TECHNOMETRICS
2018, VOL. 60, NO. 4, 511–521
https://doi.org/10.1080/00401706.2017.1408497

Taylor & Francis
Taylor & Francis Group

# Nonparametric Finite Mixture of Gaussian Graphical Models

Kevin H. Lee[a,b] and Lingzhou Xue[a,b]

[a]Western Michigan University, Kalamazoo, MI; [b]Pennsylvania State University, State College, PA

**ABSTRACT**
Graphical models have been widely used to investigate the complex dependence structure of high-dimensional data, and it is common to assume that observed data follow a homogeneous graphical model. However, observations usually come from different resources and have heterogeneous hidden commonality in real-world applications. Thus, it is of great importance to estimate heterogeneous dependencies and discover a subpopulation with certain commonality across the whole population. In this work, we introduce a novel regularized estimation scheme for learning nonparametric finite mixture of Gaussian graphical models, which extends the methodology and applicability of Gaussian graphical models and mixture models. We propose a unified penalized likelihood approach to effectively estimate nonparametric functional parameters and heterogeneous graphical parameters. We further design an efficient generalized effective expectation-maximization (EM) algorithm to address three significant challenges: high-dimensionality, nonconvexity, and label switching. Theoretically, we study both the algorithmic convergence of our proposed algorithm and the asymptotic properties of our proposed estimators. Numerically, we demonstrate the performance of our method in simulation studies and a real application to estimate human brain functional connectivity from attention deficit hyperactivity disorder (ADHD) imaging data, where two heterogeneous conditional dependencies are explained through profiling demographic variables and supported by existing scientific findings.

**ARTICLE HISTORY**
Received May 2016
Revised October 2017

**KEYWORDS**
Attention deficit hyperactivity disorder (ADHD); Brain imaging; Covariate-dependent graphical model; Finite mixture model; Label switching; Nonconvex optimization; Nonparametric estimation

## 1. Introduction

Graphical model has been widely used to investigate the complex dependence structure of high-dimensional data, and it has successful applications in various research fields. For example, in bioinformatics, the graphical model is used in exploring the patterns of association in gene expression data (Dobra et al. 2004; Schäfer and Strimmer 2005), binary genomic data (Wang, Chao, and Hsu 2011; Xue, Zou, and Cai 2012), and cell signaling data (Voorman, Shojaie, and Witten 2014) among others. Due to advances in functional magnetic resonance imaging (fMRI), investigating brain function connectivity becomes increasingly important ( Ryali et al. 2012). Gaussian graphical model has been extensively used in estimating the functional connectivity in brain imaging ( Ng et al. 2013; Varoquaux et al. 2010). The central question here is to infer conditional dependencies or independencies from high-dimensional fMRI data. In the current literature, it is common to assume that high-dimensional data come from a homogeneous resource and follow a parametric or semiparametric graphical model, for instance, Gaussian graphical model and its variants (Meinshausen and Bühlmann 2006; Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008; Peng et al. 2009; Witten, Friedman, and Simon 2011; Cai, Liu, and Luo 2011; Liu et al. 2012; Xue and Zou 2012; Chandrasekaran, Parrilo, and Willsky 2012; Ma, Xue, and Zou 2013; Danaher, Wang, and Witten 2014).

However, it is very common in real-world applications that observed data come from different resources and may have heterogeneous dependencies across the whole population. For instance, genetic variations data and gene expression data of the international HapMap project (International HapMap 3 Consortium et al. 2010) consist of four representative populations in the world. Our research is motivated by exploring the heterogeneous dependencies of human brain fMRI data to study the attention deficit hyperactivity disorder (ADHD). The famous ADHD-200 Global Competition data (Biswal et al. 2010) aggregated across eight independent imaging sites. Thus, it is very important to estimate heterogeneous dependencies and discover subpopulation with certain commonality. In the current literature, there are two different arguments about whether the gender affects the ADHD. On the one hand, some studies show that there is a gender difference in ADHD (Sauver et al. 2004). On the other hand, other studies argue that the ADHD is not systematically different between boys and girls (Bauermeister et al. 2007). It is likely that there may exist two different subpopulations with hidden commonality among the whole population, corresponding to two existing arguments, respectively. Section 6 confirms this conjecture. More specifically, we show that both arguments may be explained through investigating heterogeneous brain functional connectivity using our proposed method.

In this work, we introduce a novel regularized estimation scheme for learning nonparametric finite mixture of Gaussian graphical models, which explores the heterogeneous dependencies of high-dimensional data. Denoted by $\mathcal{G}^z(\mathbf{x}) = (\mathcal{V}, \mathcal{E}^z)$, the graphical model of random vector $\mathbf{x} \in \mathbb{R}^p$ with a

**CONTACT** Lingzhou Xue lzxue@psu.edu Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/tech.
Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/TECH.

univariate covariate $Z = z$, vertex set $\mathcal{V} = \{1, \dots, p\}$ and edge set $\mathcal{E}^z$, where edge set $\mathcal{E}^z$ may depend on $z$. Let $K$ be number of mixtures. Let $\mathcal{G}_k^z(\mathbf{x}) = (\mathcal{V}, \mathcal{E}_k^z)$ represents the Gaussian graphical model in the $k$th subpopulation. Define $\mathcal{C}$ as the latent class variable satisfying that $P(\mathcal{C} = k|Z = z) = \pi_k(z)$, where $\pi_k(z)$ is a nonparametric mixing proportion function. Throughout this article, we consider the following nonparametric finite mixture of Gaussian graphical models given some covariate $Z = z$:

$$\mathcal{G}^z(\mathbf{x}) = \pi_1(z)\mathcal{G}_1^z(\mathbf{x}) + \pi_2(z)\mathcal{G}_2^z(\mathbf{x}) + \cdots + \pi_K(z)\mathcal{G}_K^z(\mathbf{x}), \quad (1)$$

