Functional Sparse Estimation of Time Varying Graphical Model

Meilei Jiang, Yufeng Liu
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

August 26, 2016

1 Introduction

Graphical models are quite useful in many domains to uncover the dependence structure among observed variables. Typically, we consider a p-dimensional multivariate normal distributed random variable

$$\mathbf{X} = (X_1, \cdots, X_p) \sim \mathbb{N}(\mathbf{0}, \mathbf{\Sigma}),$$

where p is the number of features. Then a useful graph of these p features can be constructed based on there conditional dependence structure. More precisely, we can construct a Gaussian graphical model

$$\begin{aligned} &\mathcal{G} = (V, E), \text{ where } V = \{1, \cdots, p\} \text{ is the set of nodes,} \\ &\text{ and } E = \left\{(j,l) | X_j \text{ is conditionally dependent with } X_l, \text{given } X_{V/\{j,l\}}\right\}. \end{aligned}$$

Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = (\omega_{j,l})_{1 \leq j,l \leq p}$ be the precision matrix. Then X_j and X_l are conditionally dependent given other features if and only if $\omega_{jl} = 0$. Therefore, estimating the covariance matrix and precision matrix of X is equivalent to estimate the structure of Gaussian graphical model $\mathcal{G}$. More discussion can be found in (Lauritzen, 1996).

1.1 Estimation Sparse Precision Matrix $\mathbf{\Omega}$

Given a random sample $\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(n)}$ of $\mathbf{X}$, we aim to estimate $\mathbf{\Omega}$ and recover its support, i.e. the corresponding undirected Gaussian graph. When $n > p$, a nature estimator of $\mathbf{\Omega}$ can be $\hat{\mathbf{\Omega}}_n = \hat{\mathbf{\Sigma}}_n^{-1}$, where $\hat{\mathbf{\Sigma}}_n = \sum_{k=1}^{n} (\mathbf{X}^{(k)} - \bar{\mathbf{X}}_n)(\mathbf{X}^{(k)} - \bar{\mathbf{X}}_n)'$ is the sample covariance matrix. In the case $n < p$, which is quite common in the many applications, the estimation of $\mathbf{\Omega}$ is much more challenging since $\hat{\mathbf{\Sigma}}_n$ is no longer invertible.

There are lots of literatures discussing about the estimation of sparse precision matrix $\boldsymbol{\Omega}$ in high dimension low sample size settings, i.e. $p > n$. Generally speaking, there are three main approaches.

1. **Covariance selection approach.** There is a connection between linear regression and prediction matrix $\boldsymbol{\Omega}$:

$$\mathbf{X}_j = \mathbf{X}_{-j}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j = \sum_{l \neq j} \mathbf{X}_l \beta_{jl} + \boldsymbol{\varepsilon}_j \tag{1}$$

   It could be shown that $\beta_{jl} = \omega_{jl}/\omega_{jj}$. Thus estimating $\boldsymbol{\beta}_j$ can identify the support of ith row of $\boldsymbol{\Omega}$. Meinshausen and Bühlmann (2006) applied LASSO penalty (Tibshirani, 1996) on multivariate regression 1 to estimate the support of $\boldsymbol{\Omega}$ row by row. Peng et al. (2012) considered a joint sparse regression to estimate the support of $\boldsymbol{\Omega}$ together through an active-shooting algorithm. Yuan (2010) applied Danzig selector (Candes and Tao, 2007) to the problem 1 to estimate each column of $\boldsymbol{\Omega}$.

