf_{i\in N_t} w(i,t)$ and $\overline{w}_t = \sup_{i\in N_t} w(i,t)$. For an $n \times n$ square symmetric matrix M and an $n \times m$ rectangle matrix R, we use $\rho_i(M)$ and $\sigma_i(R)$ to denote the i-th largest eigenvalues of M and singular values of R, respectively. If $k = \text{rank}(R)$, then $\sigma_1(R) \geq \sigma_2(R) \geq \cdots \geq \sigma_k(R) > 0 = \sigma_{k+1}(R) = \cdots =$

$\sigma_{\max(m,n)}(R)$, i.e. zeros are padded to the last $\max(m,n)-k$ singular values. We denote $\rho_{\max}(M)$, $\rho_{\min}(M)$ and $\rho_{\min\neq 0}(M)$ be the maximum, minimum and nonzero minimum eigenvalues of M, respectively, and $|M|_\infty = \max_{1\le j,k\le p}|M_{jk}|$. Let

$$\rho_{\max}(M,s) = \max_{|\mathbf{b}|_0\le s, \mathbf{b}\neq \mathbf{0}} \frac{|\mathbf{b}^\top M\mathbf{b}|_2}{|\mathbf{b}|_2^2}.$$

If M is non-negative definite, then $\rho_{\max}(M,s)$ is the restricted maximum eigenvalues of M at most s columns and rows.

Before proceeding, we pause to explain the intuition behind our method. The p-dimensional coefficient vector $\boldsymbol{\beta}(t)$ is decomposed into two parts via projecting onto the $|N_t|$-dimensional linear subspace spanned by the rows of $\mathcal{X}_t$ and its orthogonal complement; see Figure 2.1(a). A key advantage of this decomposition is that the projected part can be conveniently estimated in the closed-form for example by ridge estimator since it lies in the row space of $\mathcal{X}_t$ and thus amenable for the subsequent inferential analysis. In the high-dimensional situation, this projection introduces a non-negligible shrinkage bias in estimating $\boldsymbol{\beta}(t)$ and therefore we may lose information because $p \gg |N_t|$. On the other hand, the shrinkage bias can be corrected by a consistent estimator of $\boldsymbol{\beta}(t)$. As a particular example, we shall use the Lasso estimator. However, any sparsity-promoting estimator attaining the same convergence rate as the Lasso should work. Because of the time-varying nature of the nonzero functional $\boldsymbol{\beta}(t)$, the smoothness on the row space of $\mathcal{X}_t$ along the time index t is necessary to apply nonparametric smoothing technique; see Fig. 2.1(b). As a special case when the nonzero components $\boldsymbol{\beta}(t) \equiv \boldsymbol{\beta}$ are constant functions and the error process is i.i.d. Gaussian, our algorithm is the same as [Bühlmann, 2013]. However, the emphases of this work are: (i) time-varying (i.e. non-constant) coefficient vectors; (ii) the errors are allowed to have heavy-tails by assuming milder polynomial moment conditions and to have temporal dependence, including both SRD and LRD processes. As mentioned earlier, there are other inferential methods for high-dimensional linear models such as [Zhang and Zhang, 2014] and [Javanmard and Montanari, 2014]. We do not explore specific choices here since the contribution is a general framework of combining nonparametric smoothing and bias-correction methods to make inference for high-dimensional TVCM. However, we expect that the non-stationary generalization would be feasible for those methods as well.



(a) Bias correction by projection to the row space of $\mathcal{X}_t$.

(b) Smoothly time-varying row space of $\mathcal{X}_t$.

Figure 2.1: Intuition of the proposed algorithm in Section 2.2.2.

2.2.2 Inference Algorithm

First, we estimate the projection bias $B(t)$ by $\tilde{B}(t) = (P_{\mathcal{R}_t} - I_p)\tilde{\boldsymbol{\beta}}(t)$, where $\tilde{\boldsymbol{\beta}}(t)$ is the time-varying Lasso (tv-Lasso) estimator

$$\begin{aligned} \tilde{\boldsymbol{\beta}}(t) &= \arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i\in N_t} w(i,t)(y_i - \mathbf{x}_i^\top \mathbf{b})^2 + \lambda_1 |\mathbf{b}|_1 \qquad (2.8) \\ &= \arg\min_{\mathbf{b}\in\mathbb{R}^p} |\mathcal{Y}_t - \mathcal{X}_t\mathbf{b}|_2^2 + \lambda_1|\mathbf{b}|_1. \end{aligned}$$

Next, we estimate $\boldsymbol{\theta}(t) = P_{\mathcal{R}_t}\boldsymbol{\beta}(t)$ using the time-varying ridge (tv-ridge) estimator

$$\begin{aligned} \tilde{\boldsymbol{\theta}}(t) &= \arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i\in N_t} w(i,t)(y_i - \mathbf{x}_i^\top \mathbf{b})^2 + \lambda_2|\mathbf{b}|_2^2 \\ &= (\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top \mathcal{Y}_t. \qquad (2.9) \end{aligned}$$

We shall defer the discussion of tuning parameters choice λ_1 and λ_2 in Section 2.3 and 2.4. Then, our tv-Lasso bias-corrected tv-ridge regression estimator for $\boldsymbol{\beta}(t)$ is constructed as

$$\hat{\boldsymbol{\beta}}(t) = \tilde{\boldsymbol{\theta}}(t) - \tilde{B}(t). \qquad (2.10)$$

Now, based on $\hat{\boldsymbol{\beta}}(t) = (\hat{\beta}_1(t), \cdots, \hat{\beta}_p(t))^\top$, we calculate the raw two-sided p-values for individual coefficients

$$\tilde{P}_j = 2\left[1 - \Phi\left(\frac{|\hat{\beta}_j(t)| - \lambda_1^{1-\xi}\max_{k\neq j}|(P_{\mathcal{R}_t})_{jk}|}{\Omega_{jj}^{1/2}(\lambda_2)}\right)\right], \qquad j = 1, \cdots, p, \qquad (2.11)$$

where $\xi \in [0,1)$ is user pre-specified number, which depends on the number of nonzero $\boldsymbol{\beta}(t)$, i.e. sparsity. In particular, if $|\text{supp}(\boldsymbol{\beta}(t))|$ is bounded, then we can choose $\xi = 0$. Generally, following [Bühlmann, 2013], we use $\xi = 0.05$ in our numeric examples to allow the number of nonzero components in $\boldsymbol{\beta}(t)$ diverges at proper rates. Let $\mathbf{v}(t) = (V_1(t), \cdots, V_p(t))^\top \sim N(\mathbf{0}, \Omega(\lambda_2))$ and define the distribution function

$$F(z) = \mathbb{P}\left(\min_{j \le p} 2\left[1 - \Phi\left(\Omega_{jj}^{-1/2}(\lambda_2)|V_j(t)|\right)\right] \le z\right). \tag{2.12}$$

We adjust the $\tilde{P}_j$ for multiplicity by $P_j = F(\tilde{P}_j + \zeta)$, where ζ is another predefined small number [Bühlmann, 2013] that accommodates asymptotic approximation errors. Finally, our decision rule is defined as: Reject $H_{0,j,t}$ if $P_j \le \alpha$ for $\alpha \in (0,1)$. For i.i.d. errors, since $\Sigma_e = \sigma^2 \text{Id}_n$ and

$$\Omega(\lambda_2) = \sigma^2 (\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1} \mathcal{X}_t^\top W_t \mathcal{X}_t (\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1},$$

we see that $F(\cdot)$ is independent of σ. Therefore, $F(\cdot)$ can be easily estimated by repeatedly sampling from the multivariate Gaussian distribution $N(\mathbf{0}, \Omega(\lambda_2))$. Similar observations have been pointed out in [Bühlmann, 2013].

2.3 Asymptotic Results

In this section, we present the asymptotic theory of the inference algorithm in Section 2.2.2. First, we state the main assumptions for i.i.d. Gaussian errors.

1. **Error.** The errors $e_i \sim N(0, \sigma^2)$ are independent and identically distributed (i.i.d.).

2. **Sparsity.** $\boldsymbol{\beta}(\cdot)$ is uniformly s-sparse, i.e. $\sup_{t \in [0,1]} |S_{0t}| \le s$, where $S_{0t} = \{j : \beta_j(t) \neq 0\}$ is the support set.

3. **Smoothness.**

   (a) $\boldsymbol{\beta}(\cdot)$ is twice differentiable with bounded and continuous first and second derivatives in the coordinate-wise sense, i.e. $\beta_j(\cdot) \in \mathcal{C}^2([0,1], C_0)$ for each $j = 1, \cdots, p$ and C_0 is an upper bound for the partial derivatives.

   (b) The b_n-neighborhood covariance matrix $\hat{\Sigma}_t^\diamond = |N_t|^{-1} \sum_{i \in N_t} \mathbf{x}_i \mathbf{x}_i^\top := \mathcal{X}_t^{\diamond\top} \mathcal{X}_t^\diamond$ satisfies the following conditions:

$$\rho_{\max}(\hat{\Sigma}_t^\diamond, s) \le \varepsilon_0^{-2} < \infty. \tag{2.13}$$

4. **Non-degeneracy.**

$$\liminf_{\lambda \downarrow 0} \Omega_{\min}(\lambda) > 0. \tag{2.14}$$

5. **Identifiability.**

   (a) The minimum nonzero eigenvalue condition

   $$\rho_{\min \neq 0}(\hat{\Sigma}_t^{\diamond}) \geq \varepsilon_0^2 > 0. \tag{2.15}$$

   (b) The *restricted eigenvalue condition* is met:

   $$\phi_0 = \inf \left\{ \phi > 0 : \min_{|S|=s} \inf_{|\mathbf{b}_{S^c}|_1 \leq 3|\mathbf{b}_S|_1} \frac{\mathbf{b}^\top \hat{\Sigma}_t \mathbf{b}}{|\mathbf{b}_S|_1^2} \geq \frac{\phi^2}{s} \text{ holds for all } t \in [0,1] \right\} > 0, \tag{2.16}$$

   where $\hat{\Sigma}_t = \mathcal{X}_t^\top \mathcal{X}_t$ is the kernel smoothed covariance matrix of the predictors.

6. **Kernel.** The kernel function $K(\cdot)$ is non-negative, symmetric around 0 with bounded support in $[-1, 1]$.

Here, we comment the assumptions and their implications. Assumption 1 and 6 are standard. The Gaussian distribution is non-essential and it can be relaxed to sub-Gaussian and heavier tailed distributions; see Section 2.4 for more discussions. Assumption 2 is a sparsity condition for the nonzero functional components and we allow $s \to \infty$ slower than $\min(p, n)$. It is a key condition for maintaining the low-dimensional structure when the dimension p grows fast with the sample size n. In addition, by the argument of proving [Zhou et al., 2010, Theorem 5], it also implies that the number of nonzero first and second derivatives of $\boldsymbol{\beta}(t)$ is bounded by s almost surely on $[0, 1]$. Assumption 3 ensures the smoothness of the time-varying coefficient vector and the design matrix so that nonparametric smoothing techniques are applicable. Examples of assumption 3(a) include the quadratic form $\boldsymbol{\beta}(t) = \boldsymbol{\beta} + \boldsymbol{\alpha} t + \boldsymbol{\xi} t^2/2$ and the periodic functions $\boldsymbol{\beta}(t) = \boldsymbol{\beta} + \boldsymbol{\alpha} \sin(t) + \boldsymbol{\xi} \cos(t)$ with $|\boldsymbol{\alpha}|_\infty + |\boldsymbol{\xi}|_\infty \leq C_0$. Assumption 3(b) can be viewed as the Lipschitz continuity on the local design matrix that is smoothly evolving [Zhou and Wu, 2010]. However, it is weaker than the condition that $\rho_{\max}(\hat{\Sigma}_t^{\diamond}) \leq \varepsilon_0^{-2}$ because the latter may grow to infinity much faster than the restricted form (2.13).

Assumption 4 is required for non-degenerated stochastic component of the proposed estimator which is used for the inference purpose. Assumption 5(a) and 5(b), i.e. (2.15) and (2.16), together impose the identifiability conditions for recovering the coefficient vectors. Analogous condition of the time-invariant version have been extensively used in literature to derive theoretical properties of the Lasso model; see e.g. [Bickel et al., 2009][van de Geer and Bühlmann, 2009].

Now, for the tv-lasso bias-corrected tv-ridge estimator (2.17), we establish a representation that is fundamental for the subsequent statistical inference purpose.

Theorem 2.1 (Representation)**.** *Fix $t \in \varpi$ and let*

$$L_{t,\ell} = \max_{j \le p} \left[\sum_{i \in N_t} w(t,i)^{\ell} X_{ij}^2 \right]^{1/2}, \ \ell = 1, 2, \cdots, \qquad \lambda_0 = 4\sigma L_{t,2}\sqrt{\log p}, \tag{2.17}$$

and $\lambda_1 \ge 2(\lambda_0 + 2C_0 L_{t,1} b_n (s|N_t|\overline{w}_t)^{1/2} \varepsilon_0^{-1})$. Under assumptions 1-6 and $C \le |N_t|\underline{w}_t \le |N_t|\overline{w}_t \le C^{-1}$ for some $C \in (0,1)$, our estimator $\hat{\boldsymbol{\beta}}(t)$ in (2.10) admits the following decomposition

$$\begin{aligned} \hat{\boldsymbol{\beta}}(t) &= \boldsymbol{\beta}(t) + \mathbf{z}(t) + \boldsymbol{\gamma}(t), & (2.18) \\ \mathbf{z}(t) &\sim N(\mathbf{0}, \Omega(\lambda_2)), & (2.19) \\ |\gamma_j(t)| &\le \frac{\lambda_2 |\boldsymbol{\theta}(t)|_2 + 2C_0 s^{1/2} b_n}{C\varepsilon_0^2} + \frac{4\lambda_1 s}{\phi_0^2} |P_{\mathcal{R}_t} - Id|_\infty, \quad j = 1, \cdots, p, & (2.20) \end{aligned}$$

with probability at least $1 - 2p^{-1}$. In addition, if $\beta_j(t) = 0$, then we have

$$\Omega_{jj}^{-1/2}(\lambda_2)(\hat{\beta}_j(t) - \gamma_j(t)) \overset{d}{=} N(0,1), \tag{2.21}$$

where

$$|\gamma_j(t)| \le \frac{\lambda_2 |\boldsymbol{\theta}(t)|_2 + 2C_0 s^{1/2} b_n}{C\varepsilon_0^2} + \frac{4\lambda_1 s}{\phi_0^2} \max_{k \ne j} |(P_{\mathcal{R}_t})_{jk}|. \tag{2.22}$$

Remark 1. *Our decomposition (2.18) can be viewed as a local version of the one proposed in [Bühlmann, 2013, Proposition 2]. However, due to the time-varying nature of the nonzero coefficient vectors, both the stochastic component $\mathbf{z}(t)$ in (2.19) and the bias component $\boldsymbol{\gamma}(t)$ in (2.20) of the representation (2.18) have a number of key differences from [Bühlmann, 2013]. First, our bound (2.20) for bias has three terms, arising from ridge shrinkage bias, non-stationary bias, and Lasso correction bias. All three sources of bias have localized features, depending on the bandwidth of the sliding window b_n and the smoothness parameter C_0. Second, the stochastic part (2.19) also has time-dependent features in the second-order moment ($\Omega(\lambda_2)$ implicitly depends on t though $\mathcal{X}_t$) and the scale of normal random vector is different from [Bühlmann, 2013]. Delicate balance among them allows us to perform valid statistical inference such as hypothesis testing and confidence interval construction for the coefficients and, more broadly, their lower-dimensional linear functionals.*

Example 2.1. *Consider the uniform kernel $K(x) = 0.5\mathbb{I}(|x| \le 1)$ as an important special case, which is the kernel used for our numeric experiments in Section 2.5. In this case, $\underline{w}_t = (2nb_n)^{-1}$ and $|N_t|\underline{w}_t = |N_t|\overline{w}_t = 1$. It is easily verified that conditions of Theorem 2.1 are satisfied and, under the local null hypothesis $H_{0,j,t}$, (2.22) can be simplified to*

$$\gamma_j(t) = O\left(\lambda_2|\boldsymbol{\theta}(t)|_2 + s^{1/2}b_n + \lambda_1 s \max_{k\neq j}|(P_{\mathcal{R}_t})_{jk}|\right).$$

From this, it is clear that the three terms correspond to bias of ridge-shrinkage, non-stationarity and Lasso-correction. The first and last components have dynamic features and the non-stationary bias is controlled by the bandwidth and sparsity parameters. The condition $C \le |N_t|\underline{w}_t \le |N_t|\overline{w}_t \le C^{-1}$ in Theorem 2.1 rules out the case that the kernel does not use the boundary rows in the localized window and therefore avoids any jump in the time-dependent row subspaces.

Remark 2. *In Theorem 2.1, the penalty level for the tv-Lasso estimator can be chosen as $O(\sigma L_{t,2}\sqrt{\log p} + L_{t,1}s^{1/2}b_n)$. We comment that the second term in the penalty is due to the non-stationarity of $\boldsymbol{\beta}(t)$ and the factor $s^{1/2}$ arises from the weak coordinatewise smoothness requirement on its derivatives (assumption 3(a)). In the Lasso case with $\boldsymbol{\beta}(t) \equiv \boldsymbol{\beta}$ and $w(i,t) \equiv n^{-1}$, an ideal order of the penalty level λ_1 is*

$$\sigma n^{-1}\max_{j\le p}(\sum_{i=1}^n X_{ij}^2)^{1/2}(\log p)^{1/2}$$

see e.g. [Bickel et al., 2009]. In the standardized design case $n^{-1}\sum_{i=1}^n X_{ij}^2 = 1$ so that $L_{t,1} = 1$ and $L_{t,2} = n^{-1/2}$, the Lasso penalty is $O(\sigma(n^{-1}\log p)^{1/2})$, while the tv-Lasso has an additional term $s^{1/2}b_n$ that may cause larger bias. However, in our case, we estimate the time-varying coefficient vectors by smoothing the data points in the localized window. Thus, it is unnatural to standardize the re-weighted local design matrix to have unit ℓ^2 length and the additional bias $O(s^{1/2}b_n)$ is due to non-stationarity. Consider the case that X_{ij} are i.i.d. Gaussian random variables without standardization and we interpret the linear model as conditional on X. Then, under the uniform kernel, we have $L_{t,2}^2 = O_{\mathbb{P}}(\log p/|N_t|)$ and in the Lasso case penalty level is $O(\sigma|N_t|^{-1/2}\log p)$. If $s = O(\log p)$ and the bandwidth parameter $b_n = O((\log p/n)^{1/3})$, then the choice in Theorem 2.1 has the same order as the Lasso with constant coefficient vector.