where $\pi_k(\cdot)$'s are nonparametric mixing proportion functions, and $\pi_1(z) + \cdots + \pi_K(z) = 1$ for any $z \in \mathbb{R}$. Mixture models are powerful to effectively identify subpopulations with hidden commonality within the whole population (Lindsay 1995; McLachlan and Peel 2004). The nonparametric Bayesian estimation of mixtures of Gaussian graphical models have been developed, for instance, Rodríguez et al. (2011). Under the fixed-dimensional setting, Mallapragada, Jin, and Jain (2010) and Huang, Li, and Wang (2013) studied the nonparametric finite mixture model for clustering and regression analysis, respectively. Mixture models were extensively studied in the classical low-dimensional scenarios, but received less attention in high-dimensional statistical learning. Recently, Städler, Bühlmann, and Van De Geer (2010) studied $\ell_1$ penalization for mixtures of high-dimensional regression models, and Ruan, Yuan, and Zou (2011) studied mixtures of Gaussian graphical models. However, it is a significant challenge to learn nonparametric finite mixture of Gaussian graphical models (1) in the presence of high-dimensionality, nonconvexity, and label switching. In Section 2, we propose a unified penalized likelihood approach to effectively estimate both nonparametric functional parameters and heterogeneous graphical parameters. To estimate nonparametric functional parameters, we adopt the idea of kernel regression technique and employ a local likelihood approach. Section 3 designs an efficient generalized effective expectation-maximization (EM) algorithm to address three aforementioned challenges simultaneously. Furthermore, we propose an effective criterion to choose the number of mixtures, tuning parameter, and bandwidth. We study the algorithmic convergence of our EM algorithm and the asymptotic result of our estimates in Section 4. Simulation studies and the real application to time-varying brain functional connectivity estimation are demonstrated in Sections 5 and 6, respectively. In Section 6, we discover two heterogeneous dependencies in the ADHD brain functional connectivity, which are explained through profiling demographic variables and supported by existing scientific findings. Our results provide helpful insights to study two different ADHD subpopulations with hidden commonality.

The contributions of our work can be summarized as follows. Methodologically, the proposed regularized estimation of nonparametric finite mixture of graphical models greatly extends the methodology and applicability of time-varying graphical models (Ahmed and Xing 2009; Kolar et al. 2010; Zhou, Lafferty, and Wasserman 2010) and finite mixture of graphical models (Ruan, Yuan, and Zou 2011) for the analysis of high-dimensional data. Computationally, we design a novel efficient generalized effective EM algorithm to effectively estimate the nonparametric finite mixture and functional parameters, which addresses high-dimensionality, nonconvexity, and label switching simultaneously. Theoretically, we provide a new paradigm to not only prove the local convergence of the proposed generalized effective EM algorithm but also study the asymptotic properties of the local solution to the nonparametric finite mixture for high-dimensional data.

## 2. Methodology

Given some univariate covariate $Z = z$, nonparametric finite mixture (1) entails that $\mathbf{X} = \mathbf{x}$ follows a nonparametric finite mixture of multivariate Gaussian distributions:

$$\mathbf{X} = \mathbf{x} \mid Z = z \sim_d \sum_{k=1}^{K} \pi_k(z) N_p(\boldsymbol{\mu}_k(z), \boldsymbol{\Sigma}_k(z)), \quad (2)$$

where $\mathcal{G}_k^z(\mathbf{x})$ corresponds to a multivariate normal distribution $N_p(\boldsymbol{\mu}_k(z), \boldsymbol{\Sigma}_k(z))$ for $k = 1, \dots, K$. Compared to the parametric finite mixture, the nonparametric mixing probabilities $\pi_k(z)$'s achieve the appealing robustness to model misspecification. We discuss the selection of the number of mixtures $K$ in Section 3.2. We may also follow Rodríguez et al. (2011) to consider a nonparametric Bayesian approach to study the Dirichlet process mixtures of Gaussian graphical models when the underlying clusters are unknown. For space consideration, we do not pursue this alternative approach in this article.

Let $\boldsymbol{\Theta}_k(z) = (\theta_{kij}(z))_{p\times p}$ be the precision matrix in the $k$th mixture. Then, $\theta_{kij}(z)$'s specify the graphical model $\mathcal{G}_k^z(\mathbf{x})$ in the $k$th mixture (Dempster 1972). Specifically, given $Z = z$, zeroes in $\theta_{kij}(z)$'s are equivalent to conditional independencies of $\mathbf{X}$ in the $k$th mixture. Thus, zeroes in $\theta_{kij}(z)$'s can be translated to a meaningful graphical model

$$\theta_{kij}(z) \neq 0 \iff X_i \not\perp\!\!\!\perp X_j \mid \mathbf{X} \setminus \{X_i, X_j\}, \mathcal{C} = k,$$
$$Z = z \iff \{i, j\} \in \mathcal{E}_k^z.$$

For ease of presentation, we start with the known a priori number of mixtures $K$, and we present an effective information criterion to select $K$ in Section 3.2.

Given the independent data $\{(\mathbf{x}_n, z_n), n = 1, \dots, N\}$, our goal is to estimate nonparametric functions $\pi_k(\cdot)$'s and functional parameters $\boldsymbol{\mu}_k(\cdot)$'s and $\boldsymbol{\Theta}_k(\cdot)$'s. The global average log-likelihood function for the observed data is given by

$$\ell_N = \frac{1}{N}\sum_{n=1}^{N} \log\left[\sum_{k=1}^{K} \pi_k(z_n)\phi(\mathbf{x}_n|\boldsymbol{\mu}_k(z_n), \boldsymbol{\Theta}_k(z_n))\right],$$

where $\phi(\cdot|\boldsymbol{\mu}, \boldsymbol{\Theta})$ is the density of a multivariate Gaussian distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Theta}^{-1})$.