2. **Penalized likelihood approach.** Another nature way is to estimate $\boldsymbol{\Omega}$ is the penalized likelihood approach. The log-likelihood of the parameters in $\boldsymbol{\Omega}$ is as following:

$$l(\mathbf{X}^(i), 1 \leq i \leq n|\boldsymbol{\Omega}) = -\text{tr}(\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}_n) + \log|\boldsymbol{\Omega}| \tag{2}$$

   In order to have a sparse estimation of $\boldsymbol{\Omega}$, different penalized likelihood estimator are considered. Yuan and Lin (2007) proposed the *max-det problem* to solve the LASSO-type estimator.

$$\hat{\boldsymbol{\Omega}}_L = \arg\min \text{tr}(\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}_n) - \log|\boldsymbol{\Omega}| + \lambda\|\boldsymbol{\Omega}\|_1 \tag{3}$$

   and the non-negative garrote-type estimator.

$$\begin{aligned} \hat{\boldsymbol{\Omega}}_G &= \arg\min \text{tr}(\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}_n) - \log|\boldsymbol{\Omega}| + \lambda \sum_{j \neq i} \frac{\omega_{jl}}{\tilde{\omega}_{jl}} \\ \text{subject to } & \frac{\omega_{jl}}{\tilde{\omega_{jl}}} \geq 0, \boldsymbol{\Omega} \text{ p.d.} \end{aligned} \tag{4}$$

   Banerjee et al. (2008) used a block coordinate descent algorithm to solve (3). And then Friedman et al. (2008) proposed the *graphical lasso* algorithm for (3) based on least square lasso type estimator, which is simple and fast. Rothman et al. (2008) proposed the sparse permutation invariant covariance estimator (SPICE) which could extend to the ℓ_q-type penalized likelihood estimator. Fan et al. (2009), Lam et al. (2009) studied the penalized likelihood estimator with the smoothly clipped absolute deviation (SCAD) penalty and the adaptive LASSO penalty.

3. ℓ_1 **constrained minimization approach.** Cai et al. (2011) performed a constrained ℓ_1 minimization approach to estimate sparse precision matrix (CLIME).

$$\begin{aligned} \hat{\mathbf{\Omega}}_C &= \arg\min \|\mathbf{\Omega}\|_1 \\ \text{subject to } &\|\mathbf{\Omega}\hat{\mathbf{\Sigma}} - \mathbf{I}\|_\infty \leq \lambda_n \end{aligned} \tag{5}$$

Cai et al. (2016) proposed adaptive constrained ℓ_1 minimization estimator (ACLIME), which achieved the optimal minimax rate of convergence.

1.2 Heterogeneous Data And Time Varying Graphical Model

The methods aforementioned focus on estimating a single Gaussian graph by assuming samples are identically distributed. However, in many applications it is more realistic to assume that data are heterogeneous due to batch effects or latent factors. Guo et al. (2011) reparameterized the off-diagonal entry $\omega_{jl} = \theta_{jl}\gamma_{jl}^k$ and estimated them through a penalized likelihood with the hierarchical penalty on common structures θ_{jl} and individual structureγ_{jl}^k. Danaher et al. (2014) propose the *joint graphical lasso*, to estimate multiple graphical models corresponding to distinct but related conditions. The *joint graphical lasso* utilized fussed lasso and group lasso on log-likelihood to force similarity among graphs. Lee and Liu (2015) proposed a method to estimate the common structure and unique structure through the constrained ℓ_1 minimization.

In many cases the sample indexes have orders, e.g. time, and the corresponding graphs evolve through the order. In such cases it could be quite interesting to estimate a time varying graphical model.

$$\mathbf{X}(t) \sim \mathbb{N}(\mathbf{0}, \mathbf{\Sigma}(t)) \tag{6}$$

Zhou et al. (2010) developed a nonparametric framework for estimating time varying graphical model by kernel smoothing and ℓ_1 panelty. Zhou's model assumed that the observations X^t are independent and changed smoothly.

$$\begin{gathered} \hat{\mathbf{\Omega}}(\tau) = \arg\min_{\mathbf{\Omega}} \left\{ \operatorname{tr}(\mathbf{\Omega}\hat{\mathbf{\Sigma}}(\tau)) - \log|\mathbf{\Omega}| + \lambda\|\mathbf{\Sigma}^-\|_1 \right\} \\ \text{where } \hat{\mathbf{\Sigma}}(\tau) = \sum_i \omega_i^\tau \mathbf{X}^i(\mathbf{X}^i)', \text{and } \omega_i^\tau = \frac{K_h(t_i - \tau)}{\sum_{i'} K_h(t_{i'} - \tau)} \end{gathered} \tag{7}$$

Lu et al. (2015) proposed a dynamic nonparanormal graphical model, which is more robust, by estimating a weighted Kendall's tau correlation matrix. These approaches generated estimation of similar graphs by weighting the samples among different times.