Based on Theorem 2.1, we can prove that the inference algorithm in Section 2.2.2 can asymptotically control the family-wise error rate (FWER). Let $\alpha \in (0,1)$ and $\mathrm{FP}_\alpha(t)$ be the number of false rejections of $H_{0,j,t}$ based on the adjusted p-values.

Theorem 2.2 (Pointwise inference: multiple testing)**.** *Under the conditions of Theorem 2.1 and suppose that*

$$\lambda_2|\boldsymbol{\theta}(t)|_2 + s^{1/2}b_n = o(\Omega_{\min}(\lambda_2)^{1/2}) \tag{2.23}$$

we have for each fixed $t \in \varpi$

$$\limsup_{n\to\infty} \mathbb{P}(FP_\alpha(t) > 0) \le \alpha. \tag{2.24}$$

The proof of Theorem 2.2 is achieved by combining the argument of [Bühlmann, 2013, Theorem 2] and Theorem 2.1 in the appendix. Condition (2.23) requires that the shrinkage and non-stationarity biases of the tv-ridge estimator together are dominated by the variance; see also the representation (2.18), (2.19), (2.20) and (2.21). This is mild condition for the following two reasons. First, considering that the variance of the tv-ridge estimator is lower bounded when λ_2 is small enough; c.f. (2.14), the first term is quite weak in the sense that the tv-ridge estimator acts on a much smaller subspace with dimension $|N_t|$ than the original p-dimensional vector space. Second, for the choice of penalty parameter of λ_1 in Theorem 2.1, the term $s^{1/2}b_n$ in (2.23) is at most λ_1. Hence, the bias correction (including the projection and non-stationary parts) in the inference algorithm (2.11) has a dominating effect for the second term of (2.23). Consequently, provided λ_2 is small enough, the bias correction step in computing the raw p-value asymptotically approximates the stochastic component in the tv-ridge estimator.

Next, we relax the i.i.d. Gaussian error assumption. For the Gaussian process errors, we have the following result.

Theorem 2.3. *Suppose that the error process* e_i *is a mean-zero stationary Gaussian process of form (2.3) such that* $|a_m| \le K(m+1)^{-\varrho}$ *for some* $\varrho \in (1/2, 1) \cup (1, \infty)$ *and finite constant* $K > 0$*. Under assumptions 2-6 and using the same notations in Theorem 2.1 with*

$$\lambda_0 = \begin{cases} 4\sigma L_{t,2}|\mathbf{a}|_1\sqrt{\log p} & \text{if } \varrho > 1 \\ C_{\varrho,K}\sigma L_{t,2} n^{1-\varrho}\sqrt{\log p} & \text{if } 1 > \varrho > 1/2 \end{cases}, \tag{2.25}$$

we have the same representation of $\hat{\beta}(t)$ *in (2.18)–(2.22) with probability tending to one.*

Clearly, the temporal dependence strength has a dichotomy effect on the choice of λ_0 and therefore on the asymptotic properties of $\hat{\boldsymbol{\beta}}(t)$. For e_i has SRD, we have $|\mathbf{a}|_1 < \infty$ and $\lambda_0 \asymp \sigma L_{t,2}\sqrt{\log p}$. Therefore, the bias-correction part $\boldsymbol{\gamma}(t)$ of estimating $\boldsymbol{\beta}(t)$ has the same rate of convergence as the i.i.d. error case. The temporal effect only plays a role in the long-run covariance matrix of the stochastic part $\mathbf{z}(t)$. On the other hand, if e_i has LRD, then the temporal dependence has impact on both $\boldsymbol{\gamma}(t)$ and $\mathbf{z}(t)$. In addition, choice of the bandwidth parameter b_n will be very different from the SRD and i.i.d. cases. In particular, the optimal bandwidth for $\varrho \in (1/2, 1)$ is $O((\log p/n^{\varrho})^{1/3})$ which is much larger than $O((\log p/n)^{1/3})$ in the i.i.d. and SRD cases where s is bounded.

2.4 Extensions

We assume that the noise variance-covariance matrix Σ_e is known. In the i.i.d. error case $\Sigma_e = \sigma^2 \mathrm{Id}_n$, we have seen that the distribution $F(\cdot)$ is independent of σ^2, and therefore its value does not affect the inference procedure. The noise variance only impacts the tuning parameter of the initial Lasso estimator. In practice, we use the scaled Lasso to estimate σ^2, such as in our numeric and simulation studies. Given that $|\hat{\sigma}/\sigma - 1| = o_{\mathbb{P}}(1)$ [Sun and Zhang, 2012], the theoretical properties of our estimator (2.10) remains the same if we plug in the scaled Lasso variance output to our method. For temporally dependent stationary error process, estimation of Σ_e becomes more subtle since it involves n autocovariance parameters. We propose a heuristic strategy: first, run the tv-Lasso estimator to obtain the residuals; then calculate the sample autocovariance matrix and apply a banding or tapering operation $B_h(\Sigma) = \{\sigma_{jk}\mathbf{1}(|j-k| \le h)\}_{j,k=1}^{p}$ [Bickel and Levina, 2008][Cai et al., 2010][McMurry and Politis, 2010].

We provide some justification on the heuristic strategy for SRD time series models. To simplify explanation, we consider the uniform kernel and the bandwidth $b_n = 1$. Suppose we have an oracle where $\boldsymbol{\beta}(t)$ is known and we have access to the error process $e(t)$. Let Σ_e^* be the oracle sample covariance matrix of e_i with the Toeplitz structure i.e. the h-th sub-diagonal of Σ_e^* is $\sigma_{e,h}^* = n^{-1}\sum_{i=1}^{n-h} e_i e_{i+h}$. We first compare the oracle estimator and the true error covariance matrix Σ_e. Let $\alpha > 0$ and define

$$\mathcal{T}(\alpha, C_1, C_2) = \left\{ M \in ST^{p\times p} : \sum_{k=h+1}^{p} |m_k| \le C_1 h^{-\alpha}, \rho_j(M) \in [C_2, C_2^{-1}],\ \forall j = 1, \cdots, p \right\},$$

where $ST^{p\times p}$ is the set of all $p \times p$ symmetric Toeplitz matrices. If e_i has SRD, then $\Sigma_e \in \mathcal{T}(\varrho - 1, C_1, C_2)$. By the argument in [Bickel and Levina, 2008] and Lemma A.4, we can show that

$$\begin{aligned} \rho_{\max}(B_h(\Sigma_e^*) - \Sigma_e) &\le \rho_{\max}(B_h(\Sigma_e^*) - B_h(\Sigma_e)) + \rho_{\max}(B_h(\Sigma_e) - \Sigma_e) \\ &\lesssim_{\mathbb{P}} h\sqrt{\frac{\log h}{n}} + h^{-(\varrho-1)}. \end{aligned}$$

Choosing $h^* \asymp (n/\log n)^{1/(2\varrho)}$, we get

$$\rho_{\max}(B_h(\Sigma_e^*) - \Sigma_e) = O_{\mathbb{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\varrho-1}{2\varrho}}\right).$$

This oracle rate is sharper than the one established in [Bickel and Levina, 2008] for regularizing more general band-able matrices if $n = o(p)$. Here, the improved rate is due to the Toeplitz structure in Σ_e. Since Σ_e has uniformly bounded eigenvalues from zero and infinity, the banded oracle estimator $B_h(\Sigma_e^*)$ can be used as a benchmark to assess the tv-Lasso residuals $\tilde{\mathcal{E}}_t = \mathcal{Y}_t - \mathcal{X}_t\tilde{\boldsymbol{\beta}}(t)$.

Proposition 2.1. *Suppose $\Sigma_e \in \mathcal{T}(\varrho-1, C_1, C_2)$ and conditions of Lemma A.3 are satisfied except that (e_i) is an SRD stationary Gaussian process with $\varrho > 1$. Then*

$$\rho_{\max}(B_h(\hat{\Sigma}_e) - B_h(\Sigma_e^*)) = O_{\mathbb{P}}(h\lambda_1 s^{1/2}). \tag{2.26}$$

With the choice $h^* \asymp (n'/\log n')^{1/2\varrho}$ where $n' = |N_t|$, we have

$$\rho_{\max}(B_h(\hat{\Sigma}_e) - \Sigma_e)) = O_{\mathbb{P}}\left(\left(\frac{\log n'}{n'}\right)^{\frac{\varrho-1}{2\varrho}} + \left(\frac{n'}{\log n'}\right)^{\frac{1}{2\varrho}}\left(\sqrt{\frac{s\log p}{n'}} + sb_n)\right)\right). \tag{2.27}$$

It is interesting to note that the price we pay to choose h for not knowing the error process is the second term in (2.27). Bandwidth selection for the smoothing parameter b_n is a theoretically challenging task in the high dimension. Asymptotic optimal order for the parameter is available up to some unknown constants depending on the data generation parameters. We use cross-validation (CV) in our simulation studies and real data analysis examples.

In the i.i.d. error case, the noise is assumed to be zero-mean Gaussian. First, it is easy to relax this assumption to distributions with sub-Gaussian tails and Theorem 2.1 and 2.2 continue to hold, in view that the large deviation inequality and the Gaussian approximation for a weighted partial sum of the error process only depend on the tail behavior and therefore moments of e_i. Second and more importantly, the sub-Gaussian assumption may even be knocked down by allowing the i.i.d. noise processes with algebraic tails, or equivalently e_i have moments up to a finite order. The consequence of this relaxation is that larger penalty parameter for the tv-Lasso is needed for errors with polynomial moments. This is the content of the following theorem. For simplicity, we assume $K(\cdot)$ is the uniform kernel in Theorem 2.4.

Theorem 2.4 (Heavy-tailed errors)**.** *Let conditions in Theorem 2.1 be satisfied and suppose $\mathbb{E}|e_i|^q < \infty, q >$*

2. *Choose*

$$\lambda_0 = C_q \max\left\{(p\mu_{n,q})^{1/q},\ \sigma L_t (\log p)^{1/2}\right\}, \qquad \text{for large enough } C_q > 0, \tag{2.28}$$

where $\mu_{n,q} = \sum_{i \in N_t} |w(t,i) X_{ij}|^q$. *Then, we have the same representation (2.18) and Theorem 2.2 holds with probability tending to one.*

2.5 Simulation Results

In this section, we observe the performance of the proposed time-varying bias corrected ridge inference procedure through simulation studies. We first generate an $n \times p$ design matrix using n i.i.d. rows from $\mathcal{N}_p(0, I)$, with $n = 300$ and $p \in \{300, 500\}$. The time-varying coefficient vectors $\boldsymbol{\beta}(t)$ are set up such that there are $s = 3$ non-zero elements and $p - 3$ zeros for all $t \in [0, 1]$. These non-zero elements in $\boldsymbol{\beta}(t)$ are generated by sampling nodes from a uniform distribution $\mathcal{U}(-b, b)$ at regular time points and smoothly interpolating on the interval $[0, 1]$ using cubic splines.

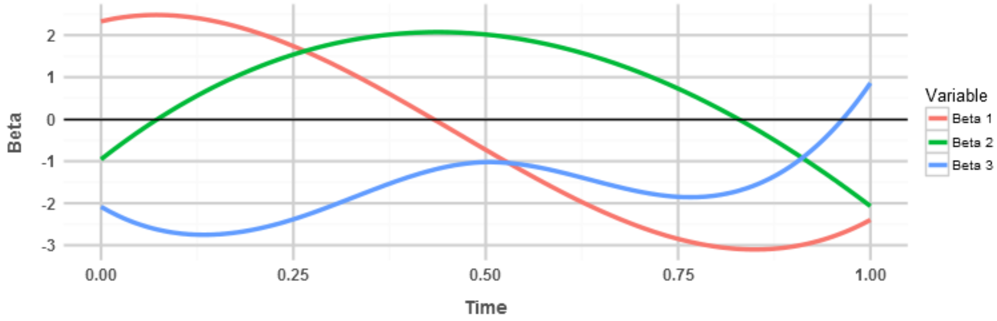


Figure 2.2: Simulated non-zero $\boldsymbol{\beta}(t)$ with $b = 2.5$

We simulated several sets of stationary error processes:

1. e_i are i.i.d. $N_n(0, I)$.

2. e_i is an AR(1) process $e_i = \varphi e_{i-1} + \epsilon_i$ where $\varphi = 0.7$ and ϵ_i are i.i.d. $N_n(0, I)$.

3. e_i are i.i.d. Student's $t(3)/\sqrt{3}$.

The remaining parameters include the kernel bandwidth $b_n = 0.1$, $\lambda_1 = \sqrt{2\log(p)/n}$, $\lambda_2 = 1/n$, and $\zeta = 0$. Again, we highlight that no tuning is required from the proposed method.

For individual testing $H_0 : \beta_j = 0$ at time t, we reject the null hypothesis at significance level α if the corresponding raw p-value $\tilde{P}_{j,t} \leq \alpha$. Using the framework above with an empty active set $\mathcal{S} = \varnothing$ demonstrates a $< 5\%$ nominal type I error rate, albeit conservatively.

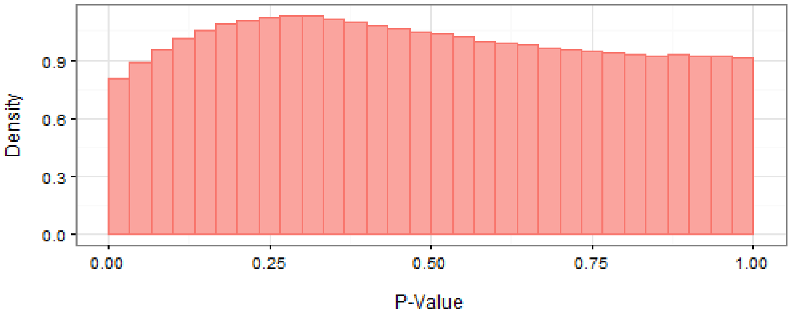


Figure 2.3: Raw p-values under Null

For multiple testing $H_0 : \beta_j = 0, j \in G$ at time t, we reject the null hypothesis at significance level α if the corresponding multiplicity-adjusted p-value $P_{j,t} \leq \alpha$. Therefore, our proposed method's false positive (FP) rate over the interval $B = [b_n, 1 - b_n]$ is written as

$$\frac{1}{n(1-b_n)(p-s)} \sum_{j \in \mathcal{S}^c} \sum_{t \in B} \mathbb{P}(P_{j,t} \leq \alpha) \tag{2.29}$$

and the false negative (FN) rate is written as

$$\frac{1}{n(1-b_n)s} \sum_{j \in \mathcal{S}} \sum_{t \in B} \mathbb{P}(P_{j,t} > \alpha) \tag{2.30}$$

The familywise error rate (FWER) in our simulation setup is defined as the proportion of multiplicity-adjusted tests with at least 1 type I error.

$$\text{FWER} = \frac{1}{n(1-b_n)} \sum_{t \in B} \mathbb{P}\left[\bigcap_{j \in \mathcal{S}^c} \{P_{j,t} | P_{j,t} \leq \alpha\} \neq \varnothing \right] \tag{2.31}$$

For each simulation setup, we report the false positive counts and rates, false negative counts and rates, family-wise error rates, and root mean square errors (RMSE) of $\boldsymbol{\beta}$. These examples are high-dimensional since we have $ns(1 - b_n) = 720$ nonzero parameters and a sample size of 300. We compared the performance of the proposed method against the following procedures:

1. (TV-Lasso) - The time-varying Lasso, an adaptation of the standard LASSO to the kernel smoothing environment with $b_n = 0.1$ and λ_1 selected using cross validation. (Generally on the order of $1.5 \times \sqrt{2\log(p)/n}$)

2. (FP-Lasso) - The false-positive Lasso, where λ_1 is tuned to match the type I error rate of the proposed method. (Generally on the order of $5 \times \sqrt{2\log(p)/n}$). This allows us to compare power at similar levels of individual and family-wise errors. In practice, such precise tuning of λ_1 on type I errors would be impossible since the active set $\mathcal{S}$ is unknown, making FP-Lasso "pseudo-oracle."

3. (FP-LPDE) - An adaptation of the de-biased LASSO [Zhang and Zhang, 2014] inference procedure with the estimator $\hat{\boldsymbol{\beta}}_{DeBias}(t) = \hat{\boldsymbol{\beta}}_{Lasso}(t) + (nb_n)^{-1}\mathbf{M}_t\mathcal{X}_t^\top(\mathcal{Y}_t - \mathcal{X}_t\hat{\boldsymbol{\beta}}_{Lasso}(t))$, where the multiplicity-adjusted significance level α is selected to yield identical type I error rates as the proposed method. Also "pseudo-oracle".

4. (Non-TV) - The original non-time-varying method of [Bühlmann, 2013] which ignores the dynamic structures.