Next, we employ the local likelihood approach (Brillinger 1977; Tibshirani and Hastie 1987) to estimate $\pi_k(z)$'s, $\boldsymbol{\mu}_k(z)$'s, and $\boldsymbol{\Theta}_k(z)$'s for any $z \in \mathbb{R}$, given smoothness assumptions of such functional parameters. The local likelihood was first mentioned by Brillinger (1977) as a method of smoothing, and it was formally introduced and studied in details by Tibshirani and Hastie (1987). The local likelihood method received considerable attention in the nonparametric regression modeling and nonparametric function estimation (Loader 1996, 2006). Now, we follow Tibshirani and Hastie (1987) to define the localized version of the average log-likelihood function as

$$\ell_z = \frac{1}{N}\sum_{n=1}^{N}\log\left[\sum_{k=1}^{K}\pi_k^z\phi\left(\mathbf{x}_n|\boldsymbol{\mu}_k^z,\boldsymbol{\Theta}_k^z\right)\right]K_h(z_n-z),$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ is a symmetric kernel function with bandwidth $h$. Given the local log-likelihood $\ell_z$, we then maximize the $\ell_1$-penalized local log-likelihood to estimate local constants $\pi_k^z$'s, vectors $\boldsymbol{\mu}_k^z$'s, and matrices $\boldsymbol{\Theta}_k^z$'s as follows:

$$\max_{\{(\pi_k^z,\boldsymbol{\mu}_k^z,\boldsymbol{\Theta}_k^z)\}_{k=1,\dots,K}}\mathcal{L}_z := \max_{\{(\pi_k^z,\boldsymbol{\mu}_k^z,\boldsymbol{\Theta}_k^z)\}_{k=1,\dots,K}}\ell_z - \lambda\sum_{k=1}^{K}\|\boldsymbol{\Theta}_k^z\|_1, \quad (3)$$

where $\|\cdot\|_1$ is the entrywise matrix $\ell_1$ norm. Instead of $\ell_1$ penalization, we may also consider the folded concave penalized estimation to encourage sparsity in $\boldsymbol{\Theta}_k^z$'s (Fan, Feng, and Wu 2009; Fan, Xue, and Zou 2014). For space consideration, we only focus on the $\ell_1$ penalization.

*Remark 2.1.* When $K = 1$, nonparametric finite mixture of graphical models (1) reduces to covariate-dependent graphical model. When the covariate $Z$ represents the varying time, $\mathcal{G}^z(\mathbf{x})$ becomes time-varying graphical model (Ahmed and Xing 2009; Kolar et al. 2010; Zhou, Lafferty, and Wasserman 2010), which is solved by using the penalized likelihood approach

$$\hat{\boldsymbol{\Theta}}^z = \arg\max_{\boldsymbol{\Theta}^z}\frac{1}{N}\sum_{n=1}^{N}\log\left[\phi(\mathbf{x}_n|\boldsymbol{\mu}^z,\boldsymbol{\Theta}^z)\right] \times K_h(z_n-z) - \lambda\|\boldsymbol{\Theta}^z\|_1.$$

*Remark 2.2.* When $\mathcal{E}_k$ does not depend on covariate $z$, nonparametric finite mixture (1) reduces to semiparametric finite mixture of graphical models. Let $\mathcal{G}_k(\mathbf{x}) = (\mathcal{V},\mathcal{E}_k)$ be the $k$th mixture. The semiparametric finite mixture becomes $\mathcal{G}^z(\mathbf{x}) = \pi_1(z)\mathcal{G}_1(\mathbf{x}) + \cdots + \pi_K(z)\mathcal{G}_K(\mathbf{x})$, where $\pi_1(z) + \cdots + \pi_K(z) = 1$ for any $z \in \mathbb{R}$. Conditioning on $Z = z$, $\mathbf{x}$ follows a semiparametric finite mixture $\sum_{k=1}^{K}\pi_k(z)N_p(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$. Now we introduce local constants $\pi_k^z$'s to approximate $\pi_k(z)$'s. Next, we may solve local estimates and global estimates of $\boldsymbol{\Theta}_k$'s. First, we solve local constants $\tilde{\pi}_k^z$'s from

$$\max_{\{(\pi_k^z,\boldsymbol{\Theta}_k)\}_{k=1,\dots,K}}\frac{1}{N}\sum_{n=1}^{N}\log\left[\sum_{k=1}^{K}\pi_k^z\phi(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Theta}_k)\right] \times K_h(z_n-z) - \lambda\sum_{k=1}^{K}\|\boldsymbol{\Theta}_k\|_1.$$

After obtaining local estimates $\tilde{\pi}_k(z_n)$'s for $n = 1,\dots,N$, we solve global estimates via

$$\max_{\{\boldsymbol{\Theta}_k\}_{k=1,\dots,K}}\frac{1}{N}\sum_{n=1}^{N}\log\left[\sum_{k=1}^{K}\tilde{\pi}_k(z_n)\phi(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Theta}_k)\right] - \lambda\sum_{k=1}^{K}\|\boldsymbol{\Theta}_k\|_1.$$

*Remark 2.3.* When there is no covariate, nonparametric finite mixture (1) further reduces to finite mixture of Gaussian graphical models (Ruan, Yuan, and Zou 2011), that is,

$$\mathcal{G}(\mathbf{x}) = \pi_1\mathcal{G}_1(\mathbf{x}) + \pi_2\mathcal{G}_2(\mathbf{x}) + \cdots + \pi_K\mathcal{G}_K(\mathbf{x}),$$

that is, $\mathbf{x} \sim_d \sum_{k=1}^{K}\pi_k N_p(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$, where $\pi_1 + \cdots + \pi_K = 1$. This can be solved by

$$\max_{\{(\pi_k,\boldsymbol{\Theta}_k)\}_{k=1,\dots,K}}\frac{1}{N}\sum_{n=1}^{N}\log\left[\sum_{k=1}^{K}\pi_k\phi(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Theta}_k)\right] - \lambda\sum_{k=1}^{K}\|\boldsymbol{\Theta}_k\|_1.$$

## 3. Computation

EM algorithm provides a powerful tool to solve latent variable problems in mixture model. Wu (1983) established some general convergence properties, and Balakrishnan, Wainwright, and Yu (2017) recently studied the statistical guarantees on both the population level and the sample level. However, we need to address two significant challenges when solving the nonparametric finite mixture model (3): (1) nonconvex optimization in high dimensions; (2) label switching issue at different grid points of covariate $z$. This section presents a generalized effective EM algorithm to address both the challenges, which enjoys some appealing convergence properties as shown in Section 4.