1.3 Varying Coefficient Model And Sparse Derivatives

Considering the samples collected from different time points as longitudinal data and estimating the time varying network from the viewpoint of regression approach, we are looking at the following model

$$X_j(t) = \mathbf{X}'_{-j}(t)\boldsymbol{\beta}_j(t) + \varepsilon_j(t) = \sum_{l \neq j} X_l(t)\beta_{jl}(t) + \varepsilon_j(t). \tag{8}$$

Model (8) is in the form of the varying coefficient model (Cleveland et al., 1992; Hastie and Tibshirani, 1993). There are lots of literature studying varying coefficient model

$$Y(t) = X(t)^T \boldsymbol{\beta}(t) + \varepsilon(t) \tag{9}$$

A good overview of the varying coefficient model study was given by Fan and Zhang (2008). For the longitude data, there are two major approaches to estimate Model (9). One approach is local kernel polynomial smoothing (Fan and Zhang, 1999; Wu and Chiang, 2000). Another approach is basis expansion (Huang et al., 2002, 2004).

Assume that we collect sample at $t_1, \cdots, t_n$, denote $X^i_j = X_j(t_i)$. Following the kernel polynomial smoothing approach, Kolar et al. (2009, 2010); Kolar and Xing (2011, 2012) proposed a local linear regression approach with "kernel ℓ_1" penalty to estimate the smoothly varying graph,

$$\hat{\boldsymbol{\beta}}_j(\tau) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p-1}}{\arg\min} \sum_i (X^i_j - \sum_{l \neq j} X^i_l \beta_l)^2 \omega^\tau_i + \lambda |\boldsymbol{\beta}|_1 \tag{10}$$

and total variation penalty to estimate graph with jumps.

$$\begin{aligned} \left\{\hat{\boldsymbol{\beta}}_j(t_1), \cdots, \hat{\boldsymbol{\beta}}_j(t_n)\right\} = \underset{\boldsymbol{\beta}(t_i), i \leq n}{\arg\min} & \sum_i (X^i_j - \sum_{l \neq j} X^i_l \beta_l(t_i))^2 \\ & + \lambda_1 \sum_i |\boldsymbol{\beta}(t_i)|_1 + \lambda_2 \sum_{i=2}^{n} |\boldsymbol{\beta}(t_i) - \boldsymbol{\beta}(t_{i-1})|_1 \end{aligned} \tag{11}$$

Current literatures of time varying networks mostly follows this approach. There are good reasons for this approach since the node X_j is indeed locally linearly dependent with other nodes $\mathbf{X}_{-j}$. However, the estimation $\hat{\boldsymbol{\beta}}_j(t)$ are usually not smooth over time. While current works focus on correctly variable selection locally, our approach focus on both variable selection and estimating the function of $\boldsymbol{\beta}(t)$. Our method of estimating Model (8) follows basis expansion approach. To best of our knowledge, there is no work following this approach to estimate time varying networks.

Moreover, in order to estimate a sparse graph, we need to gain sparse functional coefficient $\boldsymbol{\beta}(t)$. Based on the idea of FLiRTI in James, Wang and Zhu's paper James et al. (2009), we put penalty on the derivative matrices of $\boldsymbol{\beta}(t)$. Kim et al. (2009) also put the l_1 penalty on derivative of coefficient function in the trend filtering problem. Tibshirani et al. (2014) applied the generalized lasso (Tibshirani and Taylor, 2011) to solve the optimization problem in the trendfilter. We will apply ADMM algorithm to solve our optimization problem.