Our results are shown in Tables 2.5 and 2.5, with 20 replications for each setting.

i.i.d. $\mathcal{N}(0,1)$

Method	FP(%)	FP(n)	FN(%)	FN(n)	FWER	RMSE
TV-Lasso	0.0751	5355.3	0.0551	39.70	1	0.0537
FP-Lasso	1.3×10^{-4}	9.30	0.2469	177.80	0.0369	0.1122
FP-LPDE Proposed	1.3×10^{-4}	9.30	0.2184	157.25	0.0371	0.1541
Non-TV	0.5222	37224	0.1333	96	1	0.4507

AR(1), $\varphi = 0.7$

Method	FP(%)	FP(n)	FN(%)	FN(n)	FWER	RMSE
TV-Lasso	0.0755	5382.5	0.0544	39.15	1	0.0532
FP-Lasso	1.1×10^{-4}	7.80	0.2647	190.60	0.0319	0.1120
FP-LPDE Proposed	1.1×10^{-4}	7.85	0.2272	165.80	0.0321	0.1546
Non-TV	0.5337	38040	0.0667	48	1	0.4531

i.i.d. $t(3)/\sqrt{3}$

Method	FP(%)	FP(n)	FN(%)	FN(n)	FWER	RMSE
TV-Lasso	0.0693	4938.6	0.0544	39.15	1	0.0525
FP-Lasso	1.7×10^{-4}	12.00	0.2460	177.15	0.0402	0.1112
FP-LPDE Proposed	1.7×10^{-4}	12.00	0.2303	165.80	0.0410	0.1542
Non-TV	0.5244	37380	0.1667	120	1	0.4510

Table 2.1: Simulation results for the case of $n = 300, p = 300, s = 3, b = 2.5$.

i.i.d. $\mathcal{N}(0,1)$

Method	FP(%)	FP(n)	FN(%)	FN(n)	FWER	RMSE
TV-Lasso	0.0484	5776.7	0.0655	47.15	1	0.0392
FP-Lasso	3.1×10^{-5}	3.65	0.2528	182.05	0.0152	0.0798
FP-LPDE	3.1×10^{-5}	3.65	0.2296	165.30	0.0150	0.0664
Proposed	3.1×10^{-5}	3.65	0.2240	161.30	0.0150	0.1260
Non-TV	0.0299	3564	0.6000	432.00	1	0.1885

AR(1), $\varphi = 0.7$

Method	FP(%)	FP(n)	FN(%)	FN(n)	FWER	RMSE
TV-Lasso	0.0486	5801.3	0.0619	44.55	1	0.0393
FP-Lasso	3.0×10^{-5}	3.60	0.2645	190.45	0.0148	0.0810
FP-LPDE	3.0×10^{-5}	3.60	0.2431	175.05	0.0150	0.0690
Proposed	3.0×10^{-5}	3.60	0.2308	166.20	0.0150	0.1260
Non-TV	0.0330	3936	0.5667	408	1	0.1892

i.i.d. $t(3)/\sqrt{3}$

Method	FP(%)	FP(n)	FN(%)	FN(n)	FWER	RMSE
TV-Lasso	0.0469	5591.8	0.0672	48.35	1	0.0397
FP-Lasso	4.1×10^{-5}	4.85	0.3828	275.60	0.0188	0.0965
FP-LPDE	4.1×10^{-5}	4.85	0.2706	194.85	0.0149	0.0701
Proposed	4.1×10^{-5}	4.85	0.2351	169.30	0.0158	0.1260
Non-TV	0.0321	3828	0.6333	456	1	0.1893

Table 2.2: Simulation results for the case of $n = 300, p = 500, s = 3, b = 2.5$.

Empirically, we see that family-wise error control is not maintained in the non-time-varying case using [Bühlmann, 2013], since it's unable to accommodate the flip-flopping nature of $\boldsymbol{\beta}(t)$. The proposed method does maintain FWER control in all simulation setups, but is conservative in the case of $p = 500$. This FWER control naturally demands a very small false positive rate, whereas the time-varying Lasso has much greater false positive rates due to the the bias from l_1 regularization.

We also observe that the proposed method fares worse in terms of RMSE than the other time-varying methods. This is largely due to fixing λ_2 to be small in order to circumvent complications surrounding parameter tuning such as cross validation and to make the problem more amenable to the conditions of 2.1. Due to the bounds on Var $\left(\hat{\beta}_j(t)\right)$, minimizing RMSE is likely to be sub-optimal for detection, in terms of λ_2 selection.

At similar type I error levels, the proposed time-varying de-biased ridge method yields greater detection power in all simulation setups. De-biasing the Lasso helps bring the two methods closer in terms of power, but comes at the expense of substantially increased complexity and computation time. The simulations were performed using an Intel i5-4970K running R 3.2.2 for Windows with Intel MKL linear algebra libraries.

Method	Runtime
TV-Lasso	1
FP-Lasso	19
FP-LPDE	1445
Proposed (raw p-values only)	9
Proposed (adjusted p-values)	26
Non-TV	< 1

Table 2.3: Time to run 20 replications, in minutes. $p = 500$ with $\mathcal{N}(0, 1)$ i.i.d. errors

The Lasso and De-Biased Lasso are based on code from `glmnet` and `SSLasso` by J. Friedman and A. Javanmard, respectively [Friedman et al., 2010a][Javanmard and Montanari, 2014]. The reported run times for FP-Lasso and FP-LPDE includes time spent on divide and conquer λ_1 and α searches for FPR matching. The additional computation time is quite substantial in the former case and negligible in the latter. The time-varying methods above are bottle-necked by estimation of the covariances of $\mathcal{X}$ within local bandwidths, which we attempt to remedy in Chapter 3 using structure decomposition.

By varying the signal magnitude scalar b from 0.25 to 2.5, we can examine the behavior of the proposed method and its FWER control at various signal-to-noise ratios (S/N).

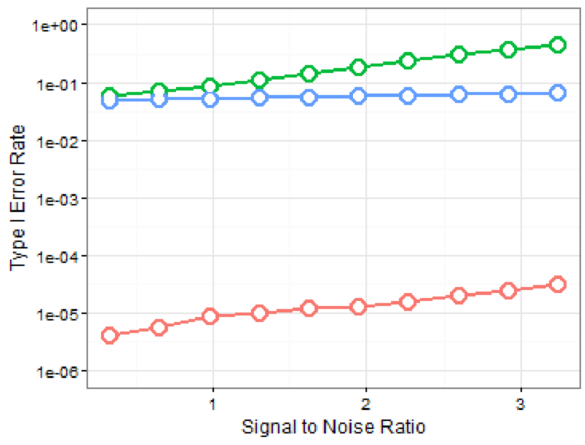


Figure 2.4: Type I Errors vs S/N

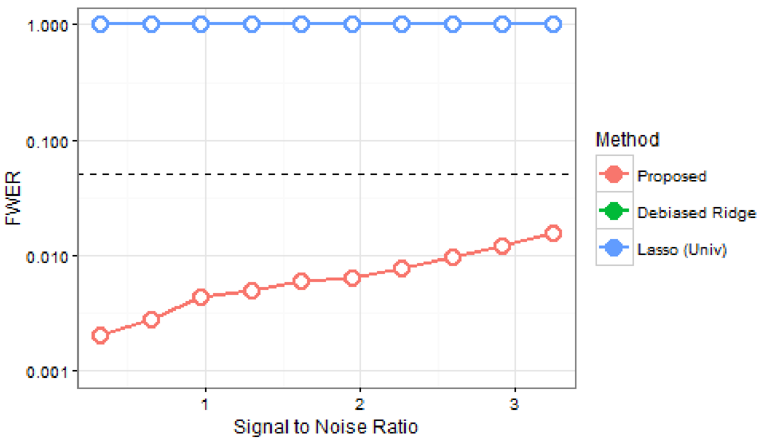


Figure 2.5: Familywise Errors vs S/N

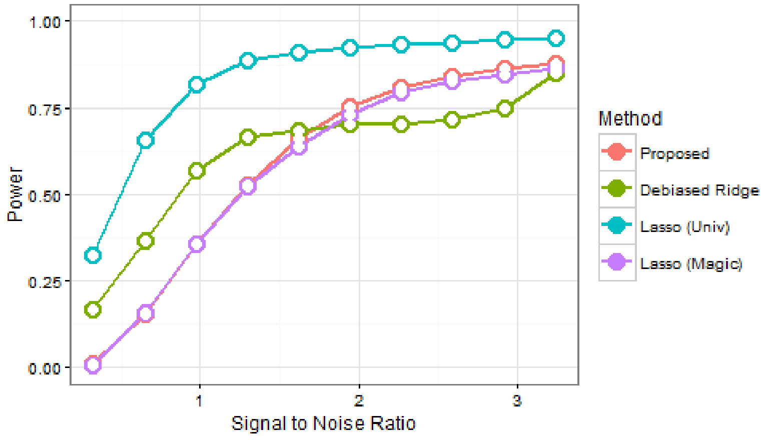


Figure 2.6: Power vs S/N

The proposed method maintains FWER control across the spectrum of signal-to-noise ratios examined above, although of course the power suffers at lower signal-to-noise ratios.

2.6 Real Data Example - Learning Brain Connectivity

We illustrate our proposed method with a problem about estimating functional brain connectivity in patients from a Parkinson's disease study. The principle of functionally segregated brain organization in humans is well established in imaging neuroscience, and connectivity is understood as a network of statistical dependencies between different regions of the nervous system.

Slowly time-varying graphs have strong implications in modeling brain connectivity networks using resting state functional magnetic resonance imaging (fMRI) data. Traditional correlation analysis of resting state blood-oxygen-level-dependent (BOLD) signals of the brain show considerable temporal variation on small timescales [CITE], and many treatments for Parkinson's disease are evaluated by medical professionals based on changes to a subject's connectivity network. For example, patients with Parkinson's disease generally have increased connectivity in the primary motor cortex, especially during "off state" times. [CITE]

Furthermore, in view of the high spatial resolution of fMRI data, brain networks of subjects at rest are believed to be structurally homogeneous with subtle fluctuations in some, but a small number, of connectivity edges [CITE]. Therefore, a popular approach to learn brain connectivity is the node-wise regression network (i.e. the neighborhood selection procedure), where the time-varying coefficients represent dynamic features of the corresponding edges. We do remark, however, that the neighborhood selection approach we adopt in this example is merely an approximation of the full multivariate distributions due to ignoring correlation among node-wise responses. This may lead to some power loss in finite samples, but is asymptotically equivalent in terms of variable selection.

Our real data example uses fMRI data collected from a study of patients with Parkinson's disease (PD) and their respective normal controls. PD is typically characterized by deviations in functional connectivity between various regions of the brain. Additionally, resting state functional connectivity has been shown as a candidate biomarker for PD progression and treatment, where more advanced stages or manifestations of PD are associated with greater deviations from normal connectivity. Each resting state data matrix in our example contains 240 time points and 52 brain regions of interest (ROI). The time points are evenly sampled and the time indices are normalized to [0,1]. Previous study of this dataset showed that the temporal and spatial connectivity patterns differ significantly between PD and control subjects [Liu et al., 2014].

The brain connectivity network is constructed using the neighborhood selection procedure. In essence, it is a sequence of time-varying linear regressions by enumerating each ROI as the response variable and sparsely regressing on all the other ROIs. Since fMRI data was collected from subjects at rest rather than subjects assigned specific tasks at certain times, we did not pool the data across subjects for analysis.

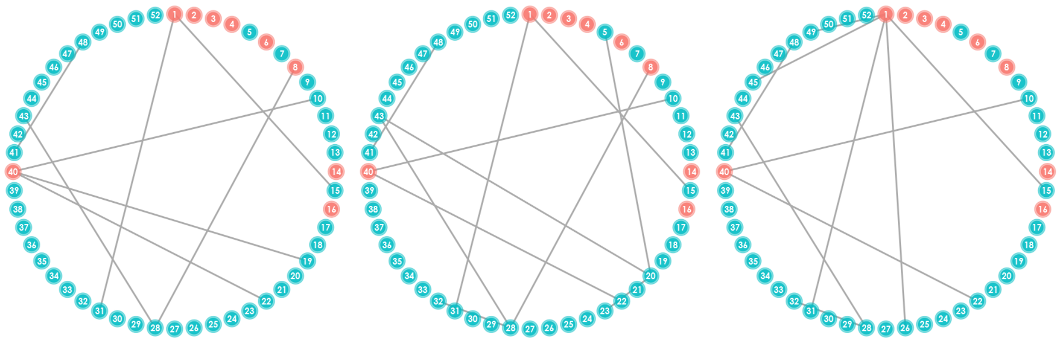


Figure 2.7: Connectivity Network in Control Subject around $t = 0.25$

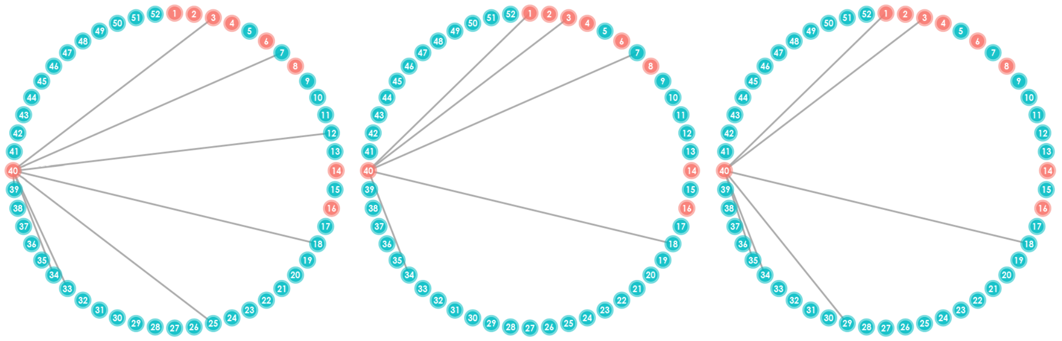


Figure 2.8: Connectivity Network in Parkinson's Subject around $t = 0.25$

In the examples above, we plot the connectivity networks of one healthy and one Parkinson's diagnosed patient for 3 sequential time points surrounding $t = 0.25$. Regions of the brain known to be associated with motor control are highlighted using red nodes. In contrast, blue nodes designate areas of the brain either known to be unrelated to motor control or whose functions in humans are not well understood. Different patterns of connectivity in the networks can be found by comparing the PD and control subjects. From the graphs, we can observe the slow change in the networks over time. Most edges are preserved on a small timescale, but there are a few number of edges changing. For instance in the PD subject, ROI 01 $\rightarrow$ 40 is unconnected in the first time point but is connected in the second and remains connected in the third. Generally, when there is substantial activity to study, we tend to observe a more diverse set of edges in healthy subjects and greater connectivity involving the motor-related regions in PD subjects.

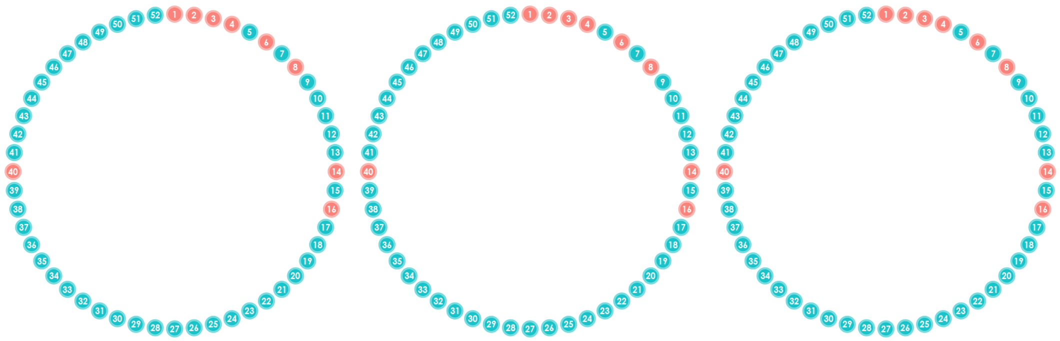


Figure 2.9: Connectivity Network in Control Subject around $t = 0.80$

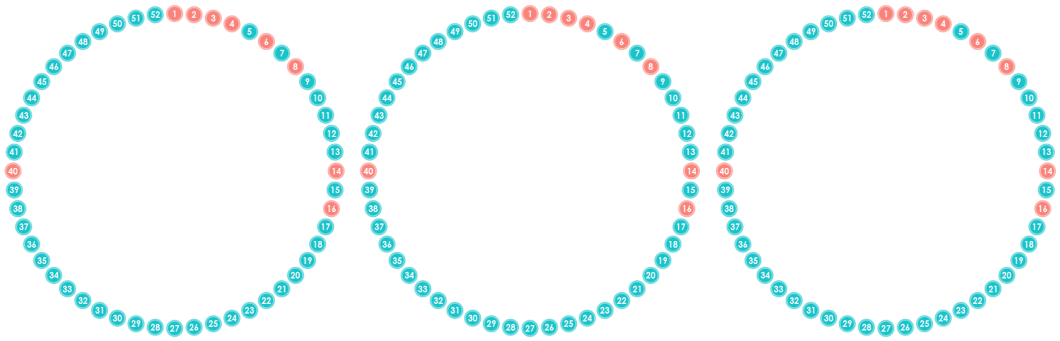


Figure 2.10: Connectivity Network in Parkinson's Subject around $t = 0.50$

There are also times when, due to the nature of stringent FWER control and variable selection process, that the estimated connectivity networks may be exceptionally sparse or empty. This serves to reinforce the importance of the time-varying design, particularly for resting state data, since traditional methods which don't study temporal variations have less power to detect differences between subjects when the signals only manifest for short periods of time.

Chapter 3

Kronecker Structured Covariance Estimation

3.1 Introduction

There has been much recent interest in the estimation of sparse graphical models in the high-dimensional setting for an $n \times p$ Gaussian matrix $\mathbf{X}$, where the observations $\mathbf{x}_1, ..., \mathbf{x}_n$ are i.i.d. $\mathcal{N}(\mu, \boldsymbol{\Sigma})$, $\mu \in \mathbb{R}^p$, and $\boldsymbol{\Sigma}$ is a positive definite $p \times p$ matrix. A graphical model can be built by using the variables or features as nodes and by using non-zero elements of the precision or inverse covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ as edges such that nodes representing features i, j are connected if $\Omega_{ij} \neq 0$. The zeroes in Ω correspond to no connecting edge, indicating that the pair of variables are conditionally independent of each other, given the other variables in $\mathbf{X}$ [Lauritzen, 1996]. These models see applications in various fields such as meteorology, communications, and genomics. For example, graphical models are frequently used in gene expression analysis to study patterns of association between different genes and create therapies targeting these pathways.