### 3.1 Proposed EM Algorithm

Following the spirit of EM algorithm, we view the collected data $(\mathbf{x}_n, z_n)$, $n = 1,\dots,N$ to be incomplete, and then define random variables $\boldsymbol{\tau}_n = (\tau_{1n},\dots,\tau_{Kn})'$ with

$$\tau_{kn} = \begin{cases} 1 & \text{if } (\mathbf{x}_n, z_n) \text{ is in the } k\text{th mixture,} \\ 0 & \text{otherwise} \end{cases}$$

to identify the mixture of $(\mathbf{x}_n, z_n)$. Given the complete data $\{(\mathbf{x}_n, z_n, \boldsymbol{\tau}_n), n = 1,\dots,N\}$, the complete log-likelihood function is written as

$$\ell_N^{\text{cmp}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}\left[\log\pi_k(z_n) + \log\phi(\mathbf{x}_n|\boldsymbol{\mu}_k(z_n),\boldsymbol{\Theta}_k(z_n))\right].$$

Let $z \in \{u_1,\dots,u_G\}$, the set of grid points. We employ kernel regression techniques to estimate $\pi_k(z_n)$, $\boldsymbol{\mu}_k(z_n)$, and $\boldsymbol{\Theta}_k(z_n)$. Define a local complete log-likelihood as

$$\ell_z^{\text{cmp}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{kn}\left[\log\pi_k^z + \log\phi(\mathbf{x}_n|\boldsymbol{\mu}_k^z,\boldsymbol{\Theta}_k^z)\right]K_h(z_n-z).$$

Next, we define the $\ell_1$-penalized local log-likelihood function for the complete data as

$$\mathcal{L}_z^{\text{cmp}} = \ell_z^{\text{cmp}} - \lambda\sum_{k=1}^{K}\|\boldsymbol{\Theta}_k^z\|_1.$$

Notice that there are potential label switching issues at any two different grid points $z, z' \in \mathcal{U} = \{u_1,\dots,u_G\}$. To solve this issue, we propose the following generalized effective EM algorithm. Given the current estimates of parameters, E-step estimates all conditional expectations (4) at observed $\{z_1,\dots,z_N\}$. M-step uses the obtained common conditional expectations to update all estimates of parameters at each grid point in $\mathcal{U}$. Hence, we effectively prevent the label switching issue at different grid points. As shown in Figure 1, two solution paths from simulation

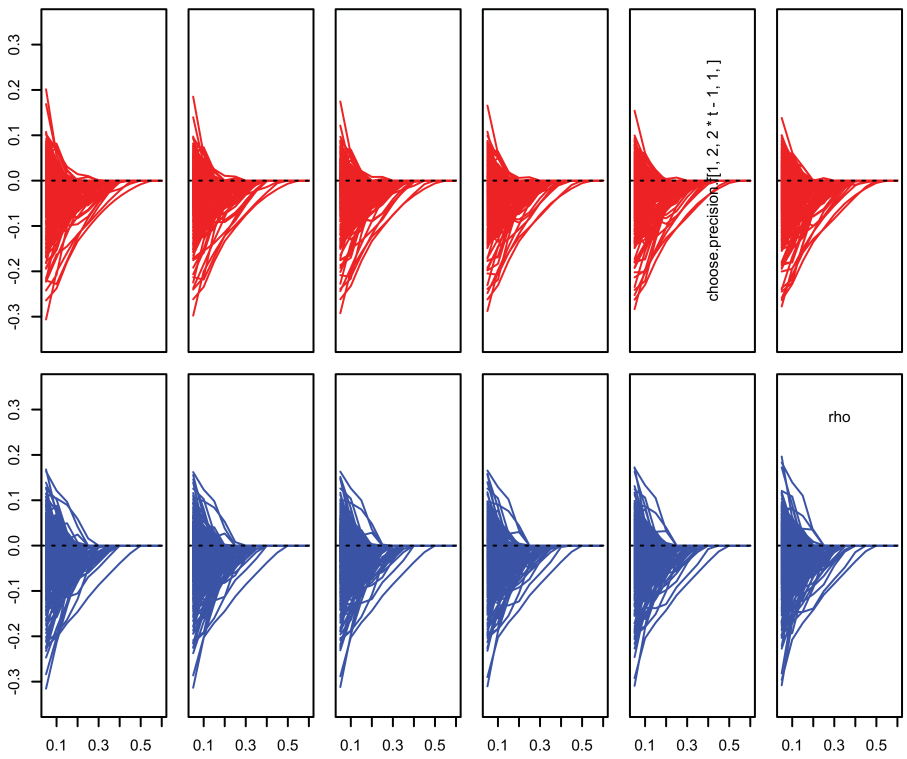

**Figure 1.** Solution path at six different grid points: 1st mixture (upper panel) and 2nd mixture (lower panel).

studies in Section 5 demonstrate that two mixtures are consistently identified at different grid points.

In what follows, we present algorithm details. Since only $(\mathbf{x}_n, z_n)$'s are observed, we treat $\boldsymbol{\tau}_n$'s as missing data. In the $(t+1)$th iteration, $t = 0, 1, 2, \ldots$, E-step employs the $t$th iterated solution $\pi_k^{(t)}(z_n)$, $\boldsymbol{\mu}_k^{(t)}(z_n)$, and $\boldsymbol{\Theta}_k^{(t)}(z_n)$ to compute the conditional expectation of $\tau_{kn}$ given the current estimates. By using Bayes' rule, we have

$$\gamma_{kn}^{(t+1)} = \frac{\pi_k^{(t)}(z_n)\phi\left(\mathbf{x}_n | \boldsymbol{\mu}_k^{(t)}(z_n), \boldsymbol{\Theta}_k^{(t)}(z_n)\right)}{\sum_{l=1}^K \pi_l^{(t)}(z_n)\phi\left(\mathbf{x}_n | \boldsymbol{\mu}_l^{(t)}(z_n), \boldsymbol{\Theta}_l^{(t)}(z_n)\right)}. \tag{4}$$