2 Methodology

2.1 Functional Undirected Graph

Consider p smooth functional continuous variables $\{X_1(t), X_2(t) \cdots, X_p(t)\}$ on the 'time' domain $\mathcal{T}$. On each t, we assume

$$(X_1(t), X_2(t) \cdots, X_p(t)) \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma}(t)).$$

Define the undirected graph

$$\begin{aligned} G(t) &= \{V, E(t)\}\,, \text{ where } V = \{1, \cdots, p\}\,, \\ \text{and } E(t) &= \left\{(j,l) \in V^2 : \operatorname{Cov}\left[X_j(t), X_l(t) | X_k(t), k \neq j, l\right] \neq 0, j \neq l\right\}. \end{aligned} \tag{12}$$

Namely, $G(t)$ is the Gaussian graphical model at each t. This model is quite flexible, which allows $G(t)$ evolve over time and includes the time dependence. Assume that data are observed at $t_1, \cdots, t_n$ and at each time point t we have n_t samples.

2.2 Functional Nearest Neighborhood Selection

Assuming data are collected at $t_1, \cdots, t_n$, and there are n_t replicates at each time point t. Consider the following functional coefficient model

$$\begin{gathered} X_j^r(t) = \mathbf{X}_{-j}^r(t)^T \boldsymbol{\beta}_j(t) + \varepsilon_j^r(t) = \sum_{l \neq j} X_l^r(t) \beta_{jl}(t) + \varepsilon_j^r(t), \\ \text{where } \mathbf{X}_{-j}^r(t) = (X_l^r(t))_{l \neq j} \in \mathbb{R}^{(p-1) \times 1}, r = 1, \cdots, n_t, t = t_1, \cdots, t_n, j = 1, \cdots, p. \end{gathered} \tag{13}$$

Moreover, for each t, let $\mathbf{X}_j(t) = (X_j^1(t), \cdots, X_j^{n_t}(t))^T \in \mathbb{R}^{n_t \times 1}, \boldsymbol{\varepsilon}(t) = (\boldsymbol{\varepsilon}_j^1(t), \cdots, \boldsymbol{\varepsilon}_j^{n_t}(t))^T \in \mathbb{R}^{n_t \times 1}$, then Model (13) can be represented as

$$\begin{aligned} \mathbf{X}_j(t) &= \mathbf{X}_{-j}(t)' \boldsymbol{\beta}_j(t) + \boldsymbol{\varepsilon}_j(t) \\ &= \sum_{l \neq j} \mathbf{X}_l(t) \beta_{jl}(t) + \boldsymbol{\varepsilon}_j(t), \end{aligned} \tag{14}$$

For each functional coefficient $\beta_{jl}(t)$, we consider the basis expansion $\mathbf{B}_{jl}(t) = (B_{jl1}(t), \cdots, B_{jlk_{jl}}(t))$:

$$\beta_{jl}(t) = \sum_{s=1}^{k_{jl}} B_{jls}(t) \gamma_{jls} + e_{jl}(t) = \mathbf{B}_{jl}(t) \boldsymbol{\gamma}_{jl} + e_{jl}(t)$$

Then we get the functional coefficient vector $\boldsymbol{\beta}_j(t) = \mathbf{B}(t)\boldsymbol{\gamma}_j + \mathbf{e}_j(t) = (\beta_{jl}|j \neq l) \in \mathbb{R}^{p-1}$, where $\mathbf{B}(t) = \text{diag}\{\mathbf{B}_{jl}(t)\} \in \mathbb{R}^{(p-1)\times\sum_{l\neq j} k_{jl}}$, $\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{jl})_{l\neq j} \in \mathbb{R}^{\sum_{l\neq j} k_{jl}\times 1}$, $\mathbf{e}(t) = (\mathbf{e}_{jl}|j \neq l) \in \mathbb{R}^{p-1}$.