The classical method of estimating the covariance matrix $\boldsymbol{\Sigma}$ is the sample covariance matrix:

$$\hat{\mathbf{S}} = \frac{1}{n-1}\mathbf{X}^\top \mathbf{X} \tag{3.1}$$

While the sample covariance matrix is an unbiased estimator of the true covariance matrix, it suffers from high variance in a high-dimensional setting where $n < p$. Furthermore, the resulting matrix is singular and incompatible with problems where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the parameter of interest. Alternatively, a low rank approximation is sometimes used where we take the r greatest principle components from an eigendecomposition of $\hat{\mathbf{S}}$.

$$\hat{\mathbf{S}}_{PCA} = \sum_{i=1}^{r} \sigma_i^2 v_i v_i^\top \tag{3.2}$$

The PCA estimator has a very useful interpretation in that it solves the rank constrained Frobenius norm minimization problem [Eckart and Young, 1936].

$$\arg \min_{\mathbf{S} \in S_{++}, \mathrm{rank}(S) \leq r} |\hat{\mathbf{S}} - \mathbf{S}|_F^2 \tag{3.3}$$

However, the PCA estimator yields substantial bias in the high-dimensional setting. [Lee et al., 2010] [Rao et al., 2008]

The maximum likelihood approach is a direct estimation method where the precision matrix estimator $\hat{\mathbf{\Omega}}_{MLE}$ is obtained by maximizing the likelihood with respect to $\mathbf{\Sigma}^{-1}$, and explained in more detail in the Graphical Lasso section below. For the purpose of graphical models, however, the maximum likelihood precision matrix estimator lacks interpretability and yields high variance in a high-dimensional setting. Specifically, $\hat{\mathbf{\Omega}_{MLE}}$ will not have any elements which are exactly zero. Consequently, graphs constructed using $\hat{\mathbf{\Omega}}_{MLE}$ would not be useful for identifying conditional independence between pairs of variables since edges exist between all pairs of nodes.

Thus arises a need for the estimation of a sparse dependence structure where many elements of the inverse covariance matrix estimator are set to zero. This literature advocating sparse estimation of covariance and inverse covariance matrices can be traced back to [Dempster, 1972]. An early approach to achieving a sparse estimator is the backwards selection method, where the least significant edges are sequentially removed from a fully connected graph until the remaining edges are all significant according to partial correlation tests. The backwards selection method did not take multiple testing into consideration, although [Drton and Perlman, 2007] later proposed a conservative procedure which did.

More recent work in graphical model estimation involves nodewise regression procedures similar to the Parkinson's data example from Chapter 1. [Meinshausen and Bühlman, 2006] proposed regressing each variable on all the other variables using ℓ_1 penalized regression, with graph edges representing the significant regression results. [Zhou et al., 2009] further extended the method to other variants of penalized regression such as the adaptive LASSO. These nodewise regression methods are able to consistently recover the support of $\mathbf{\Omega}$ to produce a directed graphical model but cannot estimate the elements of $\mathbf{\Omega}$ themselves.

A maximum likelihood approach with the ℓ_1 penalty was studied by [Yuan and Lin, 2007], [Banerjee et al., 2008], [Friedman et al., 2010b], and [d'Asprémont et al., 2008]. These approaches were later generalized to the smoothly clipped absolute deviation (SCAD) penalty [Fan et al., 2009], [Lam and Fan, 2009].

This chapter explores the algorithms, properties, and performance of Kronecker graphical lasso (KGLasso) methods for the purposes of constructing Gaussian graphical models and the estimation of Kronecker decompose-able covariance and precision matrices. We begin with ordinary KGLasso [Tsiligkaridis, 2014], followed by our extension to the joint Kronecker graphical lasso (JKGLasso) case for the estimation of covariance matrices across different groups of subjects.

3.1.1 Notation

Denote $\mathbf{I}_p$ to be the $p \times p$ identity matrix and $\mathbf{1}_p$ a vector of length p with all entries equal to 1. Define vec to be the matrix vectorization function in $\mathbb{R}^{p\times q} \to \mathbb{R}^{pq}$ such that vec($\mathbf{M}$) is the vectorized form of $\mathbf{M}$ obtained by concatenating the columns of $\mathbf{M}$. For higher tensor spaces, let vec : $\mathbb{R}^{p_1\times\cdots\times p_K} \to \mathbb{R}^{p_1\cdots p_K}$ be the tensor vectorization function where we concatenate via the array dimensions in reverse order beginning with K. Let $S(\cdot)$ denote the soft thresholding operator $S : \mathbb{R}^2 \to \mathbb{R}^1$ such that $S(x, c) = \text{sign}(x) \max(|x| - c, 0)$

Define $|\mathbf{M}|_1$ to be the L^1 norm of a matrix $\mathbf{M}$. Define $|\mathbf{M}|_F$ to be the Frobenius norm of a matrix $\mathbf{M}$. Define $S^p = \{\mathbf{M} \in \mathbb{R}^{p\times p} : \mathbf{M} = \mathbf{M}^\top\}$ to be the set of $p \times p$ symmetric matrices, and let S^p_+ denote the set of symmetric positive definite matrices and S^p_{++} the set of symmetric semi-positive definite matrices. Denote $\mathbf{M}_{ij}$ and $(\mathbf{M})_{ij}$ to be the (i, j)-th element in matrix $\mathbf{M}$ where $i, j \in \mathbb{N}^1$. Define $\mathbf{M}_{IJ}$ and $(\mathbf{M})_{IJ}$, $I \in \mathbb{N}^{p_1}, J \in \mathbb{N}^{p_2}$ to be the sub-matrix of $\mathbf{M}$ corresponding to the (i, j)-th elements in $\mathbf{M}$ such that $i \in I$ and $j \in J$.

For a sequence of positive real numbers $\{a_n\}_{n\in\mathbb{N}}$ and random variables $\{X_n\}_{n\in\mathbb{N}}$ on space $(\Omega, \mathcal{F}, P)$, define $X_n = O_P(1)$ to say that X_n is stochastically bounded: $\forall\epsilon > 0, \exists M > 0$ such that $\text{P}(|X_n| > M) < \epsilon \ \forall n$. Define $X_n = O_P(a_n)$ to say $\frac{X_n}{a_n} = O_P(1)$.

3.1.2 The Graphical Lasso

Given a set of n i.i.d. multivariate Gaussian observations $\{\mathbf{x}_j\}_{j=1}^n,\ \mathbf{x}_j \in \mathbb{R}^p$ with mean zero (without loss of generality), positive definite covariance matrix $\mathbf{\Sigma} \in S^p_{++}$, and sample covariance matrix $\hat{\mathbf{S}} = \frac{1}{n}\sum_{j=1}^n \mathbf{x}_j\mathbf{x}_j^\top$, then the log-likelihood $l(\mathbf{\Sigma})$ is written

$$l(\mathbf{\Sigma}) = \log\det(\mathbf{\Sigma}^{-1}) - \text{tr}(\mathbf{\Sigma}^{-1}\hat{\mathbf{S}}) \tag{3.4}$$

The late 2000's saw an interest in ℓ_1-regularized maximum likelihood estimators [Banerjee et al., 2008] [Friedman et al., 2010b] [Ravikumar et al., 2010] for $\mathbf{\Sigma}$ known as the graphical Lasso (GLasso) estimators and are obtained by solving the ℓ_1-regularized minimization problem

$$\hat{\mathbf{\Sigma}} = \arg\min_{\mathbf{\Sigma}\in S^p_{++}} -l(\mathbf{\Sigma}) + \lambda|\mathbf{\Sigma}^{-1}|_1 \tag{3.5}$$

$$= \arg\min_{\mathbf{\Sigma}\in S^p_{++}} \operatorname{tr}(\hat{\mathbf{S}}\mathbf{\Sigma}^{-1}) - \log\det\left(\mathbf{\Sigma}^{-1}\right) + \lambda|\mathbf{\Sigma}^{-1}|_1 \tag{3.6}$$

where $\lambda \geq 0$ is a tuning parameter. [Friedman et al., 2010a] introduced an iterative algorithm using block coordinate descent to obtain the GLasso estimators in $\mathcal{O}(p^4)$ time, and $\mathcal{O}(p^3)$ in the sparse case. An alternative algorithm was introduced by [Hsieh et al., 2011] with the same computational complexity.

[Rothman et al., 2008] and [Zhou et al., 2010] showed the high dimensional consistency in Frobenius norm of the GLasso estimators for an appropriate choice of λ

$$\|\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_F = O_P\left(\sqrt{\frac{(p+s)\log(p)}{n}}\right) \tag{3.7}$$

where $\hat{\mathbf{\Sigma}}$ is the GLasso estimator and s is a measure of sparsity denoting an upper bound on the number of nonzero off-diagonal entries in $\mathbf{\Sigma}^{-1}$

3.2 The Kronecker Covariance Framework

We consider the Kronecker covariance model:

$$\mathbf{\Sigma} = \mathbf{A} \otimes \mathbf{B} \tag{3.8}$$

Where $\mathbf{\Sigma}$ is the $p \times p$ covariance matrix for the observed data and $\mathbf{A}$ and $\mathbf{B}$ are $p_A \times p_A$ and $p_B \times p_B$ positive definite matrices, respectively, such that $p_A p_B = p$. This type of low-dimensional Kronecker covariance matrix representation can be found in communications to model signal propagation from systems with multiple input multiple output (MIMO) radio antenna arrays such as modern home wi-fi routers [Werner and Jansson, 2007][Werner et al., 2008], in genomics to estimate correlations between genes and their associated factors [Yin and Li, 2012], facial recognition [Zhang and Schneider, 2010], recommendation systems, and missing data imputation [Allen and Tibshirani, 2010]. Typically, the motivation behind using the Kronecker model is the pursuit of computational and mathematical tractability or a very low dimensional representation given the model assumptions.

An immediate concern when using the Kronecker covariance model is that estimators for $\mathbf{A}$ and $\mathbf{B}$ are not unique, since each is identifiable only up to a constant. This is true of the methods introduced in this chapter, although the resulting estimator for $\mathbf{\Sigma} = \mathbf{A} \otimes \mathbf{B}$ is unique.

[Allen and Tibshirani, 2010] referred to this model as a transpose-able model, where both the rows and columns are considered to be features of interest. Transpose-able models and the Kronecker covariance model are special cases of the matrix variate normal distribution from [Efron, 2009], where separate covariance matrices are used for the rows and columns. The authors provided a concept of a movie recommendation engine which used this model such that the relationship between Customer A's rating of Movie 1 and Customer B's rating of Movie 2 is modeled using the interaction between Customers A and B, and Movies 1 and 2.

Generally speaking, the simple Kronecker product covariance structure follows when we have a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with row means $\mathbf{v} \in \mathbb{R}^n$, column means $\mathbf{u} \in \mathbb{R}^p$, row covariance $\mathbf{A} \in \mathbb{R}^{n \times n}$, and column covariance $\mathbf{B} \in \mathbb{R}^{p \times p}$. By vectorizing the data matrix $\mathbf{X}$, we have

$$\operatorname{vec}(\mathbf{X}) \sim \mathcal{N}\left(\operatorname{vec}(\mathbf{v}\mathbf{1}_p^\top + \mathbf{1}_n \mathbf{u}^\top), \mathbf{A} \otimes \mathbf{B}\right) \tag{3.9}$$

where $\mathbf{1}_k$ is a vector of length k with all entries equal to 1. Therefore, an element X_{ij} in $\mathbf{X}$ is distributed $X_{ij} \sim \mathcal{N}(v_i + u_j, \sigma_{ij})$ and follows a mixed effects model without an assumption of independence between errors from the rows and columns.

The model can be readily extended beyond 2 dimensional arrays for $\mathbf{X}$ using appropriate vectorization. [Flaxman et al., 2015] and [Bonilla et al., 2008] compared the Kronecker covariance structure to tensor Gaussian products.

3.3 The Kronecker Graphical Lasso for One Group

Suppose we have n i.i.d. observations from a multivariate Gaussian distribution with mean zero and covariance $\mathbf{\Sigma} = \mathbf{A} \otimes \mathbf{B}$ where $\mathbf{A} \in S_{++}^{p_A}, \mathbf{B} \in S_{++}^{p_B}$ then the ℓ_1-penalized maximum likelihood estimator is obtained by solving

$$\begin{aligned}
\hat{\mathbf{\Sigma}} &= \arg \min_{\mathbf{\Sigma} \in S_{++}^p} \operatorname{tr}(\hat{\mathbf{S}}\mathbf{\Sigma}^{-1}) - \log\det\left(\mathbf{\Sigma}^{-1}\right) + \lambda |\mathbf{\Sigma}^{-1}|_1 && (3.10)\\
&= \arg \min_{\mathbf{\Sigma} \in S_{++}^p} \operatorname{tr}\left[\hat{\mathbf{S}}(\mathbf{A}^{-1} \otimes \mathbf{B}^{-1})\right] - p_B \log\det\left(\mathbf{A}^{-1}\right) - p_A \log\det\left(\mathbf{B}^{-1}\right) + \lambda |\mathbf{A}^{-1}|_1 |\mathbf{B}^{-1}|_1 && (3.11)
\end{aligned}$$

where $\lambda \geq 0$ is a regularization parameter. When $\mathbf{A}$ or $\mathbf{B}$ is fixed in the objective minimization function above, then the function is convex with respect to the other argument [Tsiligkaridis and Hero, 2013]. Therefore, we consider an alternating or "flip-flop" approach to handling the dual problem of fixing one half of the Kronecker covariance matrix and optimizing over the other half.

The original penalized flip-flop (FFP) algorithm introduced by [Tsiligkaridis, 2014] and [Allen and Tibshirani, 2010] is given below. Both minimization steps are solved using GLasso.

Flip Flop Algorithm for KGLasso

```
Input $\hat{\mathbf{S}}, \lambda > 0, \epsilon > 0$
Initialize $\hat{\mathbf{\Sigma}}^{-1} = \mathbf{I}_p$ and $\mathbf{A}^{-1} \in S_{++}^{p_A}$
repeat
    $\hat{\mathbf{\Sigma}}_{prev}^{-1} \leftarrow \hat{\mathbf{\Sigma}}^{-1}$
    $\mathbf{T}_A \leftarrow \frac{1}{p_A} \sum_{i,j=1}^{p_A} \mathbf{A}_{i,j}^{-1} \hat{\mathbf{S}}_{j,i}$
    $\lambda_A \leftarrow \frac{\lambda |\mathbf{A}^{-1}|_1}{p_A}$
    $\mathbf{B}^{-1} \leftarrow \arg\min_{\mathbf{B} \in S_{++}^{p_B}} \left[ \mathrm{tr}(\mathbf{B}^{-1}\mathbf{T}_A) - \log\det\left(\mathbf{B}^{-1}\right) + \lambda_A |\mathbf{B}^{-1}|_1 \right]$
    $\mathbf{T}_B \leftarrow \frac{1}{p_B} \sum_{i,j=1}^{p_B} \mathbf{B}_{i,j}^{-1} \hat{\mathbf{S}}_{j,i}$
    $\lambda_B \leftarrow \frac{\lambda |\mathbf{B}^{-1}|_1}{p_B}$
    $\mathbf{A}^{-1} \leftarrow \arg\min_{\mathbf{A} \in S_{++}^{p_A}} \left[ \mathrm{tr}(\mathbf{A}^{-1}\mathbf{T}_B) - \log\det\left(\mathbf{A}^{-1}\right) + \lambda_B |\mathbf{A}^{-1}|_1 \right]$
    $\hat{\mathbf{\Sigma}}^{-1} \leftarrow \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$
until $\|\hat{\mathbf{\Sigma}}^{-1} - \hat{\mathbf{\Sigma}}_{prev}^{-1}\| \leq \epsilon$
```

The flip flop algorithm for KGLasso has complexity in $\mathcal{O}(p_A^4 + p_B^4)$ compared to GLasso's $\mathcal{O}(p_A^4 p_B^4)$.

3.3.1 Simulation Results

We consider moderately sparse covariance matrices $\mathbf{\Sigma}$ with dimension $p = 400$, decompose-able into smaller matrices $\mathbf{\Sigma} = \mathbf{A} \otimes \mathbf{B}$ with dimensions $p_A = p_B = 20$. We construct $\mathbf{A}, \mathbf{B}$ from different distributions including a block Toeplitz structure whose block structure is likely to favor the Kronecker product representation, and a more general structure based on a positive definite Erdös - Rényi graph [Erdös and Rényi, 1960]. A visualization of the latter is given below, where greys represent zeros, lighter shades represent more positive values, and darker shades represent more negative values.