Next, M-step obtains the estimates of parameters from maximizing

$$\frac{1}{N}\sum_{n=1}^N \sum_{k=1}^K \gamma_{kn}^{(t+1)} \left[\log \pi_k^z + \log \phi\left(\mathbf{x}_n | \boldsymbol{\mu}_k^z, \boldsymbol{\Theta}_k^z\right)\right] \times K_h(z_n - z) - \lambda \sum_{k=1}^K \|\boldsymbol{\Theta}_k^z\|_1,$$

subject to the constraint that $\sum_{k=1}^K \pi_k^z = 1$. It is equivalent to maximizing

$$\frac{1}{N}\sum_{n=1}^N \sum_{k=1}^K \gamma_{kn}^{(t+1)} \log \pi_k^z K_h(z_n - z), \tag{5}$$

subject to $\sum_{k=1}^K \pi_k^z = 1$, and for $k = 1, \ldots, K$, maximizing

$$\frac{1}{N}\sum_{n=1}^N \gamma_{kn}^{(t+1)} \log \phi(\mathbf{x}_n | \boldsymbol{\mu}_k^z, \boldsymbol{\Theta}_k^z) K_h(z_n - z) - \lambda \|\boldsymbol{\Theta}_k^z\|_1, \tag{6}$$

respectively. To solve the subproblem (5), we introduce Lagrange multiplier $\alpha$ with the constraint $\sum_{k=1}^K \pi_k^z = 1$. Then in the $(t+1)$th cycle, for $z \in \{u_g, g = 1, \ldots, G\}$, we update $\pi_k^z$ by

$$\pi_k^{z(t+1)} = \sum_{n=1}^N \frac{\gamma_{kn}^{(t+1)} K_h(z_n - z)}{\sum_{n'=1}^N K_h(z_{n'} - z)}. \tag{7}$$

In order to solve the subproblem (6), we first simplify (6) as

$$\frac{1}{N}\sum_{n=1}^N \gamma_{kn}^{(t+1)} \left[\frac{1}{2}\log|\boldsymbol{\Theta}_k^z| - \frac{1}{2}\left(\mathbf{x}_n - \boldsymbol{\mu}_k^z\right)' \boldsymbol{\Theta}_k^z \left(\mathbf{x}_n - \boldsymbol{\mu}_k^z\right)\right] \times K_h(z_n - z) - \lambda \|\boldsymbol{\Theta}_k^z\|_1.$$

Then, it is easy to obtain the closed-form update for $\boldsymbol{\mu}_k^z$, that is

$$\boldsymbol{\mu}_k^{z(t+1)} = \sum_{n=1}^N \frac{\gamma_{kn}^{(t+1)} K_h(z_n - z)\mathbf{x}_n}{\sum_{n'=1}^N \gamma_{kn'}^{(t+1)} K_h(z_{n'} - z)}. \tag{8}$$

Next, we employ the state-of-art optimization algorithm such as Friedman, Hastie, and Tibshirani (2008), Witten, Friedman, and Simon (2011), and Goldfarb, Ma, and Scheinberg (2013) and to solve $\boldsymbol{\Theta}_k^z$ from

$$\boldsymbol{\Theta}_k^{z(t+1)} = \underset{\boldsymbol{\Theta}_k^z}{\arg\max} \left\{\log|\boldsymbol{\Theta}_k^z| - \mathrm{tr}\left(\boldsymbol{\Theta}_k^z \mathbf{A}_k^{z(t+1)}\right) - \lambda \|\boldsymbol{\Theta}_k^z\|_1\right\}, \tag{9}$$

where $\mathbf{A}_k^{z(t+1)} = \sum_{n=1}^N \frac{\gamma_{kn}^{(t+1)} K_h(z_n - z)}{\sum_{n'=1}^N \gamma_{kn'}^{(t+1)} K_h(z_{n'} - z)} (\mathbf{x}_n - \boldsymbol{\mu}_k^{z(t+1)})(\mathbf{x}_n - \boldsymbol{\mu}_k^{z(t+1)})'$. Furthermore, we update $\pi_k^{(t+1)}(z_n)$, $\boldsymbol{\mu}_k^{(t+1)}(z_n)$, and $\boldsymbol{\Theta}_k^{(t+1)}(z_n)$, $n = 1, \ldots, N$ by linear interpolating $\pi_k^{u_g(t+1)}$, $\boldsymbol{\mu}_k^{u_g(t+1)}$, and $\boldsymbol{\Theta}_k^{u_g(t+1)}$, $g = 1, \ldots, G$, respectively. More details about (7), (8), and (9) are presented in the online Appendix.

Now, we summarize the details of our proposed algorithm in Algorithm 1.

**Algorithm 1** Proposed generalized effective EM algorithm.

```
• Initialization of $\pi_k^{(0)}(z_n)$, $\boldsymbol{\mu}_k^{(0)}(z_n)$, and $\boldsymbol{\Theta}_k^{(0)}(z_n)$ for all $k$.
• Iteratively, solve E-step and M-step with $t = 0, 1, \ldots,$ till convergence:
    – E-step: solve $\gamma_{kn}^{(t+1)}$ from (4) for all $k$ and $n$.
    – M-step: solve $\pi_k^{z(t+1)}$, $\boldsymbol{\mu}_k^{z(t+1)}$, $\boldsymbol{\Theta}_k^{z(t+1)}$ from (7)–(9) for all $k$ and $z$.
```

### 3.2. Selection of Tuning Parameters

We need to select three tuning parameters: number of mixtures $K$, penalization parameter $\lambda$, and bandwidth $h$. To determine them, we consider the information criterion approach. Bayesian information criterion (BIC) has the general form of $-2\mathcal{L} + \delta \times df$, where $\mathcal{L}$ is the maximum log-likelihood, $\delta = \log N$, and $df$ is the degree of freedom to measure model complexity. To specify the degree of freedom in nonparametric finite mixture (2), we follow Fan, Zhang, and Zhang (2001) and Huang, Li, and Wang (2013) to derive the degree of freedom. Denoted by $df = \tau_K h^{-1}|\mathcal{Z}|(K(0) - \frac{1}{2}\int K^2(t)\,dt)$ the degree of freedom of a univariate nonparametric function, where $\mathcal{Z}$ is the support of the covariate $Z$, and $\tau_K = \frac{K(0)-\frac{1}{2}\int K^2(t)\,dt}{\int (K(t)-\frac{1}{2}K*K(t))^2\,dt}$. Hence, for each pair of $(K, \lambda, h)$, the BIC score is defined as

$$\mathrm{BIC}(K, \lambda, h) = -2\mathcal{L} + \log(N) \times df(K, \lambda, h),$$

where

$$df(K, \lambda, h) = \left[(K-1) + \frac{1}{G}\sum_{z\in\mathcal{U}}\left(Kp + \sum_{k=1}^{K}\sum_{i\le j} I_{\{\hat{\Theta}_{ijk}^{z}\neq 0\}}\right)\right] \times df.$$