Thus Equation (14) can be represented as

$$
\begin{aligned}
\mathbf{X}_j(t) &= \sum_{l\neq j}\sum_{s=1}^{k_{jl}} \mathbf{X}_l(t) B_{jls}(t)\gamma_{jls} + \tilde{\boldsymbol{\varepsilon}}(t) \\
&= \mathbf{X}_{-j}(t)\mathbf{B}_j(t)\boldsymbol{\gamma}_j + \tilde{\boldsymbol{\varepsilon}}(t) \\
&= \mathbf{U}(t)\boldsymbol{\gamma}_j + \tilde{\boldsymbol{\varepsilon}}(t) \\
\text{where } \mathbf{U}(t) &= \mathbf{X}_{-j}(t)\mathbf{B}_j(t) \in \mathbb{R}^{\sum_{t=1}^{n} n_t \times \sum_{j\neq i} k_{jl}}, \\
\tilde{\boldsymbol{\varepsilon}}(t) &= \mathbf{X}_{-j}(t)\mathbf{e}_j(t) + \boldsymbol{\varepsilon}(t) \\
&= \sum_{l\neq j} \mathbf{X}_l(t) e_{jl}(t) + \boldsymbol{\varepsilon}(t) \in \mathbb{R}^{n_t}, \\
t &= t_1, \cdots, t_n, j = 1, \cdots, p.
\end{aligned}
\tag{15}
$$

As seen in Equation (15), our model is quite flexible since the basis of each functional coefficient can be various for different nodes.

Combing the data from $t_1, \cdots, t_n$, we can get the matrix form of Model 13.

$$
\begin{aligned}
\mathbf{X}_j &= \mathbf{U}_j\boldsymbol{\gamma}_j + \tilde{\boldsymbol{\varepsilon}}_j, j = 1, \cdots, p. \\
\text{where } \mathbf{X}_j &= (X_j(t_1)', \cdots, X_j(t_n)')' \in \mathbb{R}^{\sum_{t=1}^{n} n_t \times 1} \\
\mathbf{U}_j &= (\mathbf{U}_j(t_1)', \cdots, \mathbf{U}_j(t_n)')' \in \mathbb{R}^{\sum_{t=1}^{n} n_t \times \sum_{j\neq i} k_{jl}} \\
\tilde{\boldsymbol{\varepsilon}}_j &= (\tilde{\boldsymbol{\varepsilon}}_j(t_1), \cdots, \tilde{\boldsymbol{\varepsilon}}_j(t_n))^T \in \mathbb{R}^{\sum_{t=1}^{n} n_t \times 1}
\end{aligned}
\tag{16}
$$

For each node j, least square of coefficient vector $\boldsymbol{\gamma}_j$ in the model 16 is

$$
\begin{aligned}
l(\boldsymbol{\gamma}_j) &= (\mathbf{X}_j - \mathbf{U}_j\boldsymbol{\gamma}_j)'\mathbf{W}(\mathbf{X}_j - \mathbf{U}_j\boldsymbol{\gamma}_j) \\
&= \sum_{i=1}^{n} (\mathbf{X}_j(t_i) - \mathbf{U}_j(t_i)\boldsymbol{\gamma}_j)^2 w_i \\
&= \sum_{i=1}^{n} (\mathbf{X}_j(t_i) - \mathbf{X}_{-j}(t_i)\boldsymbol{\beta}_j(t_i))^2 w_i
\end{aligned}
$$

2.3 Control The Sparsity Of Derivatives

In Model (16), we want to estimate a sparse graph as well as control the smoothness of coefficient functions. Typically, for each j and $l \neq j$, we want to control the sparsity and smoothness of $\beta_{jl}(t) \approx \sum_{s=1}^{k_{jl}} B_{jls}(t)\gamma_{jls}$. Our approach is to choose m-degree basis function and put l_1 penalty on $\beta_{jl}(t)$ and total variation penalty on $\beta_{jl}^{(m)} = \frac{\mathrm{d}^{\mathrm{m}}}{\mathrm{d}t^m}\beta_{jl}(t) \approx \frac{\mathrm{d}^{\mathrm{m}}}{\mathrm{d}t^m}\mathbf{B}_{jl}(t)^T\boldsymbol{\gamma}_{jl}$.