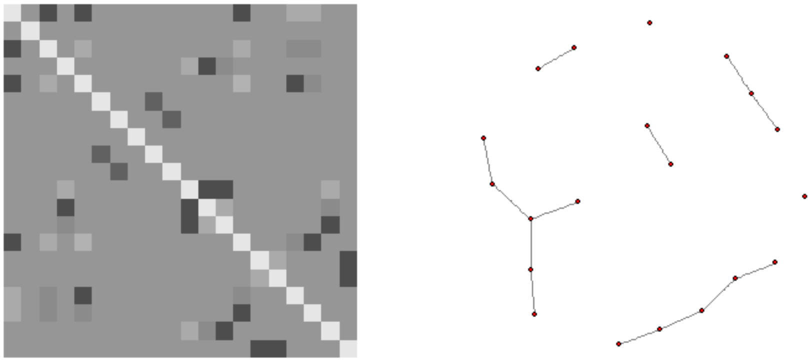


Figure 3.1: A covariance map and graph of **A**

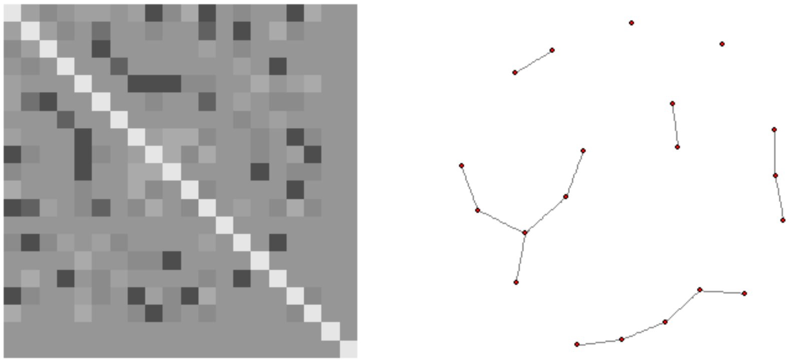


Figure 3.2: A covariance map and graph of **B**

Taking the Kronecker product, we obtain $\mathbf{\Sigma}$ which demonstrates some clustered connectivity.

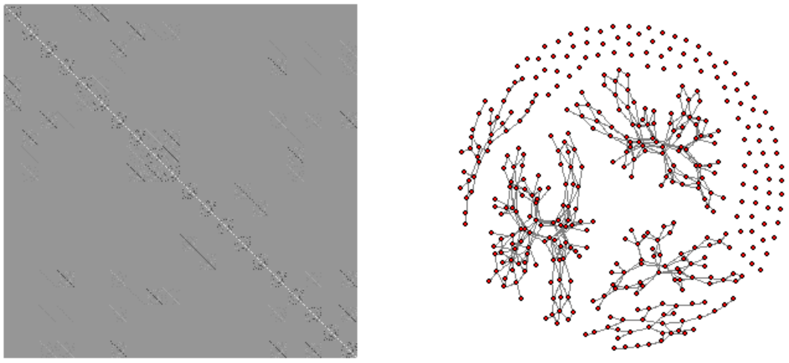


Figure 3.3: A covariance map and graph of $\mathbf{\Sigma} = \mathbf{A} \otimes \mathbf{B}$

We compared the performance of 4 different methods: The KGLasso, the naïve GLasso without consideration of the Kronecker covariance structure, the non-sparsified KGLasso based on the flip flop algorithm of the corresponding un-penalized maximum likelihood, and the CLIME by [Cai et al., 2011]. We evaluated the performance using Frobenius norm losses on the covariance and precision matrices for $n \in \{10, 25, 50, 100, 200, 400, 800, 1000\}$, and the tuning parameter λ was chosen experimentally and separately for each method to minimize Frobenius norm loss on the covariance matrix (Usually close to $\lambda = 0.4$). In the case of CLIME, λ was selected automatically using the `fastclime` package in R and its precision matrix estimator was further thresholded below to promote sparsity. The threshold cutoff was selected experimentally to minimize Frobenius norm loss on the covariance matrix. 100 trials were run for each method for each value of n.

A sample visualization of the graph produced by each method is given below for the case of $n = 100$.

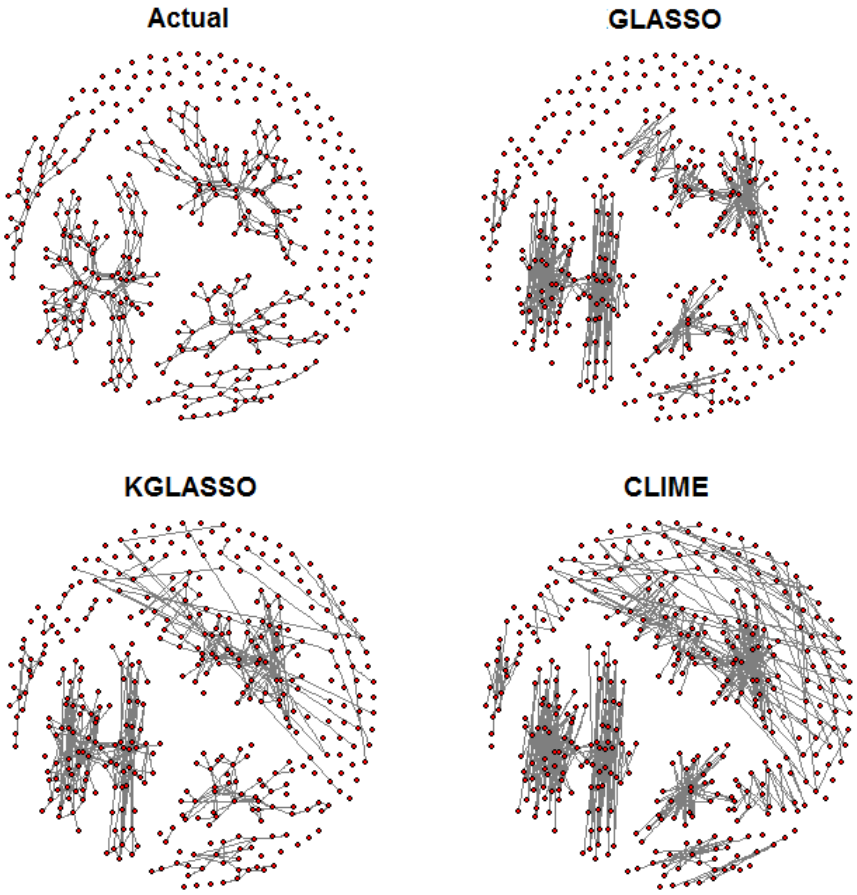


Figure 3.4: Graphs of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ from GLasso, KGLasso, and CLIME

From these graphs we can see that all 3 methods above are able to discriminate between clusters of connected nodes, although edge detection within those clusters is far from perfect. We can also observe that by not imposing a Kronecker product structure, the GLasso does not falsely detect many edges outside of clusters, in contrast to the more heavily connected upper hemisphere of the KGLasso graph. This is likely due to an edge being incorrectly detected in either of the component covariance matrices $\mathbf{A}$ or $\mathbf{B}$ and amplified through taking the Kronecker product. On the other hand, such structure enables KGLasso to more closely approximate networks within clusters than GLasso since information can be extrapolated outside of hubs.

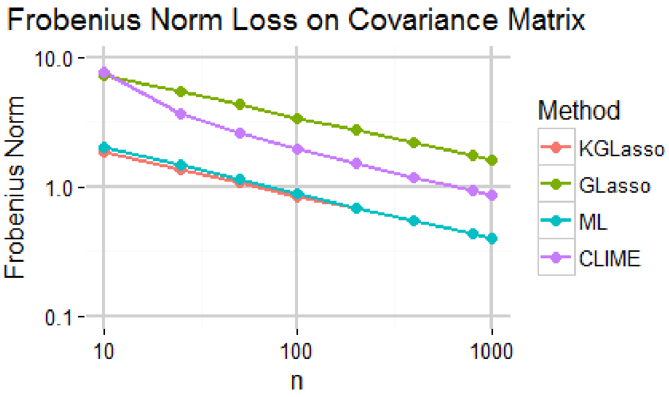


Figure 3.5: Frobenius norm loss for covariance matrix $\mathbf{\Sigma}$

From the Frobenius norm losses on the covariance matrix $\mathbf{\Sigma}$, we observe that the methods which assume a Kronecker covariance structure perform well. Both KGLasso and it's un-sparsified maximum likelihood form yield almost identical performance. KGLasso outperforms naïve GLasso and CLIME for all n considered above with respect to Frobenius norm loss on the covariance matrix.

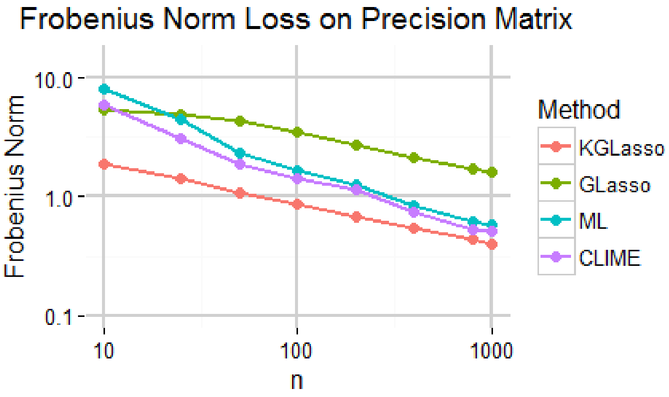


Figure 3.6: Frobenius norm loss for precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$

For precision matrix estimation, KGLasso outperforms the other methods as measured by Frobenius norm loss. Despite taking advantage of the underlying Kronecker structure, the un-sparsified maximum likelihood estimator is outperformed by CLIME and KGLasso, the former of which does not contain any structure decomposition information.

Method	Precision	Recall	Run Time (1000 runs)
KGLasso	0.683	0.763	71 seconds
GLasso	0.275	0.189	85 seconds
CLIME	0.224	0.908	610 seconds

Table 3.1: Precision, Recall, and Run Time for $n = 100$

We can see that for $n = 100$, KGLasso outperforms naïve GLasso in terms of precision and recall for edge detection. CLIME obtains better recall through over-selection and consequently yields poor precision. On average during our trials, CLIME selected over 3 times as many edges as KGLasso (1894 vs 535). The average number of edges in the underlying models was 320.

In terms of computational cost, both KGLasso and GLasso are very efficient. The overhead caused by the flip flop algorithm and dual problem appears to almost negate the complexity advantage of KGLasso for $p_A = p_B = 20$, but greater dimensionality would better demonstrate the advantages of KGLasso's simpler representation.

On the Selection of Tuning Parameters

The selection of tuning parameters λ_A and λ_B provides an interesting challenge for the Kronecker covariance model. Most commonly, the practice of selecting an appropriate λ involves optimizing a scalar c in $\lambda = cf(s, p)$ over a metric such as squared error loss or false discovery rate via cross-validation [Li et al., 2013]. However, the inter-connectivity of the Kronecker model complicates the partitioning of data into training and testing sets, which we will attempt to address in our real data example using the Joint Kronecker Graphical Lasso algorithm.

3.4 The Joint Kronecker Graphical Lasso for Multiple Groups

We consider the problem of joint covariance and precision matrix estimation in a high dimensional setting. In the traditional KGLasso setup, all observations are assumed to originate from the same Gaussian distribution. However, many datasets do not reflect such an assumption, and include observations which may come from different groups. For example, a genomics researcher may be studying gene co-expression networks in subjects with and without a certain gene. Those without the gene or those with an inactive copy will be lacking the respective hub in the co-expression networks versus those with an active gene, yet other parts of the networks should largely be the same. In section, we introduce an algorithm for obtaining joint covariance and precision matrix estimates and compare its performance via simulations and a real data example to

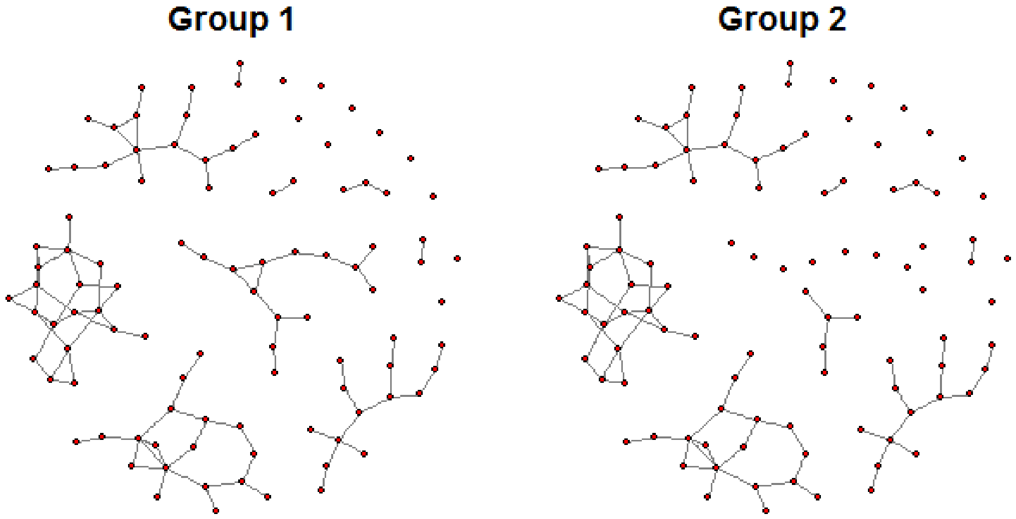


Figure 3.7: Two similar networks. Note that Group 2 differs from Group 1 in that a sub-network near the center of the image is present in Group 1 but not in Group 2.

other graphical methods including KGLasso applied separately to each group.

Given K sets of n_k i.i.d. multivariate Gaussian observations $\{\mathbf{x}_{jk}\}_{j=1}^{n_k}$, $\mathbf{x}_{jk} \in \mathbb{R}^p$ with mean zero (without loss of generality) and positive definite covariance matrix $\mathbf{\Sigma}_k \in S_{++}^p$, then the joint Kronecker graphical lasso (JKGLasso) covariance model is given by

$$\mathbf{\Sigma}_k = \mathbf{A}_k \otimes \mathbf{B}_k \quad k = 1, ..., K \tag{3.12}$$

and the log-likelihood takes form proportional to

$$\begin{aligned} l(\mathbf{\Sigma}) &= \sum_{k=1}^{K} \log\det\left(\mathbf{\Sigma}_k^{-1}\right) - \operatorname{tr}(\hat{\mathbf{S}}_k \mathbf{\Sigma}_k^{-1}) && (3.13) \\ &= \sum_{k=1}^{K} p_B \log\det\left(\mathbf{A}_k^{-1}\right) + p_A \log\det\left(\mathbf{B}_k^{-1}\right) - \operatorname{tr}\left[\hat{\mathbf{S}}_k(\mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1})\right] && (3.14) \end{aligned}$$

However, the maximum likelihood estimates yield poor performance in the sense that the they present high variance when $p \sim n_k$. In the high dimensional setting where $p > n_k$, the maximum likelihood estimates may be singular or ill-defined, although the lower dimensional Kronecker structure may alleviate this somewhat. Lastly, the lack of sparsity enforcement when using maximum likelihood makes $\hat{\mathbf{\Sigma}}_k^{-1}$ difficult to interpret, especially from a graphical standpoint, since all entries in $\hat{\mathbf{\Sigma}}_k^{-1}$ would be nonzero and the every pair of nodes would contain a connecting edge in its corresponding graph. Therefore, we consider a penalization scheme similar to the ordinary Kronecker Graphical Lasso. However, since the Joint Kronecker Graphical Lasso studies K groups with similar covariance matrices $\mathbf{\Sigma}_k$, we'd like to penalize the log-likelihood using a term that both promotes sparsity and similarity between the $\mathbf{\Sigma}_k^{-1}$ estimators.

The group Lasso penalty accomplishes these goals. The resulting objective function for the Joint Kronecker Graphical Lasso (JKGLasso) is convex and encourages both sparsity within $\mathbf{\Sigma}_k^{-1}$ and similarity between, and takes the form

$$f = \sum_{k=1}^{K} \left[\mathrm{tr}(\hat{\mathbf{S}}_k \mathbf{\Sigma}_k^{-1}) - \log\det\left(\mathbf{\Sigma}_k^{-1}\right) \right] + \left[\lambda_1 \sum_{k=1}^{K} |\mathbf{\Sigma}_k^{-1}|_1 + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{k=1}^{K} [(\mathbf{\Sigma}_k^{-1})_{i,j}]^2} \right] \tag{3.15}$$

$$= \sum_{k=1}^{K} \left[\mathrm{tr}(\hat{\mathbf{S}}_k \mathbf{\Sigma}_k^{-1}) - \log\det\left(\mathbf{\Sigma}_k^{-1}\right) + \lambda_1 |\mathbf{\Sigma}_k^{-1}|_1 \right] + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{k=1}^{K} [(\mathbf{\Sigma}_k^{-1})_{i,j}]^2} \tag{3.16}$$

where $i, j = 1, ..., p$, and $(\mathbf{\Sigma}_k^{-1})_{i,j}$ denotes the element from $\mathbf{\Sigma}_k^{-1}$ in row i, column j. Furthermore, the group Lasso penalty separates the objective function into two parts above, the left of which allows isolation by group k and is amenable to many fast convex optimization methods. Extending to the Kronecker covariance model, we obtain the objective function

$$f = \sum_{k=1}^{K} \left[\mathrm{tr}\left(\hat{\mathbf{S}}_k (\mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1}) \right) - \log\det\left(\mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1} \right) + \lambda_1 \left(|\mathbf{A}_k^{-1}|_1 |\mathbf{B}_k^{-1}|_1 \right) \right] \tag{3.17}$$

$$+ \lambda_2 \left[\sum_{i_A \neq j_A} \sqrt{\sum_{k=1}^{K} [(\mathbf{A}_k^{-1})_{i_A, j_A}]^2} + \sum_{i_B \neq j_B} \sqrt{\sum_{k=1}^{K} [(\mathbf{B}_k^{-1})_{iB, jB}]^2} \right] \tag{3.18}$$

where $i_A, i_B = 1, ..., p_A$ and $i_B, j_B = 1, ..., p_B$. Note that we impose the group Lasso penalty separately on $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ to reinforce similarity in $\mathbf{A}_k^{-1}$ across k and $\mathbf{B}_k^{-1}$ across k. While this naturally produces similarity in $\mathbf{\Sigma}^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$, such a penalty is not imposed directly on $\mathbf{\Sigma}^{-1}$ because $\sum_{i \neq j} \sqrt{\sum_{k=1}^{K} [(\mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1})_{i,j}]^2}$ is not jointly convex.

