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

March 22, 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 $\mathbf{X} = (X_1, \cdots, X_p) \sim \mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 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 $\mathcal{G} = (V, E)$, where $V = \{1, \cdots, p\}$ is the set of nodes and $E = \{(i,j)|X_i \text{ is conditionally dependent with } X_b, \text{given } X_{V/\{i,j\}}\}$.

Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\omega_{i,j})_{1 \le i,j \le p}$ be the precision matrix. Then X_i and X_j are conditionally dependent given other features if and only if $\omega_{ij} = 0$. 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 [11].

There are lots of literatures discussing about estimating $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$. Utilizing the idea of LASSO from Tibshirani [15], Meinshausen and Bühlmann [13] used a regression approach, which is a computationally attractive method, to perform the nearest neighborhood selection of non-zero elements of $\boldsymbol{\Omega}$ at each node in the graph. Another nature way is to estimate $\boldsymbol{\Sigma}$ and Ω is the penalized likelihood approach. Friedman, Hastie nd Tibshirani [6] proposed the graphical lasso. Fan, Feng and Wu [3] studied the penalized likelihood methods with the SCAD penalty and the adaptive LASSO penalty. Cai, Liu and Luo [1] performed a constrained l_1 minimization approach to estimate sparse precision matrix (CLIME).

In particular, we are interested in estimating multiple graphs from different locations or time points. Peng et al. [14] proposed the regression approach to estimate multiple graphs through an active-shooting algorithm. Danaher et al. [2] 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 to force similarity among graphs. These approaches generated estimation of similar graphs by forcing the similarity of parameters.

Zhou, Laffery and Wasserman [16] developed a nonparametric framework for estimating time varying graphical model for estimating time varying graphical structure for multivariate Gaussian distributions $\mathbf{X}^t \sim \mathbb{N}(0, \Sigma(t))$ using l_1 regularization method. Zhou's model assumed that the observations X^t are independent and changed smoothly. Kolar and Xing [10] showed the model selection consistency for the procedure proposed in Zhou et al. [16] and for the modified neighborhood selection procedure of Meinshausen and Bühlmann [13]. Lu, Kolar and Liu [12] 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.

All these methods mentioned above can only estimate multiple graphs on several fix points. If we are interested in estimating a dynamic graph which is smoothing over a time region, a functional model which consider each feature as functional data over time is very attractive. Since the structure of Gaussian graph can be recovered though regression approach, we can consider dynamic graphical model under the context of varying coefficient model [7, 4, 5]. Typically, we are interested in a nonparametric approach to estimate the varying coefficient models in Huang, Wu and Zhou's paper [8].

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

Then we are looking at the following model

$$X_i(t) = \mathbf{X}_{-i}^T(t)\boldsymbol{\beta}_i(t) + \varepsilon_i(t) = \sum_{j \neq i} X_j(t)\beta_{ij}(t) + \varepsilon_i(t).$$

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 [9], we put penalty on the derivative matrices of $\boldsymbol{\beta}(t)$.

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

$$\begin{aligned} G(t) &= \Big\{V, E(t)\Big\}, \text{ where } V = \Big\{1, \cdots, p\Big\}, \\ \text{and } E(t) &= \Big\{(i,j) \in V^2 : \mathrm{Cov}\Big[X_i(t), X_j(t) | X_k(t), k \neq i, j\Big] \neq 0, i \neq j\Big\}. \end{aligned} \tag{1}$$

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

Consider the following functional linear model

$$X_i^r(t) = \mathbf{X}_{-i}^r(t)^T \boldsymbol{\beta}_i(t) + \varepsilon_i^r(t) = \sum_{j \neq i} X_j^r(t)\beta_{ij}(t) + \varepsilon_i^r(t),$$
$$\text{where } \mathbf{X}_{-i}^r(t) = (X_j^r(t))_{j \neq i} \in \mathbb{R}^{(p-1)\times 1}, r = 1, \cdots, n_t, t = t_1, \cdots, t_n, i = 1, \cdots, p. \tag{2}$$

For each functional coefficient $\beta_{ij}(t)$, we consider the basis $\mathbf{B}_{ij}(t) = (B_{ij1}(t), \cdots, B_{ijk_{ij}}(t))$ and then

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

Thus Equation (2) can be represented as

$$X_i^r(t) = \sum_{j \neq i} \sum_{s=1}^{k_{ij}} X_j^r(t) B_{ijs}(t)\gamma_{ijs} + \tilde{\varepsilon}^r(t),$$
$$\text{where } \tilde{\varepsilon}^r(t) = \sum_{j \neq i} X_j^r(t) e_{ij}^r(t) + \varepsilon^r(t), r = 1, \cdots, n_t, t = t_1, \cdots, t_n, i = 1, \cdots, p. \tag{3}$$

As seen in Equation (3), our model is quite flexible since the basis of each functional coefficient can be different.