Let

$$\mathbf{A}_{jl} = \left(\frac{\mathrm{d}^{\mathrm{m}}}{\mathrm{d}t^m} \mathbf{B}_{jl}(t_1), \cdots, \frac{\mathrm{d}^{\mathrm{m}}}{\mathrm{d}t^m} \mathbf{B}_{jl}(t_n) \right)^T \in \mathbb{R}^{n \times k_{jl}}, \tag{17}$$

Next, set

$$\boldsymbol{\eta}_{jl} = \mathbf{A}_{jl} \boldsymbol{\gamma}_{jl} \in \mathbb{R}^{n \times 1} \tag{18}$$

Then $\boldsymbol{\eta}_{jl} \approx (\beta_{jl}^{(m)}(t_k))_{1 \le k \le n}$. Moreover, we denote

$$\begin{gathered} \boldsymbol{\eta}_j = (\boldsymbol{\eta}_{jl})_{l \neq j} = \mathbf{A}_j \boldsymbol{\gamma}_j \in \mathbb{R}^{n(p-1)}, \\ \text{where } \mathbf{A}_j = \operatorname{diag}(\mathbf{A}_{jl})_{l \neq j} \in \mathbb{R}^{n(p-1) \times \sum_{l \neq j} k_{jl}}. \end{gathered} \tag{19}$$

We want to put the sparsity penalty on the $\boldsymbol{\eta}_j$. Then Model (16) can be expressed as the following ℓ_1 optimization problem, which is a generalized lasso problem.

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{j,L} &= \arg\min_{\boldsymbol{\gamma}_j} \frac{1}{2} \|\mathbf{X}_j - \mathbf{U}_j \boldsymbol{\gamma}_j\|_2^2 \\ \text{subject to } \|\boldsymbol{\eta}_j\|_1 &= \|\mathbf{A}_j \boldsymbol{\gamma}_j\|_1 \le t \\ \text{i.e. } \hat{\boldsymbol{\gamma}}_{j,L} &= \arg\min_{\boldsymbol{\gamma}_j} \frac{1}{2} \|\mathbf{X}_j - \mathbf{U}_j \boldsymbol{\gamma}_j\|_2^2 + \lambda \|\mathbf{A}_j \boldsymbol{\gamma}_j\|_1 \end{aligned} \tag{20}$$

Moreover, if we want to control the sparsity of multiple derivatives of $\boldsymbol{\beta}_{jl}(t)$, say we want both $\boldsymbol{\beta}_{jl}^{(0)}(t) = 0$ and $\boldsymbol{\beta}_{jl}^{(2)}(t) = 0$ in large area.

B-splines Basis A typical choice of basis function is the B-splines basis, which is defined recursively on nodes $t_1, \cdots, t_n$.

- $B_i^1(x) = \mathbb{1}_{[t_i, t_{i+1})}(x), i = 1, \cdots, n-1$.
- $B_i^k(x) = \frac{x - t_i}{t_{i+k-1} - t_i} B_i^{k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1}^{k-1}(x)$

Then the derivatives of B-splines have the form:

$$\frac{\mathrm{d}}{\mathrm{d}x} B_i^k(x) = \frac{k-1}{t_{i+k-1} - t_i} B_i^{k-1}(x) - \frac{k-1}{t_{i+k} - t_{i+1}} B_{i+1}^{k-1}(x)$$

When the knots are evenly spaced with gap Δt, the derivatives can be written as

$$\frac{\mathrm{d}}{\mathrm{d}x} B_i^k(x) = \frac{B_i^{k-1}(x) - B_{i+1}^{k-1}(x)}{\Delta t}.$$

3 Optimization

In order to solve Model (16), an Alternative Direction Methods of Multiplier (ADMM) approach has been developed.