3.4.1 The JKGLasso Algorithm

The separability of the objective function f by Kronecker component matrices $\mathbf{A}$ and $\mathbf{B}$ yields the sub-expression containing $\mathbf{A}$:

$$f_A = \sum_{k=1}^{K} \left[\mathrm{tr}\left(\hat{\mathbf{S}}_k (\mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1}) \right) - p_B \log\det\left(\mathbf{A}_k^{-1}\right) + \lambda_1 |\mathbf{A}_k^{-1}|_1 \right] \tag{3.19}$$

$$+ \lambda_2 \sum_{i_A \neq j_A} \sqrt{\sum_{k=1}^{K} [(\mathbf{A}_k^{-1})_{i_A, j_A}]^2} \qquad i_A, j_A = 1, ..., p_A \tag{3.20}$$

and a corresponding sub-expression f_B for $\mathbf{B}$. Since both the negative log likelihood and penalty terms are bi-convex, the objective function f is bi-convex and iterative block coordinate-wise minimization of f with respect to $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$. Since the Kronecker covariance maximum likelihood estimator exists for $n \geq \max\left(\frac{p_A}{p_B}, \frac{p_B}{p_A}\right) + 1$ [Dutilleul, 1999], the objective function $f(\mathbf{A}^{-1}, \mathbf{B}^{-1})$ is bounded below. The iterative block coordinate-wise minimization of f yields a decreasing sequence over iterations, so the algorithm converges. Additionally, the algorithm converges to a local minimum on f and each iteration yields strict descent. The proof mirrors Theorem II.6 of [Tsiligkaridis, 2014].

Let, $\{\mathbf{A}^{-1}\} = \{\mathbf{A}_1^{-1}, ..., \mathbf{A}_K^{-1}\}$ and $\{\mathbf{B}^{-1}\} = \{\mathbf{B}_1^{-1}, ..., \mathbf{B}_K^{-1}\}$. The Joint Kronecker Graphical Lasso algorithm introduced above can be summarized by:

```
Input $\hat{\mathbf{S}_k}, \lambda_1 > 0, \lambda_2 > 0, \epsilon > 0$
Initialize $\hat{\mathbf{\Sigma}}^{-1} = \mathbf{I}_p$ and $\mathbf{A}^{-1} \in S_{++}^{p_A}$
repeat
    $\hat{\mathbf{\Sigma}}_{k,prev}^{-1} \leftarrow \hat{\mathbf{\Sigma}}_k^{-1} \quad k = 1, ..., K$
    $\{\mathbf{B}^{-1}\} \leftarrow \arg\min_{\{\mathbf{B}^{-1}\}} f(\{\mathbf{A}^{-1}\}, \{\mathbf{B}^{-1}\}) = \arg\min_{\{\mathbf{B}^{-1}\}} f_B(\{\mathbf{A}^{-1}\}, \{\mathbf{B}^{-1}\})$
    $\{\mathbf{A}^{-1}\} \leftarrow \arg\min_{\{\mathbf{A}^{-1}\}} f(\{\mathbf{A}^{-1}\}, \{\mathbf{B}^{-1}\}) = \arg\min_{\{\mathbf{A}^{-1}\}} f_B(\{\mathbf{A}^{-1}\}, \{\mathbf{B}^{-1}\})$
    $\hat{\mathbf{\Sigma}}_k^{-1} \leftarrow \mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1} \quad k = 1, ..., K$
until $\sum_{k=1}^{K} \|\hat{\mathbf{\Sigma}}_k^{-1} - \hat{\mathbf{\Sigma}}_{k,prev}^{-1}\| \leq \epsilon$
```

Each update of $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ is handled using the Alternating Directions Method of Multipliers (ADMM) algorithm introduced by [Boyd et al., 2010].

The ADMM approach rewrites f_A as a Lagrangian dual problem by adding Lagrange multipliers to handle additional constraints. Introducing k additional variables $\mathbf{Z}_k$, $k = 1, ..., K$ subject to $\mathbf{Z}_k = \mathbf{A}_k$ allows separation of the log-likelihood of individual groups from the group penalty so that each group can be handled individually in the primal problem. We minimize

$$g_A = \sum_{k=1}^{K} \left[\operatorname{tr}\left(\hat{\mathbf{S}}_k (\mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1}) \right) - p_B \log\det\left(\mathbf{A}_k^{-1}\right) \right] \tag{3.21}$$

$$+ \lambda_1 \sum_{k=1}^{K} |\mathbf{Z}_k^{-1}|_1 + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{k=1}^{K} [(\mathbf{Z}_k^{-1})_{i,j}]^2} \qquad i,j = 1, ..., p_A \tag{3.22}$$

subject to positive definiteness $\mathbf{A}_k \in S_{++}^{p}$ and $\mathbf{Z}_k = \mathbf{A}_k$ for $k = 1, ..., K$. The corresponding Lagrangian takes the form

$$L(\{\mathbf{A}^{-1}\}, \{\mathbf{Z}^{-1}\}, \{\mathbf{U}\}) = \sum_{k=1}^{K} \left[\operatorname{tr}\left(\hat{\mathbf{S}}_k (\mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1}) \right) - p_B \log\det\left(\mathbf{A}_k^{-1}\right) \right] \tag{3.23}$$

$$+ \lambda_1 \sum_{k=1}^{K} |\mathbf{Z}_k^{-1}|_1 + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{k=1}^{K} [(\mathbf{Z}_k^{-1})_{i,j}]^2} \qquad i,j = 1, ..., p_A \tag{3.24}$$

$$+ \rho \sum_{k=1}^{K} |\mathbf{A_k}^{-1} - \mathbf{Z}_k^{-1} + \mathbf{U}_k|_F^2 \tag{3.25}$$

where $\mathbf{U}_k$ is the dual variable. Let $\mathbf{A}_{k(i)}^{-1}, \mathbf{Z}_{k(i)}, \mathbf{U}_{k(i)}$ denote the i-th iteration of $\mathbf{A}_k^{-1}, \mathbf{Z}_k, \mathbf{U}_k$, respectively. Let $\{\mathbf{A}_{(i)}^{-1}\}, \{\mathbf{Z}_{(i)}\}, \{\mathbf{U}_{(i)}\}$ denote the i-th iteration of $\{\mathbf{A}^{-1}\}, \{\mathbf{Z}\}, \{\mathbf{U}\}$, respectively. Each iteration, $\{\mathbf{A}^{-1}\}$ and $\{\mathbf{Z}^{-1}\}$ are updated sequentially to minimize the Lagrangian until the duality gap closes and the ADMM algorithm converges so the resulting estimators $\{\hat{\mathbf{A}}\}$ can be returned to the flip stop step.

To update $\{\mathbf{A}_{(i)}^{-1}\} = \arg\min_{\{\mathbf{A}^{-1}\}} L(\{\mathbf{A}^{-1}\}, \{\mathbf{Z}_{(i-1)}^{-1}\}, \{\mathbf{U}_{(i-1)}\})$, we minimize the relevant parts of the Lagrangian containing $\mathbf{A}_k^{-1}$

$$L_A(\{\mathbf{A}^{-1}\}, \{\mathbf{Z}^{-1}\}, \{\mathbf{U}\}) = \sum_{k=1}^{K} \left[\operatorname{tr}\left(\hat{\mathbf{S}}_k (\mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1}) \right) - p_B \log\det\left(\mathbf{A}_k^{-1}\right) \right] \tag{3.26}$$

$$+ \rho \sum_{k=1}^{K} |\mathbf{A}_k^{-1} - \mathbf{Z}_k^{-1} + \mathbf{U}_k|_F^2 \tag{3.27}$$

Since each term corresponding to a unique k is separable, the solution amounts to K separate, Frobenius-penalized ordinary Graphical Lasso problems which can be solved using the eigen-decomposition approach of [Witten and Tibshirani, 2009] and [Allen and Tibshirani, 2010].

We have the scaled augmented Lagrangian for some fixed k,

$$\left[\operatorname{tr}\left(\hat{\mathbf{S}}_k (\mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1}) \right) - p_B \log\det\left(\mathbf{A}_k^{-1}\right) \right] p_B^{-1} + \rho n_k^{-1} |\mathbf{A}_k^{-1} - \mathbf{Z}_k^{-1} + \mathbf{U}_k|_F^2 \tag{3.28}$$

Let $\hat{\mathbf{S}}_{[i,j]} \in \mathbb{R}^{p_B \times p_B}$ denote the sub-matrix in $\hat{\mathbf{S}}$ corresponding to block row i and block column j where each block is $p_A \times p_A$, and let $\mathbf{C}_k \in \mathbb{R}^{p_A \times p_A}$ denote the matrix whose elements are $(\mathbf{C}_k)_{i,j} = \mathrm{tr}(\hat{\mathbf{S}}_{[i,j]}\mathbf{B}_k)$ for $i = 1, ..., p_B \;\; j = 1, ..., p_B$. Then, we have the first order equation

$$0 = \mathbf{C}n_k - p_B n_k \mathbf{A}_k + \rho(\mathbf{A}_k^{-1} - \mathbf{Z}_k^{-1} + \mathbf{U}_k) \tag{3.29}$$

$$= \frac{\rho}{p_B n_k}\mathbf{A}_k^{-2} + \left(\frac{C}{p_B} - \frac{\rho \mathbf{Z}_k^{-1}}{p_b n_k} + \frac{\rho \mathbf{U}_k}{p_B n_k}\right)\mathbf{A}_k^{-1} - \mathbf{I}_{p_A} \tag{3.30}$$

Let $\mathbf{V}_k\mathbf{D}_k\mathbf{V}_k^\top$ be the eigen-decomposition of $\dfrac{C}{p_B} - \dfrac{\rho \mathbf{Z}_k^{-1}}{p_b n_k} + \dfrac{\rho \mathbf{U}_k}{p_B n_k}$.

Then the solution to the first order equation is given by $\hat{\mathbf{A}}_k^{-1} = \mathbf{V}_k\hat{\mathbf{D}}_k{\mathbf{V}_k}^\top$, where $\hat{\mathbf{D}}_k$ is a diagonal matrix with entries

$$(\hat{\mathbf{D}}_k)_{j,j} = \frac{p_B n_k}{2\rho}\left(-(\mathbf{D}_k)_{j,j} + \sqrt{(\mathbf{D}_k)_{j,j}^2 + \frac{4\rho}{p_B n_k}}\right) \tag{3.31}$$

The Kronecker structure is particularly important in this solution since the eigen-decomposition of a large $p \times p$ matrix can be very expensive compared to the eigen-decompositions of several much smaller matrices of sizes $p_A \times p_A$ and $p_B \times p_B$. [Witten et al., 2011]

To update $\{\mathbf{Z}_{(i)}\} = \arg\min_{\{\mathbf{Z}\}} L(\{\mathbf{A}_{(i)}^{-1}\}, \{\mathbf{Z}^{-1}\}, \{\mathbf{U}_{(i-1)}\})$, we minimize the relevant parts of the Lagrangian containing $\mathbf{Z}_k$

$$L(\{\mathbf{A}^{-1}\}, \{\mathbf{Z}\}, \{\mathbf{U}\}) = \lambda_1 \sum_{k=1}^{K} |\mathbf{Z}_k^{-1}|_1 + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{k=1}^{K} [(\mathbf{Z}_k^{-1})_{i,j}]^2} \qquad i, j = 1, ..., p_A \tag{3.32}$$

$$+ \rho \sum_{k=1}^{K} |{\mathbf{A_k}}^{-1} - \mathbf{Z}_k^{-1} + \mathbf{U}_k|_F^2 \tag{3.33}$$

which is the ordinary group Lasso penalty and has solution:

$$(\hat{\mathbf{Z}}_k^{-1})_{i,j} = \begin{cases} (\hat{\mathbf{A}}_k^{-1})_{i,j} & i = j \\ S\left((\hat{\mathbf{A}}_k^{-1})_{i,j}, 2\lambda_1/\rho\right)\left(1 - \dfrac{\lambda_2}{2\rho\sqrt{\sum_{k=1}^{K} S\left((\hat{\mathbf{A}}_k^{-1})_{i,j}, 2\lambda_1/\rho\right)^2}}\right) & \end{cases} \tag{3.34}$$

The dual variable update is trivial. $\{\mathbf{U}_{(i)}\} = \{\mathbf{A}_{(i)}^{-1}\} - \{\mathbf{Z}_{(i)}^{-1}\} + \{\mathbf{U}_{(i-1)}\}$

3.4.2 Simulation Results

In this subsection, we compare simulation performances of the Joint Kronecker Graphical Lasso (JK-GLasso) with several other methods including naïve Graphical Lasso (GLasso), the Kronecker Graphical Lasso (KGLasso) where all subjects are considered to come from a single group, the Joint Graphical Lasso (JGLasso) where the Kronecker covariance assumption is withheld, universal thresholding (Thresh), the constrained ℓ_1-minimization for inverse matrix estimation (CLIME) [Cai et al., 2011], the neighborhood pursuit method (Neigh) of [Meinshausen and Bühlman, 2006], and the time-varying ridge method introduced in Chapter 2 (TV-Ridge).

We evaluated the performance of these methods using precision and recall of edge selection. For methods which provide inverse covariance matrix estimators (Excludes Neigh and TV-Ridge), we compared total Frobenius norm error loss on $\mathbf{\Sigma}_k$ and $\mathbf{\Sigma}_k^{-1}$, summed across groups.

In our simulations, we considered setups with $K = 3$ and 50 groups. We generated $\mathbf{\Sigma}_k^{-1} = \mathbf{A}_k^{-1} \otimes \mathbf{B}_k^{-1}$ where $\mathbf{A}_k^{-1}$ and $\mathbf{B}_k^{-1}$ are 100×100 inverse covariance matrices. $\mathbf{A}_1^{-1}$ and $\mathbf{B}_1^{-1}$ were generated using the cluster technique from the high dimensional undirected graph estimation package (`huge`) in R. Subsequent groups were generated using the previous group with uniform probability to add a cluster, remove a cluster, or remain unchanged. The record of group ordering was kept and treated as a time variable for the TV-Ridge method. A previous example is given again below to illustrate.

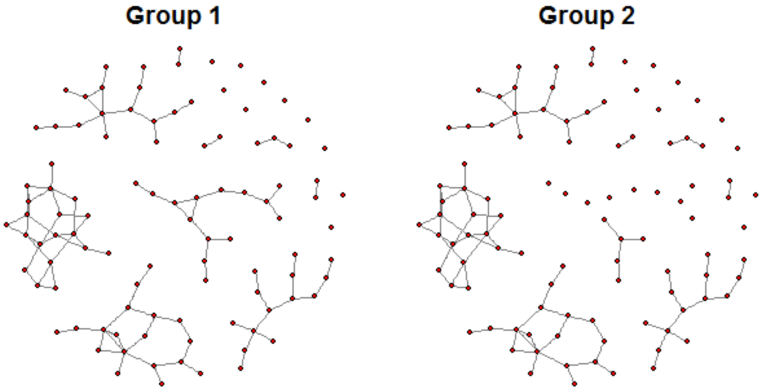


Figure 3.8: Graphs of $\mathbf{A}_1^{-1}$ and $\mathbf{A}_2^{-1}$

The KJGLasso algorithm generally converged within 1000 iterations. The example below shows the behavior of the overall objective function through iterations. Spikes in the plot correspond to the initialization stages of the ADMM algorithm. Behavior between spikes is characterized by the ADMM optimization routine itself. Behavior of the spikes is characterized by the flip-flop routine.

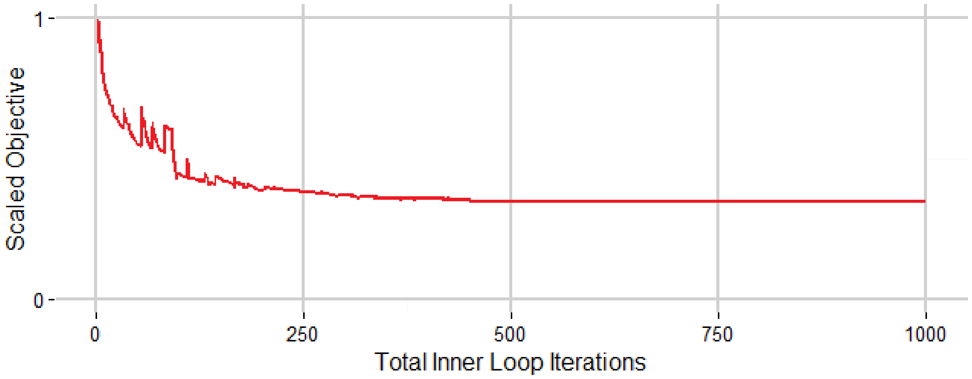


Figure 3.9: The scaled objective function over the JKGLasso algorithm's inner loop iterations

Appendix A

Proofs

A.1 Lemmas

Lemma A.1. *Let X be an $n \times p$ matrix and $D = diag(d_1, \cdots, d_n)$ with $|d_i| \leq b$ and $b \geq 0$. Then*

$$\rho_{\max}(X^\top D X, s) \leq 2b\rho_{\max}(X^\top X, s).$$

If $d_i \in [0, b]$, then $\rho_{\max}(X^\top D X, s) \leq b\rho_{\max}(X^\top X, s)$.