To get the matrix form of Equation (3), denote

$$\begin{aligned}
&\mathbf{B}(t) = \text{diag}\{\mathbf{B}_{ij}(t)\} \in \mathbb{R}^{(p-1)\times \sum_{j\neq i} k_{ij}},\\
&\mathbf{U}_i^r(t) = \mathbf{B}(t)^T X_{-i}^r(t) \in \mathbb{R}^{\sum_{j\neq i} k_{ij} \times 1},\\
&\mathbf{U}_i(t) = (\mathbf{U}_i^1(t), \cdots, \mathbf{U}_i^{n_t}(t))^T \in \mathbb{R}^{n_t \times \sum_{j\neq i} k_{ij}},\\
&\mathbf{U}_i = (\mathbf{U}_i(t_1), \cdots, \mathbf{U}_i(t_n))^T \in \mathbb{R}^{\sum_{t=1}^n n_t \times \sum_{j\neq i} k_{ij}},\\
&\mathbf{X}_i(t) = (X_i^1(t), \cdots, X_i^{n_t}(t))^T \in \mathbb{R}^{n_t \times 1},\\
&\mathbf{X}_i = (X_i(t_1), \cdots, X_i(t_n))^T \in \mathbb{R}^{\sum_{t=1}^n n_t \times 1},\\
&\boldsymbol{\varepsilon}_i(t) = (\varepsilon_i^1(t), \cdots, \varepsilon_i^{n_t}(t))^T \in \mathbb{R}^{n_t \times 1},\\
&\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_1), \cdots, \varepsilon_i(t_n))^T \in \mathbb{R}^{\sum_{t=1}^n n_t \times 1},\\
&\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{ij})_{j\neq i} \in \mathbb{R}^{\sum_{j\neq i} k_{ij} \times 1}.
\end{aligned}$$

Then the Equation (3) can be expressed as

$$\mathbf{X}_i = \mathbf{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i, i = 1, \cdots, p. \tag{4}$$

2.3 Control The Sparsity Of Derivatives

In Model (4), we want to estimate a sparse graph and interpretable coefficient functions. For each i and $j \neq i$, we want to control the sparsity of $\beta_{ij}^{(m)} = \frac{\mathrm{d}^\mathrm{m}}{\mathrm{d}t^m}\beta_{ij}(t) \approx \frac{\mathrm{d}^\mathrm{m}}{\mathrm{d}t^m}\mathbf{B}_{ij}(t)^T\boldsymbol{\gamma}_{ij}$ for some m. Say assume $\beta_{ij}^{(0)}(t) = 0$ and $\beta_{ij}^{(2)}(t) = 0$ in large area, then $\beta_{ij}(t)$ is zero in many region and linear in the left regions.

Let

$$\mathbf{A}_{ij} = \begin{bmatrix} D^m\mathbf{B}_{ij}(t_1) & \cdots & D^m\mathbf{B}_{ij}(t_n) \end{bmatrix}^T \in \mathbb{R}^{n\times k_{ij}}, \tag{5}$$

where D^m is the mth finite difference operator, i.e., $D\mathbf{B}_{ij}(t_k) = [\mathbf{B}_{ij}(t_k) - \mathbf{B}_{ij}(t_{k-1})]/[t_k - t_{k-1}]$, $D^2\mathbf{B}_{ij}(t_k) = [D\mathbf{B}_{ij}(t_k) - D\mathbf{B}_{ij}(t_{k-1})]/[t_k - t_{k-1}]$, etc.

Next, set

$$\boldsymbol{\eta}_{ij} = \mathbf{A}_{ij}\boldsymbol{\gamma}_{ij} \in \mathbb{R}^{n\times 1} \tag{6}$$

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

$$\begin{aligned} \boldsymbol{\eta}_i = (\boldsymbol{\eta}_{ij})_{j\neq i} = \mathbf{A}_i\boldsymbol{\gamma}_i \in \mathbb{R}^{n(p-1)}, \\ \text{where } \mathbf{A}_i = \mathrm{diag}(\mathbf{A}_{ij})_{j\neq i} \in \mathbb{R}^{n(p-1)\times\sum_{j\neq i}k_{ij}}. \end{aligned} \tag{7}$$

We want to put the sparsity penalty on the $\boldsymbol{\eta}_i$. Then Model 4 can be expressed as the following optimization problem, which is a generalized lasso problem.

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{i,L} &= \arg\min_{\boldsymbol{\gamma}_i} \frac{1}{2}\|\mathbf{X}_i - \mathbf{U}_i\boldsymbol{\gamma}_i\|_2^2 \\ \text{subject to } \|\boldsymbol{\eta}_i\|_1 &= \|\mathbf{A}_i\boldsymbol{\gamma}_i\|_1 \leq t \\ \text{i.e. } \hat{\boldsymbol{\gamma}}_{i,L} &= \arg\min_{\boldsymbol{\gamma}_i} \frac{1}{2}\|\mathbf{X}_i - \mathbf{U}_i\boldsymbol{\gamma}_i\|_2^2 + \lambda\|\mathbf{A}_i\boldsymbol{\gamma}_i\|_1 \end{aligned} \tag{8}$$

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

There is a connection between Model 8 and fused lasso.

References

[1] Tony Cai, Weidong Liu, and Xi Luo. A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[2] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[3] Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive lasso and scad penalties. *The annals of applied statistics*, 3(2):521, 2009.

[4] Jianqing Fan and Wenyang Zhang. Statistical estimation in varying coefficient models. *Annals of Statistics*, pages 1491–1518, 1999.

[5] Jianqing Fan and Wenyang Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179, 2008.

[6] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[7] Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796, 1993.

[8] Jianhua Z Huang, Colin O Wu, and Lan Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128, 2002.

[9] Gareth M James, Jing Wang, and Ji Zhu. Functional linear regression that's interpretable. *The Annals of Statistics*, pages 2083–2108, 2009.

[10] Mladen Kolar and Eric P Xing. On time varying undirected graphs. 2011.

[11] Steffen L Lauritzen. *Graphical models*. Clarendon Press, 1996.

[12] Junwei Lu, Mladen Kolar, and Han Liu. Post-regularization inference for dynamic nonparanormal graphical models. *arXiv preprint arXiv:1512.08298*, 2015.

[13] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[14] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 2012.

[15] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[16] Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319, 2010.