For each j, Model (16) can be rewritten as

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{X}_j - \mathbf{U}_j\gamma_j\|_2^2 + \lambda\|\boldsymbol{\eta}_j\|_1 \\ &\text{Subject to } \boldsymbol{\eta}_j = \mathbf{A}_j\boldsymbol{\gamma}_j \end{aligned} \tag{21}$$

The augmented Lagrange of Model (21) is as following:

$$L_\rho(\boldsymbol{\gamma}_j, \boldsymbol{\eta}_j, \mathbf{y}) = \frac{1}{2}\|\mathbf{X}_j - \mathbf{U}_j\boldsymbol{\gamma}_j\|_2^2 + \lambda\|\boldsymbol{\eta}_j\|_1 + \rho\mathbf{y}'(\boldsymbol{\eta}_j - \mathbf{A}_j\boldsymbol{\gamma}_j) + \frac{\rho}{2}\|\boldsymbol{\eta}_j - \mathbf{A}_j\boldsymbol{\gamma}_j\|_2^2 \tag{22}$$

Then the updating rule for $\boldsymbol{\gamma}_j^{(k)}$ is

$$\begin{aligned} \boldsymbol{\gamma}_j^{(k)} &= \arg\min_{\boldsymbol{\gamma}} L_\rho(\boldsymbol{\gamma}, \boldsymbol{\eta}_j^{(k-1)}, \mathbf{y}^{(k-1)}) \\ &= \arg\min_{\boldsymbol{\gamma}} \frac{1}{2}\|\mathbf{X}_j - \mathbf{U}_j\boldsymbol{\gamma}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{\eta}_j^{(k)} - \mathbf{A}_j\boldsymbol{\gamma} + \mathbf{y}^{(k-1)}\|_2^2 \\ &= (\mathbf{U}_j'\mathbf{U}_j + \rho\mathbf{A}_j'\mathbf{A}_j)^+ \left[\mathbf{U}_j'\mathbf{X}_j + \rho\mathbf{A}_j'(\boldsymbol{\eta}^{(k-1)} + \mathbf{y}^{(k-1)})\right] \end{aligned} \tag{23}$$

The updating rule for $\boldsymbol{\eta}_j^{(k)}$ is

$$\begin{aligned} \boldsymbol{\eta}_j^{(k)} &= \arg\min_{\boldsymbol{\eta}} L_\rho(\boldsymbol{\gamma}_j^{(k)}, \boldsymbol{\eta}, \mathbf{y}^{(k-1)}) \\ &= \arg\min_{\boldsymbol{\eta}} \lambda\|\boldsymbol{\eta}\|_1 + \frac{\rho}{2}\|\boldsymbol{\eta} - \mathbf{A}_j\boldsymbol{\gamma}_j^{(k)} + \mathbf{y}^{(k-1)}\|_2^2 \\ &= \text{prox}_{\lambda/\rho|\cdot|_1}(\mathbf{A}_j\boldsymbol{\gamma}_j^{(k)} - \mathbf{y}^{(k-1)}) \\ &= \mathcal{S}_{\lambda/\rho}(\mathbf{A}_j\boldsymbol{\gamma}_j^{(k)} - \mathbf{y}^{(k-1)}) \\ \text{where } &\mathcal{S}_u(\mathbf{x}) = (\mathbf{x} - u)_+ - (-\mathbf{x} - u)_+ \end{aligned} \tag{24}$$

The updating rule for $\mathbf{y}^{(k)}$ is

$$\mathbf{y}^{(k)} = \mathbf{y}^{(k-1)} + \rho(\boldsymbol{\eta}^{(k)} - \mathbf{A}_j\boldsymbol{\gamma}_j^{(k)}) \tag{25}$$

4 Tuning Parameter

Akaike information criterion (AIC) and Bayes information criterion (BIC) are common model selection criteria to achieve optimal prediction error. However these methods usually have good theoretical properties in low dimensions and are not suitable for high dimensional problems. Stability selection is a method for tunning parameter of estimation of structure.