We first select $K$ and $\lambda$ by minimizing the BIC score, and then choose $h$ by cross-validation (CV). We choose $K$ by minimizing the best available BIC score for each choice of $K$ over different choices of $\lambda$ and $h$. Namely,

$$\hat{K} = \arg\min_{(\lambda,h)} \mathrm{BIC}(K, \lambda, h).$$

After fixing $K = \hat{K}$, we choose $\lambda$ by minimizing the best available BIC score for each $\lambda$ over different choices of $h$. Namely,

$$\hat{\lambda} = \arg\min_{h} \mathrm{BIC}(\hat{K}, \lambda, h).$$

Last, we use the log-likelihood to construct the CV loss, and choose $h$ by CV.

## 4. Theoretical Properties

This section will first establish the algorithmic convergence of our proposed algorithm, and then prove the asymptotic properties of our proposed estimator.

### 4.1. Algorithmic Convergence

Recall that $\mathcal{L}_z(\boldsymbol{\pi}^z, \boldsymbol{\mu}^z, \boldsymbol{\Theta}^z) = \ell_z(\boldsymbol{\pi}^z, \boldsymbol{\mu}^z, \boldsymbol{\Theta}^z) - \lambda\sum_{k=1}^{K}\|\boldsymbol{\Theta}_k^z\|_1$ is the objective function (3). Let $\{(\boldsymbol{\pi}^{z(t)}, \boldsymbol{\mu}^{z(t)}, \boldsymbol{\Theta}^{z(t)}) : t = 0, 1, 2, \ldots,\}$ be the sequence generated by Algorithm 1. Since the objective function is nonconcave and nondifferentiable, we study the local convergence of our proposed algorithm. To this end, we first define the *cluster point* in the sequence $\{(\boldsymbol{\pi}^{z(t)}, \boldsymbol{\mu}^{z(t)}, \boldsymbol{\Theta}^{z(t)}) : t = 0, 1, 2, \ldots,\}$ generated by Algorithm 1, which is also known as "accumulation point" in mathematics or optimization theory.

*Definition 4.1.* For any given grid point $z$, $(\bar{\boldsymbol{\pi}}^z, \bar{\boldsymbol{\mu}}^z, \bar{\boldsymbol{\Theta}}^z)$ is called a *cluster point* in the sequence $\{(\boldsymbol{\pi}^{z(t)}, \boldsymbol{\mu}^{z(t)}, \boldsymbol{\Theta}^{z(t)}) : t = 0, 1, 2, \ldots,\}$ if there exists a convergent subsequence $\{(\boldsymbol{\pi}^{z(t_j)}, \boldsymbol{\mu}^{z(t_j)}, \boldsymbol{\Theta}^{z(t_j)}) : j = 1, 2, \ldots,\}$ such that $(\boldsymbol{\pi}^{z(t_j)}, \boldsymbol{\mu}^{z(t_j)}, \boldsymbol{\Theta}^{z(t_j)}) \to (\bar{\boldsymbol{\pi}}^z, \bar{\boldsymbol{\mu}}^z, \bar{\boldsymbol{\Theta}}^z)$.

Next, we follow Tseng (2001) and Städler, Bühlmann, and Van De Geer (2010) to define the stationary point under the nonconcave and nondifferentiable maximization.

*Definition 4.2.* Let $u$ be a function defined on some open set $U$. A point $s \in U$ is called a *stationary point* if the directional derivative $u'(s; d) = \lim_{\alpha\downarrow 0}\frac{u(s+\alpha d)-u(s)}{\alpha} \le 0, \quad \forall d$.

We first show that Algorithm 1 preserves the nice ascent property as the classical EM algorithm with probability tending to 1.

*Theorem 4.1.* Suppose $h \to 0$ and $Nh \to \infty$ as $N \to \infty$. For any given point $z$ and $t = 0, 1, 2, \ldots,$ with probability tending to 1, we always have

$$\mathcal{L}_z(\boldsymbol{\pi}^{z(t+1)}, \boldsymbol{\mu}^{z(t+1)}, \boldsymbol{\Theta}^{z(t+1)}) \ge \mathcal{L}_z(\boldsymbol{\pi}^{z(t)}, \boldsymbol{\mu}^{z(t)}, \boldsymbol{\Theta}^{z(t)}),$$

and $\{\mathcal{L}_z(\boldsymbol{\pi}^{z(t)}, \boldsymbol{\mu}^{z(t)}, \boldsymbol{\Theta}^{z(t)}) : t = 0, 1, 2, \ldots,\}$ increases monotonically to some $\bar{\mathcal{L}} < \infty$.

Given the ascent property in Theorem 4.1, we are ready to prove the local convergence result of our proposed EM algorithm in the following theorem. Theorem 4.2 extends the local convergence result of Städler, Bühlmann, and Van De Geer (2010) to the nonparametric finite mixture (1).

*Theorem 4.2.* Under the same conditions of Theorem 4.1, with probability tending to 1, our proposed generalized effective EM algorithm (i.e., Algorithm 1) achieves the local convergence. More specifically, for any given $z$, every cluster point $(\bar{\boldsymbol{\pi}}^z, \bar{\boldsymbol{\mu}}^z, \bar{\boldsymbol{\Theta}}^z)$ in the sequence $\{(\boldsymbol{\pi}^{z(t)}, \boldsymbol{\mu}^{z(t)}, \boldsymbol{\Theta}^{z(t)}) : t = 0, 1, 2, \ldots,\}$ is a stationary point of the objective function $\mathcal{L}_z(\boldsymbol{\pi}^z, \boldsymbol{\mu}^z, \boldsymbol{\Theta}^z)$ in (3) with probability tending to 1.