Proof. Let $\mathcal{A}_s = \{\mathbf{a} \in \mathbb{R}^p : |\mathbf{a}|_2 \leq 1, |\mathbf{a}|_0 \leq s\}$. Write $d_i = d_i^+ - d_i^-$, where $d_i^+ = \max(d_i, 0)$ and $d_i^- = \max(-d_i, 0)$ are the positive and negative parts, respectively. By definition

$$\begin{aligned}
\rho_{\max}(X^\top D X, s) &= \max_{\mathbf{a} \in \mathcal{A}_s} |\mathbf{a}^\top X^\top D X \mathbf{a}| = \max_{\mathbf{a} \in \mathcal{A}_s} |\mathrm{tr}(D(X\mathbf{a}\mathbf{a}^\top X^\top))| \\
&= \max_{\mathbf{a} \in \mathcal{A}_s} \left| \sum_{i=1}^{n} (d_i^+ - d_i^-)(X\mathbf{a}\mathbf{a}^\top X^\top)_{ii} \right| \leq 2b \max_{\mathbf{a} \in \mathcal{A}_s} \sum_{i=1}^{n} (X\mathbf{a}\mathbf{a}^\top X^\top)_{ii} \\
&= 2b \max_{\mathbf{a} \in \mathcal{A}_s} \mathrm{tr}(X\mathbf{a}\mathbf{a}^\top X^\top) = 2b \max_{\mathbf{a} \in \mathcal{A}_s} \mathbf{a}^\top X^\top X \mathbf{a} = 2b\rho_{\max}(X^\top X, s),
\end{aligned}$$

because $X^\top X$ is nonnegative definite. The second claim follows from the same lines with $d_i^- = 0$. □

Lemma A.2. *Let $t \in \varpi$ and $\hat{\Sigma}_t$ be the kernel smoothed sample covariance at time t and $\hat{\Sigma}_t^\diamond = \mathcal{X}_t^{\diamond\top} \mathcal{X}_t^\diamond$. Suppose that $\mathcal{X}_t^\diamond$ has full row rank. Assume further (2.15), (2.13) and assumption 6 hold, then we have*

$$\rho_{\min\neq 0}(\hat{\Sigma}_t) \geq |N_t| \underline{w}_t \varepsilon_0^2 \tag{A.1}$$

$$\rho_{\max}(\hat{\Sigma}_t, s) \leq |N_t| \overline{w}_t \varepsilon_0^{-2}. \tag{A.2}$$

Proof. Since $\mathcal{X}_t^\diamond$ is of full row rank, $r = |N_t|$. Note that $\mathcal{X}_t = (|N_t|W_t)^{1/2}\mathcal{X}_t^\diamond$, $\rho_i(\hat{\Sigma}_t) = \sigma_i^2(\mathcal{X}_t)$ and $\rho_i(\hat{\Sigma}_t^\diamond) = \sigma_i^2(\mathcal{X}_t^\diamond)$. By the generalized Marshall-Olkin inequality, see e.g. [Wang and Zhang, 1992, Theorem 4], assumption 6 and (2.15), we have

$$\begin{aligned}\rho_{\min\neq 0}(\hat{\Sigma}_t) &= \rho_{\min}(\mathcal{X}_t\mathcal{X}_t^\top) = |N_t|\rho_{\min}(W_t^{1/2}\mathcal{X}_t^\diamond\mathcal{X}_t^{\diamond\top}W_t^{1/2}) \\ &= |N_t|\rho_{\min}(\mathcal{X}_t^\diamond\mathcal{X}_t^{\diamond\top}W_t) \geq |N_t|\rho_{\min}(W_t)\rho_{\min}(\mathcal{X}_t^\diamond\mathcal{X}_t^{\diamond\top}) \geq |N_t|\underline{w}_t\varepsilon_0^2.\end{aligned}$$

The second inequality (A.2) follows from assumption 3(b) and Lemma A.1 applying to $\hat{\Sigma}_t = |N_t|\mathcal{X}_t^{\diamond\top}W_t\mathcal{X}_t^\diamond$ and $W_t \geq 0$. □

Lemma A.3. *Suppose assumption 1, 2, 3 and 5(a) hold. Let $t \in \varpi$ be fixed and λ_0 be defined in (2.17). Then, for $\lambda_1 \geq 2(\lambda_0 + 2C_0L_{t,1}s^{1/2}\varepsilon_0^{-2}b_n|N_t|\overline{w}_t)$ where λ_0 is defined in (2.17), we have, with probability $1-2p^{-1}$,*

$$|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}(t)]|_2^2 + \lambda_1|\tilde{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}(t)|_1 \leq 4\lambda_1^2\frac{s}{\phi_0^2}. \tag{A.3}$$

Proof. By definition (2.8),

$$|\mathcal{Y}_t - \mathcal{X}_t\tilde{\boldsymbol{\beta}}(t)|_2^2 + \lambda_1|\tilde{\boldsymbol{\beta}}(t)|_1 \leq |\mathcal{Y}_t - \mathcal{X}_t\boldsymbol{\beta}(t)|_2^2 + \lambda_1|\boldsymbol{\beta}(t)|_1,$$

which implies that

$$|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}(t)]|_2^2 + \lambda_1|\tilde{\boldsymbol{\beta}}(t)|_1 \leq \lambda_1|\boldsymbol{\beta}(t)|_1 + 2\left\langle \mathcal{Y}_t - \mathcal{X}_t\boldsymbol{\beta}(t), \mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}(t)]\right\rangle.$$

By assumption 2 and Taylor's expansion in the b_n-neighborhood of t, we see that

$$\mathcal{Y}_t - \mathcal{X}_t\boldsymbol{\beta}(t) = \mathcal{E}_t + M_t\mathcal{X}_t\boldsymbol{\beta}'(t) + \mathcal{X}_t\boldsymbol{\xi}, \tag{A.4}$$

where $M_t = \mathrm{diag}((t_i-t)_{i\in N_t})$ and $\boldsymbol{\xi}$ is a vector such that $|\boldsymbol{\xi}|_\infty \leq C_0b_n^2/2$ and $|\boldsymbol{\xi}|_0 \leq s$. Let $\mathcal{J} = \{2|\mathcal{E}_t^\top\mathcal{X}_t|_\infty \leq \lambda_0\}$. Observe that $|\mathcal{E}_t^\top\mathcal{X}_t|_\infty = \max_{j\leq p}|\sum_{i\in N_t} w(t,i)X_{ij}e_i|$ and, by assumption 1,

$$\sum_{i\in N_t} w(t,i)X_{ij}e_i \sim N\left(0, \sigma^2\sum_{i\in N_t} w(t,i)^2X_{ij}^2\right). \tag{A.5}$$

Then, by the standard Gaussian tail bound and the union inequality, we obtain that

$$\mathbb{P}\left(\max_{j\leq p}\left|\frac{\sum_{i\in N_t} w(t,i)X_{ij}e_i}{\sigma L_{t,2}}\right| \geq \sqrt{\varepsilon^2 + 2\log p}\right) \leq \mathbb{P}(\max_{j\leq p}|Z_j| \geq \sqrt{\varepsilon^2+2\log p}) \leq 2\exp\left(-\frac{\varepsilon^2}{2}\right)$$

for all $\varepsilon > 0$, where $Z_j \sim N(0,1)$. Now, choose $\varepsilon = (2\log p)^{1/2}$ and $\lambda_0 = 4\sigma L_{t,2}(\log p)^{1/2}$, we have $\mathbb{P}(\mathcal{J}) \geq 1 - 2p^{-1}$. Further, we have

$$\begin{aligned}
& |\boldsymbol{\beta}'(t)^\top \mathcal{X}_t^\top M_t \mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]| \leq |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 |\mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t)|_\infty \\
\leq \quad & |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 \max_{j \leq p} \left(\sum_{i \in N_t} w(t,i) X_{ij}^2 \right)^{1/2} \left[\boldsymbol{\beta}'(t)^\top \mathcal{X}_t^\top M_t^2 \mathcal{X}_t \boldsymbol{\beta}'(t)\right]^{1/2} \quad \text{(Cauchy-Schwarz)} \\
\leq \quad & |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 L_{t,1} \sqrt{\rho_{\max}(\mathcal{X}_t^\top M_t^2 \mathcal{X}_t, s)} |\boldsymbol{\beta}'(t)|_2 \quad \text{(assumption 2)} \\
\leq \quad & |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 L_{t,1} C_0 s^{1/2} b_n \sqrt{\rho_{\max}(\mathcal{X}_t^\top \mathcal{X}_t, s)} \quad \text{(Lemma A.1, assumption 2 and 3)} \\
\leq \quad & |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 L_{t,1} C_0 (|N_t| \overline{w}_t s)^{1/2} b_n \varepsilon_0^{-1} \quad \text{(Lemma A.2, equation (A.2)).}
\end{aligned}$$

Similarly, we can show that $|\boldsymbol{\xi}^\top \mathcal{X}_t^\top \mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]| = O(L_{t,1}(|N_t|\overline{w}_t s)^{1/2} b_n^2 \varepsilon_0^{-1} |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1)$. Therefore, it follows that, with probability at least $(1 - 2p^{-1})$,

$$\left| \left\langle \mathcal{Y}_t - \mathcal{X}_t \boldsymbol{\beta}(t), \mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)] \right\rangle \right| \leq \left[\lambda_0 + 2L_{t,1} C_0 (|N_t| \overline{w}_t s)^{1/2} b_n \varepsilon_0^{-1} (1 + o(1)) \right] |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1.$$

Now, choose $\lambda_1 \geq 2(\lambda_0 + 2L_{t,1} C_0 (|N_t| \overline{w}_t s)^{1/2} b_n \varepsilon_0^{-1})$, we get

$$2|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 + 2\lambda_1 |\tilde{\boldsymbol{\beta}}(t)|_1 \leq \lambda_1 |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 + 2\lambda_1 |\boldsymbol{\beta}(t)|_1.$$

Denote $S_0 := S_0(t) = \text{supp}(\boldsymbol{\beta}(t))$. By the same argument as [Bühlmann and van de Geer, 2011, Lemma 6.3], it is easy to see that, on $\mathcal{J}$,

$$2|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 + \lambda_1 |\tilde{\boldsymbol{\beta}}_{S_0^c}(t)|_1 \leq 3\lambda_1 |\tilde{\boldsymbol{\beta}}_{S_0}(t) - \boldsymbol{\beta}_{S_0}(t)|_1.$$

But then, (A.3) follows from the restricted eigenvalue condition (assumption 4) with the elementary inequality $4ab \leq a^2 + 4b^2$ that

$$2|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 + \lambda_1 |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 \leq 4\lambda_1 |\tilde{\boldsymbol{\beta}}_{S_0}(t) - \boldsymbol{\beta}_{S_0}(t)|_1 \leq |\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 + 4\lambda_1^2 s / \phi_0^2.$$

□

Lemma A.4. *Let ξ_i be a sub-Gaussian random variable with mean zero and variance factor σ^2, and $e_i = \sum_{m=0}^{\infty} a_m \xi_{i-m}$ be a linear process. Let $\mathbf{w} = (w_1, \cdots, w_n)$ be a real vector and $S_n = \sum_{i=1}^{n} w_i e_i$ be the weighted partial sum of e_i.*

1. *(Short-range dependence). If $|\mathbf{a}|_1 = \sum_{i=0}^{\infty} |a_i| < \infty$, then for all $x > 0$ we have*

$$\mathbb{P}(|S_n| \geq x) \leq 2 \exp\left(-\frac{x^2}{2|\mathbf{w}|^2 |\mathbf{a}|_1^2 \sigma^2} \right). \tag{A.6}$$

2. *(Long-range dependence). Suppose $K = \sup_{m \geq 0} |a_m|(m+1)^{\varrho} < \infty$, where $1/2 < \varrho < 1$. Then, there exists a constant C_ϱ that only depends on ϱ such that*

$$\mathbb{P}(|S_n| \geq x) \leq 2 \exp\left(-\frac{C_\varrho x^2}{|\mathbf{w}|^2 n^{2(1-\varrho)} \sigma^2 K^2} \right). \tag{A.7}$$

Proof. Put $a_m = 0$ if $m < 0$ and we may write $S_n = \sum_{m \in \mathbb{Z}} b_m \xi_m$, where $b_m = \sum_{i=1}^{n} w_i a_{i-m}$. By the Cauchy-Schwartz inequality,

$$\sum_{m \in \mathbb{Z}} b_m^2 \leq \sum_{m \in \mathbb{Z}} \left(\sum_{i=1}^{n} w_i^2 |a_{i-m}| \right) \left(\sum_{i=1}^{n} |a_{i-m}| \right) \leq |\mathbf{w}|^2 |\mathbf{a}|_1^2.$$

Then, (A.6) follows from the Cramér-Chernoff bound [Boucheron et al., 2013]. Let $\bar{a}_m = \max_{l \geq m} |a_l|$ and $A_m = \sum_{l=0}^{m} |a_l|$. Note that $A_n \leq K \sum_{l=0}^{n} (l+1)^{-\varrho} \leq C_\varrho K (n+1)^{1-\varrho}$, where $C_\varrho = (1-\varrho)^{-1}$. Then, we have

$$\sum_{m=1-n}^{n} b_m^2 \leq \sum_{m=1-n}^{n} \left(\sum_{i=1}^{n} w_i^2 |a_{i-m}| \right) \left(\sum_{i=1}^{n} |a_{i-m}| \right) \leq |\mathbf{w}|^2 A_{2n}^2.$$

If $m \leq -n$, then $|b_m| \leq |\mathbf{w}|_1 \bar{a}_{1-m}$ and therefore

$$\sum_{m \leq -n} b_m^2 \leq |\mathbf{w}|_1^2 \sum_{m \leq -n} \bar{a}_{1-m}^2 \leq C_\varrho n |\mathbf{w}|^2 K^2 n^{1-2\varrho},$$

where the last inequality follows from Karamata's theorem; see e.g. [Resnick, 1987]. Hence, the proof is complete by invoking the Cramér-Chernoff bound for sub-Gaussian random variables. □

A.2 Proof of Theorem 2.1

Observe that $\mathcal{X}_t\boldsymbol{\beta}(t) = \mathcal{X}_t\boldsymbol{\theta}(t)$ since $\boldsymbol{\theta}(t) = P_{\mathcal{R}_t}\boldsymbol{\beta}(t)$. Using the closed-form formulae for the tv-ridge estimator (2.9) and by (A.4), we have

$$\text{bias}(\tilde{\boldsymbol{\theta}}(t)) = \mathbb{E}(\tilde{\boldsymbol{\theta}}(t)) - \boldsymbol{\theta}(t) = (\mathcal{X}_t^\top\mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top[\mathcal{X}_t\boldsymbol{\theta}(t) + M_t\mathcal{X}_t\boldsymbol{\beta}'(t) + \mathcal{X}_t\boldsymbol{\xi}] - \boldsymbol{\theta}(t), \tag{A.8}$$

where $|\boldsymbol{\xi}|_\infty \le C_0 b_n^2/2$ and $|\boldsymbol{\xi}|_0 \le s$ almost surely $t \in \varpi$. First, we bound the shrinkage bias of the tv-ridge estimator. By the argument in Section 3 of [Shao and Deng, 2012], we can show that

$$(\mathcal{X}_t^\top\mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top\mathcal{X}_t\boldsymbol{\theta}(t) - \boldsymbol{\theta}(t) = -Q(\lambda_2^{-1}D^2 + I_r)^{-1}Q^\top\boldsymbol{\theta}(t).$$

Then, it follows from Lemma A.2 that

$$\begin{aligned} |Q(\lambda_2^{-1}D^2 + I_r)^{-1}Q^\top\boldsymbol{\theta}(t)|_2 &\le \frac{|\boldsymbol{\theta}(t)|_2}{\rho_{\min}(\lambda_2^{-1}D^2 + I_r)} \\ &= \left(\frac{\lambda_2}{\lambda_2 + \min_{j\le r} d_j^2}\right)|\boldsymbol{\theta}(t)|_2 \le \frac{\lambda_2|\boldsymbol{\theta}(t)|_2}{\rho_{\min\neq 0}(\hat{\Sigma}_t)} \le \frac{\lambda_2|\boldsymbol{\theta}(t)|_2}{|N_T|\underline{w}_t\varepsilon_0^2}, \end{aligned} \tag{A.9}$$

where $d_j^2 = \rho_j(\hat{\Sigma}_t), j = 1, \cdots, r$. Next, we deal with the non-stationary bias of the tv-ridge estimator (A.8) by a similar argument for (A.9). Indeed, let $Q_\perp$ be the orthogonal complement of Q such that $Q_\perp^\top Q_\perp = I_{p-r}$ and $Q_\perp^\top Q = \mathbf{0}_{(p-r)\times r}$. Let $\Gamma = [Q; Q_\perp]$; then, $\Gamma\Gamma^\top = \Gamma^\top\Gamma = I_p$. By the SVD of $\mathcal{X}_t$, equation (2.5), we may write

$$\begin{aligned} &(\mathcal{X}_t^\top\mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top M_t\mathcal{X}_t\boldsymbol{\beta}'(t) = \Gamma\left(\Gamma^\top(QD^2Q^\top + \lambda_2 I_p)\Gamma\right)^{-1}\Gamma^\top\mathcal{X}_t^\top M_t\mathcal{X}_t\boldsymbol{\beta}'(t) \\ &= [Q; Q^\perp]\left(\begin{bmatrix} Q^\top \\ Q_\perp^\top \end{bmatrix}(QD^2Q^\top + \lambda_2 I_p)[Q; Q_\perp]\right)^{-1}\begin{bmatrix} Q^\top \\ Q_\perp^\top \end{bmatrix} QDP^\top M_t\mathcal{X}_t\boldsymbol{\beta}'(t) \\ &= [Q; Q^\perp]\begin{pmatrix} (D^2 + \lambda_2 I_r)^{-1} & \mathbf{0} \\ \mathbf{0} & \lambda_2^{-1}I_{p-r} \end{pmatrix}\begin{bmatrix} DP^\top M_t\mathcal{X}_t\boldsymbol{\beta}'(t) \\ \mathbf{0} \end{bmatrix} \\ &= Q(D + \lambda_2 D^{-1})^{-1}P^\top M_t\mathcal{X}_t\boldsymbol{\beta}'(t). \end{aligned}$$