References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008), "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *Journal of Machine Learning Research*, 9, 485–516.

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, T. T., Liu, W., and Zhou, H. H. (2016), "Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation," *The Annals of Statistics*, 44, 455–488.

Candes, E. and Tao, T. (2007), "The Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, 2313–2351.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992), "Local regression models," *Statistical models in S*, 2, 309–376.

Danaher, P., Wang, P., and Witten, D. M. (2014), "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 373–397.

Fan, J., Feng, Y., and Wu, Y. (2009), "Network exploration via the adaptive LASSO and SCAD penalties," *The annals of applied statistics*, 3, 521.

Fan, J. and Zhang, W. (1999), "Statistical estimation in varying coefficient models," *Annals of Statistics*, 1491–1518.

— (2008), "Statistical methods with varying coefficient models," *Statistics and its Interface*, 1, 179.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011), "Joint estimation of multiple graphical models," *Biometrika*, asq060.

Hastie, T. and Tibshirani, R. (1993), "Varying-coefficient models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002), "Varying-coefficient models and basis function approximations for the analysis of repeated measurements," *Biometrika*, 89, 111–128.

— (2004), "Polynomial spline estimation and inference for varying coefficient models with longitudinal data," *Statistica Sinica*, 763–788.

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

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009), "ℓ_1 Trend Filtering," *SIAM review*, 51, 339–360.

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010), "Estimating time-varying networks," *The Annals of Applied Statistics*, 94–123.

Kolar, M., Song, L., and Xing, E. P. (2009), "Sparsistent learning of varying-coefficient models with structural changes," in *Advances in Neural Information Processing Systems*, pp. 1006–1014.

Kolar, M. and Xing, E. P. (2011), "On time varying undirected graphs," *Journal of Machine Learning Research*, 15, 407–415.

— (2012), "Estimating networks with jumps," *Electronic journal of statistics*, 6, 2069.

Lam, C., Fan, J., et al. (2009), "Sparsistency and rates of convergence in large covariance matrix estimation," *The Annals of Statistics*, 37, 4254–4278.

Lauritzen, S. L. (1996), *Graphical models*, Clarendon Press.

Lee, W. and Liu, Y. (2015), "Joint Estimation of Multiple Precision Matrices with Common Structures," *Journal of Machine Learning Research*, 16, 1035–1062.

Lu, J., Kolar, M., and Liu, H. (2015), "Post-regularization Inference for Dynamic Nonparanormal Graphical Models," *arXiv preprint arXiv:1512.08298.*

Meinshausen, N. and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, 1436–1462.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2012), "Partial correlation estimation by joint sparse regression models," *Journal of the American Statistical Association.*

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

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

Tibshirani, R. J. and Taylor, J. (2011), "THE SOLUTION PATH OF THE GENERALIZED LASSO," *The Annals of Statistics*, 1335–1371.

Tibshirani, R. J. et al. (2014), "Adaptive piecewise polynomial estimation via trend filtering," *The Annals of Statistics*, 42, 285–323.

Wu, C. O. and Chiang, C.-T. (2000), "Kernel smoothing on varying coefficient models with longitudinal dependent variable," *Statistica Sinica*, 433–456.

Yuan, M. (2010), "High dimensional inverse covariance matrix estimation via linear programming," *The Journal of Machine Learning Research*, 11, 2261–2286.

Yuan, M. and Lin, Y. (2007), "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94, 19–35.

Zhou, S., Lafferty, J., and Wasserman, L. (2010), "Time varying undirected graphs," *Machine Learning*, 80, 295–319.