*Remark 4.1.* When learning the nonparametric finite mixture of Gaussian graphical models, Theorems 4.1 and 4.2 proved that both the ascent property and the local convergence hold for Algorithm 1 with probability tending to 1. It is obvious that Theorems 4.1 and 4.2 can be easily extended to the semiparametric finite mixture $\mathcal{G}^z(\mathbf{x}) = \pi_1(z)\mathcal{G}_1(\mathbf{x}) + \cdots + \pi_K(z)\mathcal{G}_K(\mathbf{x})$ in Remark 2.2. When there is no covariate $z$, we may further extend Theorems 4.1 and 4.2 and obtain the exact ascent property that $\mathcal{L}(\boldsymbol{\pi}^{(t+1)}, \boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Theta}^{(t+1)}) \ge \mathcal{L}(\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Theta}^{(t)})$ and the exact local convergence for the finite mixture of Gaussian graphical models $\mathcal{G}(\mathbf{x}) = \pi_1 \mathcal{G}_1(\mathbf{x}) + \cdots + \pi_K \mathcal{G}_K(\mathbf{x})$ proposed by Ruan, Yuan, and Zou (2011) in Remark 2.3.

### 4.2. Asymptotic Properties

Let $\boldsymbol{\omega}(z) = (\boldsymbol{\pi}(z), \boldsymbol{\mu}(z), \boldsymbol{\Theta}(z))$ be the true functional parameters in model (2). Define $\text{vec}(\cdot)$ as the vectorization of a matrix. We introduce some regularity conditions.

A. $\{(\mathbf{x}_n, z_n), n = 1, \ldots, N\}$ are independent and identically distributed as $(\mathbf{X}, Z)$. The support for $Z$, denoted by $\mathcal{Z}$, is compact subset of $\mathbb{R}^1$.

B. $\boldsymbol{\omega}(z)$ have continuous second derivatives, and $\pi_k(z) > 0$ for any $z \in \mathcal{Z}$.

C. The marginal density $f(z)$ of $Z$ is twice continuously differentiable and positive.

D. The kernel function $K(\cdot)$ is symmetric, continuous, and has a closed and bounded support and satisfy following conditions:

$$\int K(u)\,du = 1, \quad \int uK(u)\,du = 0, \quad \int u^2 K(u)\,du < \infty,$$
$$\int K^2(u)\,du < \infty, \quad \int |K^3(u)|\,du < \infty.$$

E. There exists a function $M(\mathbf{x})$ with $\mathrm{E}[M(\mathbf{X})] < \infty$, such that for all $\mathbf{x}$ and all $\boldsymbol{\omega}^z$ in a neighborhood of $\boldsymbol{\omega}(z)$, $|\partial^3 \ell(\boldsymbol{\omega}(z), \mathbf{x})/\partial\omega_i \partial\omega_j \partial\omega_l| < M(\mathbf{x})$ holds.

F. $\mathrm{E}(|\frac{\partial \ell(\boldsymbol{\omega}(z),\mathbf{X})}{\partial \omega_i}|^3) < \infty$ and $\mathrm{E}(\{\frac{\partial^2 \ell(\boldsymbol{\omega}(z),\mathbf{X})}{\partial\omega_i \partial\omega_j}\}^2) < \infty$ hold for all $i$ and $j$.

G. $h \to 0$, $Nh \to \infty$ and $Nh^5 = O(1)$ as $N \to \infty$.

Conditions A.–G. are mild, and they have been used in Mack and Silverman (1982), Zhou, Lafferty, and Wasserman (2010), Huang, Li, and Wang (2013), and many others. In particular, G. is a standard condition in the literature of local-likelihood estimation such as Loader (1996, 2006) and Huang, Li, and Wang (2013).

In the following theorem, we prove the asymptotic properties of the local solution $(\hat{\boldsymbol{\pi}}(z), \hat{\boldsymbol{\mu}}(z), \hat{\boldsymbol{\Theta}}(z))$ for the nonparametric finite mixture (2).

*Theorem 4.3.* If $\lambda = O((Nh)^{-1/2})$ as $N \to \infty$, under conditions A.–G., there exists a local maximizer $(\hat{\boldsymbol{\pi}}(z), \hat{\boldsymbol{\mu}}(z), \hat{\boldsymbol{\Theta}}(z))$ of (3) such that

$$\sqrt{Nh}(\hat{\boldsymbol{\pi}}(z) - \boldsymbol{\pi}(z)) = O_p(1), \quad \sqrt{Nh}(\hat{\boldsymbol{\mu}}(z) - \boldsymbol{\mu}(z)) = O_p(1),$$

and

$$\sqrt{Nh}(\text{vec}(\hat{\boldsymbol{\Theta}}(z)) - \text{vec}(\boldsymbol{\Theta}(z))) = O_p(1),$$

where $O_p(1)$ denotes the stochastic boundedness or boundedness in probability.

*Remark 4.2.* In view of Theorem 4.3, our proposed method delivers a nice local solution to estimate nonparametric mixing proportions $\boldsymbol{\pi}(z)$ and heterogenous graphical parameters $\boldsymbol{\mu}(z)$ and $\boldsymbol{\Theta}(z)$. In particular, when $\min_k \min_{(i,j)\in\mathcal{E}_k^z} |\theta_{kij}(z)| \gg (Nh)^{-1/2}$ is satisfied, the local maximizer $(\hat{\boldsymbol{\pi}}(z), \hat{\boldsymbol{\mu}}(z), \hat{\boldsymbol{\Theta}}(z))$ of (3) would recover all true edges and achieve the selection consistency, that is, $|\hat{\theta}_{kij}(z)| > 0$ for any $k$ and $(i, j) \in \mathcal{E}_k^z$.