Hence, by Lemma A.2 we have

$$|(\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t)|_2 \leq \frac{b_n|\mathcal{X}_t\boldsymbol{\beta}'(t)|_2}{\rho_{\min}(D+\lambda_2 D^{-1})} \leq \frac{C_0 b_n (s|N_t|\overline{w}_t)^{1/2}\varepsilon_0^{-1}}{\min_{j\leq r}(d_j+\lambda_2/d_j)},$$

where $\overline{w}_t = \sup_{i\in N_t} w(i,t)$. Since $\lambda_2 = o(1)$ and $d_j \geq (|N_t|\underline{w}_t)^{1/2}\varepsilon_0$, the denominator of last expression is lower bounded by $[(|N_t|\underline{w}_t)^{1/2}\varepsilon_0 + \lambda_2/((|N_t|\underline{w}_t)^{1/2}\varepsilon_0)]$ for large enough n. Therefore, we have

$$|(\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t)|_2 \leq \frac{C_0 b_n (s|N_t|\overline{w}_t)^{1/2}}{(|N_t|\underline{w}_t)^{1/2}\varepsilon_0^2} \leq \frac{C_0 b_n s^{1/2}}{C\varepsilon_0^2}. \tag{A.10}$$

Similarly, upper bound for the remainder term of (A.8) can be established as follows

$$|(\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\xi}|_2 \leq \frac{C_0 b_n^2 s^{1/2}}{2C\varepsilon_0^2}, \quad \text{for almost surely } t \in \varpi. \tag{A.11}$$

In addition, $\tilde{\boldsymbol{\theta}}(t) - \mathbb{E}[\tilde{\boldsymbol{\theta}}(t)] = (\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top \mathcal{E}_t$ is the stochastic part of the tv-ridge estimator. Since $e_i \sim N(0, \sigma^2 I_n)$ are i.i.d., $\mathcal{E}_t \sim N(\mathbf{0}, \sigma^2 W_t)$. Hence, $\tilde{\boldsymbol{\theta}}(t) - \mathbb{E}[\tilde{\boldsymbol{\theta}}(t)] \sim N(\mathbf{0}, \Omega(\lambda))$, where $\Omega(\lambda)$ is defined in (2.7), and thus

$$\mathrm{Var}(\tilde{\theta}_j(t)) = \sigma^2\Omega_{jj}(\lambda_2) \geq \sigma^2\Omega_{\min}(\lambda_2). \tag{A.12}$$

Now, we consider the initial tv-lasso estimator. By Lemma A.3,

$$|\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 \leq 4\phi_0^{-2}\lambda_1 s. \tag{A.13}$$

Then, (2.18), (2.19) and (2.20) follow by assembling (A.9), (A.10), (A.11) and (A.13) into (2.10)

$$\hat{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}(t) + \mathrm{bias}(\tilde{\boldsymbol{\theta}}(t)) + \{\tilde{\boldsymbol{\theta}}(t) - \mathbb{E}[\tilde{\boldsymbol{\theta}}(t)]\} - \{(P_{\mathcal{R}_t} - I_p)[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]\}.$$

The marginal representation (2.21) and (2.22) follow from similar argument by noting that $B_j(t) = \sum_{k\neq j}(P_{\mathcal{R}_t})_{jk}\beta_k(t)$ under $H_{0,j,t}$. □

A.3 Proof of Theorem 2.3

The proof of Theorem 2.3 is similar to Theorem 2.1 so we only highlight the difference involving the error process. First, $\mathrm{Cov}(\mathcal{E}_t) = W_t^{1/2}\Sigma_{e,t}W_t^{1/2}$. Second, instead of using (A.5) in proving Lemma A.3, we shall use Lemma A.4 to get for all $\lambda > 0$

$$\mathbb{P}\left(\max_{j\le p}\left|\sum_{i\in N_t} w(t,i)X_{ij}e_i\right| \ge \lambda\right) \le 2p\exp\left(-\frac{\lambda^2}{2L_t^2|\mathbf{a}|_1^2\sigma^2}\right) \quad \text{if } \varrho > 1;$$

and

$$\mathbb{P}\left(\max_{j\le p}\left|\sum_{i\in N_t} w(t,i)X_{ij}e_i\right| \ge \lambda\right) \le 2p\exp\left(-\frac{C_\varrho\lambda^2}{L_t^2 n^{2(1-\varrho)}\sigma^2K^2}\right) \quad \text{if } \varrho \in (1/2,1).$$

□

A.4 Proof of Theorem 2.4

The proof essentially follows the lines in Theorem 2.1, however with two key differences of requiring a larger penalty parameter λ_1 of the tv-Lasso. First, by the Nagaev inequality [Nagaev, 1979], we have for any $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{j\le p}\left|\sum_{i\in N_t} w(t,i)X_{ij}e_i\right| \ge \sigma L_t\varepsilon\right) \le (1+2/q)^q\kappa_q\frac{p\mu_{n,q}}{(\sigma L_t\varepsilon)^q} + 2p\exp\left(-c_q\varepsilon^2\right),$$

where $c_q = 2e^{-q}(q+2)^{-2}$ and κ_q is the q-th absolute moment of e_1. Then, choosing

$$\varepsilon = C_q\max\left\{\frac{(p\mu_{n,q})^{1/q}}{\sigma L_t},\ (\log p)^{1/2}\right\} \quad \text{for large enough } C_q > 0,$$

we have $\max_{j\le p}|\sum_{i\in N_t} w(t,i)X_{ij}e_i| = O_{\mathbb{P}}(\lambda_0)$. Second, let $\Xi = (\mathcal{X}_t^\top\mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top W_t^{1/2}$ and $\mathcal{E}_t^\diamond = (e_i)_{i\in N_t}^\top$. Recall that $\tilde{\boldsymbol{\theta}}(t) - \mathbb{E}[\tilde{\boldsymbol{\theta}}(t)] = \Xi\mathcal{E}_t^\diamond$ and denote $\boldsymbol{\xi}_j$ be the columnized j-th row of Ξ. By the Gaussian approximation, e.g. [Shao, 1995, Theorem B], there exist i.i.d. Gaussian random variables $g_i \sim N(0,\sigma^2\xi_{ji}^2)$ defined on a richer probability space such that for every $t > 0$

$$\mathbb{P}\left(\left|\tilde{\theta}_j(t) - \mathbb{E}[\tilde{\theta}_j(t)] - \sum_{i\in N_t} g_i\right| \ge t\right) \le (Cq)^q\frac{\sum_{i\in N_t}\mathbb{E}|\xi_{ji}e_i|^q}{t^q}.$$

Thus, it follows that

$$\tilde{\theta}_j(t) - \mathbb{E}[\tilde{\theta}_j(t)] = N(0,\sigma^2\Omega_{jj}(\lambda_2)) + O_{\mathbb{P}}(|\boldsymbol{\xi}_j|_q).$$

Then, the proof is complete if $|\boldsymbol{\xi}_j|_q = o(\Omega_{jj}^{1/2}(\lambda_2))$. Since $K(\cdot)$ is the uniform kernel such that $W_t = |N_t|^{-1} I_{|N_t|}$ and $\mathcal{X}_t = \mathcal{X}_t^{\diamond}$, this follows from $\Xi = (\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1} \mathcal{X}_t^\top W_t^{1/2} = |N_t|^{-1/2} Q(D + \lambda_2 D^{-1})^{-1} P^\top$ and

$$\rho_{\max}^{1/2}(\Omega(\lambda_2)) \quad \leq \quad \frac{\varepsilon_0 \sigma}{|N_t|^{1/2}(\varepsilon_0^2 + \lambda_2)} \to 0, \qquad \text{as} \quad n \to \infty.$$

□

A.5 Proof of Proposition 2.1

Since we consider the uniform kernel, we may assume $b_n = 1, |N_t| = n$ and then rescale. Observe that

$$\begin{aligned}
\max_{|k| \leq h} |\hat{\sigma}_{e,k}^2 - \sigma_{e,k}^{*2}| \quad &= \quad \max_{|k| \leq h} \frac{1}{n} \left| \sum_{i=1}^{n-k} (\hat{e}_i \hat{e}_{i+k} - e_i e_{i+k}) \right| \\
&\leq \quad \max_{|k| \leq h} \frac{1}{n} \left| \sum_{i=1}^{n-k} \hat{e}_i (\hat{e}_{i+k} - e_{i+k}) \right| + \left| \sum_{i=1}^{n-k} e_{i+k} (\hat{e}_i - e_i) \right| \\
&\leq \quad \max_{|k| \leq h} \frac{1}{n} \left(\sum_{i=1}^{n-k} \hat{e}_i^2 \right)^{1/2} \left(\sum_{i=1}^{n-k} (\hat{e}_{i+k} - e_{i+k})^2 \right)^{1/2} \\
&\qquad + \max_{|k| \leq h} \frac{1}{n} \left(\sum_{i=1}^{n-k} e_{i+k}^2 \right)^{1/2} \left(\sum_{i=1}^{n-k} (\hat{e}_i - e_i)^2 \right)^{1/2} \\
&\leq \quad \left[\left(\frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^2 \right)^{1/2} + \left(\frac{1}{n} \sum_{i=1}^{n} e_i^2 \right)^{1/2} \right] \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{e}_i - e_i)^2 \right)^{1/2}.
\end{aligned}$$

By Lemma A.3,

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{e}_i - e_i)^2 = |\tilde{\mathcal{E}}_t - \mathcal{E}_t|_2^2 = |\mathcal{X}_t [\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 = O_{\mathbb{P}}(\lambda_1^2 s).$$

Then, it follows from the last expression and $n^{-1} \sum_{i=1}^{n} e_i^2 = O_{\mathbb{P}}(1)$ that

$$\max_{|k| \leq h} |\hat{\sigma}_{e,k}^2 - \sigma_{e,k}^{*2}| = O_{\mathbb{P}}(\lambda_1 s^{1/2}).$$

Therefore

$$\rho_{\max}(B_h(\hat{\Sigma}_e) - B_h(\Sigma_e^*)) \lesssim h \max_{|k| \leq h} |\hat{\sigma}_{e,k}^2 - \sigma_{e,k}^{*2}| = O_{\mathbb{P}}(h \lambda_1 s^{1/2}).$$

□

Appendix B

References

[Allen and Tibshirani, 2010] Allen, G. I. and Tibshirani, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4:764–790.

[Banerjee et al., 2008] Banerjee, O., Ghaoui, L. E., and d'Asprémont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

[Bellman, 1961] Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour.* Princeton University Press, Princeton.

[Bickel and Levina, 2008] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36:2577–2604.

[Bickel et al., 2009] Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37:1705–1732.

[Bonilla et al., 2008] Bonilla, E., Chai, K. M., and WIlliams, C. (2008). Multi-task gaussian process prediction. *Advances in Neural Information Processing Systems*, 20:153–160.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford.

[Boyd et al., 2010] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122.

[Bühlmann, 2013] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19:1212–1242.

[Bühlmann and Mandozzi, 2014] Bühlmann, P. and Mandozzi, J. (2014). High-dimensional variable screening and bias in subsequent inference. *Computational Statistics*, 29:407–430.

[Bühlmann and van de Geer, 2011] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics.

[Cai et al., 2011] Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607.

[Cai et al., 2010] Cai, T., Zhang, C., and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38:2118–2144.

[Chen et al., 2013] Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41:2994–3021.

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

[Dempster, 1972] Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28:157–175.

[Dezeure et al., 2015] Dezeure, R., Bü"hlmann, P., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values, and r software hdi. *Statistical Science*, 30:533–558.

[Donoho, 2000] Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *Lecture Notes*.

[Drton and Perlman, 2007] Drton, M. and Perlman, M. D. (2007). Multiple testing and error control in gaussian graphical model selection. *Statistical Science*, 22:430–449.

[Dutilleul, 1999] Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *The Journal of Statistical Computation and Simulation*, 64:105–123.

[Eckart and Young, 1936] Eckart, G. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218.

[Efron, 2009] Efron, B. (2009). Are a set of microarrays independent of each other? *The Annals of Applied Statistics*, 3:922–942.

[Erdös and Rényi, 1960] Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Mathematics Institute of Hungarian Academic Sciences*, 5:17–60.

[Fan et al., 2009] Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3:521–541.

[Fan and Wenyang, 1999] Fan, J. and Wenyang, Z. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, 27:1491–1518.

[Flaxman et al., 2015] Flaxman, S., WIlson, A. G., Neill, D., Nickisch, H., and Smola, A. J. (2015). Fast kronecker inference in gaussian processes with non-gaussian likelihoods. *International Conference on Machine Learning*.

[Friedman et al., 2010a] Friedman, J., Hastie, T., and Tibshirani, R. (2010a). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

[Friedman et al., 2010b] Friedman, J., Hastie, T., and Tibshirani, R. (2010b). Sparse inverse covariance estimation with the graphical lasso. *Journal for Statistical Software*, 33:1–22.

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

[Hoover et al., 1998] Hoover, D., Rice, J., Wu, C., and Yang, L. (1998). Nonparametric smoothing estimates of the time-varying coefficient models with longitudinal data. *Biometrika*, 89:111–128.

[Hsieh et al., 2011] Hsieh, C., Dhillon, I. S., and Ravikumar, P. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. *Advanced in Neural Information Processing Systems*, 24:2330–2338.

[Ibragimov and Khasminskii, 1981] Ibragimov, I. and Khasminskii, Z. (1981). *Statistical estimation: Asymptotic theory*. Springer.

[Javanmard and Montanari, 2014] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

[Lam and Fan, 2009] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *The Annals of Statistics*, 37:4254–4278.

[Lauritzen, 1996] Lauritzen, S. (1996). *Graphical Models*. Oxford University Press, Oxford.

[Lee et al., 2010] Lee, S., Huang, J. Z., and Hu, J. (2010). Sparse logistic principal components analysis for binary data. *The Annals of Applied Statistics*, 4:1579–1601.

[Li et al., 2013] Li, S., Hsu, L., Peng, J., and Wang, P. (2013). Bootstrap inference for network construction. *The Annals of Applied Statistics*, 7.

[Liu et al., 2014] Liu, A., Chen, X., McKeown, M. J., and Wang, Z. J. (2014). A sticky weighted regression model for time-varying resting state brain connectivity estimation. *IEEE Transactions on Biomedical Engineering*.

[McMurry and Politis, 2010] McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, 31:471–482.

[Meinshausen and Bühlman, 2006] Meinshausen, N. and Bühlman, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462.

[Meinshausen et al., 2009] Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681.

[Meinshausen and Yu, 2009] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37:246–270.

[Nagaev, 1979] Nagaev, S. V. (1979). Large deviations of sums of independent random variables. *The Annals of Probability*, 7:745–789.

[Orbe et al., 2005] Orbe, S., Ferreira, E., and Rodriguez-Poo, J. (2005). Nonparametric estimation of time varying parameters under shape restrictions. *Journal of Econometrics*, 126:53–77.

[Rao et al., 2008] Rao, N. R., Mingo, J. A., Speicher, R., and Edelman, A. (2008). Statistical eigen-inference from large wishart matrices. *The Annals of Statistics*, 36:2850–2885.

[Ravikumar et al., 2010] Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2010). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

[Resnick, 1987] Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes.* Applied Probability. Springer-Verlag.

[Robinson, 1989] Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. *Statistical Analysis and Forecasting of Economic Structural Change*, pages 164–253.

[Rothman et al., 2008] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

[Shao and Deng, 2012] Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, 40:812–831.

[Shao, 1995] Shao, Q. (1995). Strong approximation theorems for independent random variables and their applications. *Journal of Multivariate Analysis*, 52:107–130.

[Sun and Zhang, 2012] Sun, T. and Zhang, C. (2012). Scaled sparse linear regression. *Biometrika*, 99:879–898.

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

[Tsiligkaridis, 2014] Tsiligkaridis, T. (2014). *High Dimensional Separable Representations for Statistical Estimation and Controlled Sensing.* US Army Research Laboratory.

[Tsiligkaridis and Hero, 2013] Tsiligkaridis, T. and Hero, A. (2013). Covariance estimation in high dimensions via kronecker product expansions. *IEEE Transactions on Signal Processing*, 61:5347–5360.

[van de Geer and Bühlmann, 2009] van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

[Wang and Zhang, 1992] Wang, B. and Zhang, F. (1992). Some inequalities for the eigenvalues of the product of positive semidefinite hermitian matrices. *Linear Algebra and Its Applications*, 160:113–118.

[Wang et al., 2014] Wang, L., Xue, L., Qu, A., and Liang, H. (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics*, 42:592–624.

[Wei and Huang, 2010] Wei, F. and Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16:1369–1384.

[Werner and Jansson, 2007] Werner, K. and Jansson, M. (2007). Estimation of kronecker structured channel covariances using training data. *In Proceedings of EUSIPCO.*

[Werner et al., 2008] Werner, K., Jansson, M., and Stoica, P. (2008). On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing*, 56:478–491.

[Witten et al., 2011] Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20:892–900.

[Witten and Tibshirani, 2009] Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *Journal of the Royal Statistical Society: Series B*, 71:615–636.

[Yin and Li, 2012] Yin, J. and Li, H. (2012). Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, 107:119–140.

[Yuan and Lin, 2007] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94:19–35.

[Zhang and Zhang, 2014] Zhang, C. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B*, 76:217–242.

[Zhang and Wu, 2012] Zhang, T. and Wu, B. (2012). Inference of time-varying regression models. *The Annals of Statistics*, 40:1376–1402.

[Zhang et al., 2002] Zhang, W., Lee, S. Y., and Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *Journal of Multivariate Analysis*, 82:166–188.

[Zhang and Schneider, 2010] Zhang, Y. and Schneider, J. (2010). Learning multiple tasks with a sparse matrix-normal penalty. *Advances in Neural Information Processing Systems*, 23:2550–2558.

[Zhao and Yu, 2006] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

[Zhou et al., 2010] Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time-varying undirected graphs. *Machine Learning*, 80:295–319.

[Zhou et al., 2009] Zhou, S., van de Geer, S., and B uhlmann, P. (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. *The Annals of Statistics*, 42:532–562.

[Zhou and Wu, 2010] Zhou, Z. and Wu, W. B. (2010). Simultaneous inference of linear models with time varying coefficients. *Journal of the Royal Statistical Society, Series B*, 72:513–531.