## 5. Simulation Studies

This section presents two simulation studies. Section 5.1 considers a nonparametric finite mixture of autoregressive (AR) and block diagonal dependencies, and Section 5.2 considers a nonparametric finite mixture of AR and random sparse dependencies. For both simulation studies, we consider $K = 2$, $\mathcal{Z} = [0, 1]$, and two mixing proportion functions:

$$\pi_1(z) = \frac{\exp(0.5z)}{1 + \exp(0.5z)} \quad \text{and} \quad \pi_2(z) = 1 - \pi_1(z).$$

We also consider different mixing probabilities $\pi_1(z) = \frac{1}{12}\cos(2\pi z) + \frac{7}{12}$ and $\pi_2(z) = 1 - \pi_1(z)$, which result in similar numerical performance. For space consideration, we only present simulation studies with logistic mixing probabilities.

We generate data at the equally spaced grid points in $\mathcal{Z}$, and generate 50 observations at each grid point. We consider dimension $p = 50, 100$. The Epanechnikov kernel is used, and the initial values are obtained by using the method of Fraley and Raftery (2002) that was implemented in the R package “mclust.” To compare numerical performance, we define some average metrics over 100 replications:

- The averaged spectral norm loss: ASL $= \frac{1}{G}\sum_{z\in\mathcal{U}}\sum_{k=1}^{K} \|\hat{\boldsymbol{\Theta}}_k^z - \boldsymbol{\Theta}_k^z\|_2$.
- The averaged Frobenius norm loss: AFL $= \frac{1}{G}\sum_{z\in\mathcal{U}}\sum_{k=1}^{K} \|\hat{\boldsymbol{\Theta}}_k^z - \boldsymbol{\Theta}_k^z\|_\mathrm{F}$.
- The averaged Kullback–Leibler loss: AKL $= \frac{1}{G}\sum_{z\in\mathcal{U}}\sum_{k=1}^{K} \mathrm{KL}(\boldsymbol{\Theta}_k^{-1,z}, \hat{\boldsymbol{\Theta}}_k^{-1,z})$, where $\mathrm{KL}(\boldsymbol{\Theta}^{-1}, \hat{\boldsymbol{\Theta}}^{-1}) = \mathrm{tr}(\boldsymbol{\Theta}^{-1}\hat{\boldsymbol{\Theta}}) - \log|\boldsymbol{\Theta}^{-1}\hat{\boldsymbol{\Theta}}| - p$.
- The average squared error (RASE) for estimated mixing proportions:

$$\text{RASE}_\pi^2 = \frac{1}{G}\sum_{z\in\mathcal{U}}\sum_{k=1}^{K}\left(\hat{\pi}_k^z - \pi_k^z\right)^2.$$

- The average true positive rate (ATPR) and average false positive rate (AFPR):

$$\text{ATPR} = \frac{1}{G}\sum_{z\in\mathcal{U}}\frac{1}{K}\sum_{k=1}^{K}\text{TPR}_k^z \quad \text{and}$$
$$\text{AFPR} = \frac{1}{G}\sum_{z\in\mathcal{U}}\frac{1}{K}\sum_{k=1}^{K}\text{FPR}_k^z.$$

### 5.1. Mixture of AR and Block Diagonal Dependencies

In the first simulation study, we construct the mixture of AR(1) ($\boldsymbol{\Sigma}_1^0 = 0.4^{|i-j|}$) and a two-block diagonal structure $\boldsymbol{\Theta}_2^0$ when $z = 0$. When $p = 50$ ($p = 100$), we randomly specify 25 (50) edges from the first and the second block of $\boldsymbol{\Theta}_2^0$, respectively. The corresponding entries in $\boldsymbol{\Theta}_2^0$ are uniformly generated from $[-0.2, -0.1]$ and the diagonal elements are set to be 0.25. We fix the zero mean vector and 11-equally spaced grid points. Next, we consider the smooth evolving of $\boldsymbol{\Theta}_1^z$ and $\boldsymbol{\Theta}_2^z$ at the remaining 10 grid points: we randomly add 5 (20) new edges to first mixture toward the AR(2) structure, and add five new edges to

**Table 1.** Performance of our proposed BIC criterion in the 1st simulation study. The true number of mixtures is $K_0 = 2$.

| | $p = 50$ | | | $p = 100$ | | |
|---|---|---|---|---|---|---|
| $h$ | $K = 1$ | $K = 2$ | $K = 3$ | $K = 1$ | $K = 2$ | $K = 3$ |
| 0.65 | 82 | 18 | 0 | 91 | 9 | 0 |
| 0.85 | 16 | 84 | 0 | 3 | 97 | 0 |
| 1.05 | 2 | 96 | 2 | 0 | 100 | 0 |
| 1.25 | 1 | 97 | 2 | 1 | 99 | 0 |
| 1.45 | 1 | 97 | 2 | 0 | 100 | 0 |
| $\min_{\lambda,h}$ BIC | 1 | 97 | 2 | 0 | 100 | 0 |

each block of second mixture when $p = 50$ ($p = 100$). If necessary, we increase diagonal elements to keep the precision matrices positive definite.

First of all, we check the performance of choosing the number of mixtures using our proposed BIC criterion. We report the frequencies of $\min_\lambda$ BIC over 100 repeats. As shown in Table 1, undersmoothing in $h$ may cause an underestimated $K$. Hence, we should take into account different bandwidths, instead of some given bandwidth. Our proposed $\min_{\lambda,h}$ BIC performs best in choosing correct number of mixtures.

Next, we examine the performance of graphical model selection using ATPR and AFPR and of estimation using ASL, AFL, AKL, and RASE$_\pi$. We compare our proposed method to Zhou, Lafferty, and Wasserman (2010) and Ruan, Yuan, and Zou (2011), which are two special cases of our proposed method. The results are summarized in Tables 2 and 3. Overall, our proposed method clearly outperform two competitors, especially in terms of ATPR and estimation performance. It achieves the highest ATPR and a reasonably low AFPR. Both estimation- and selection- performances are less sensitive to bandwidths. Since two graphs evolves smoothly, it tends to give better result with larger bandwidth. But oversmoothing in $h$ may lead to biased estimation of mixing proportions.

**Table 2.** Comparison of graphical model selection performance in the 1st simulation study.

| | $p = 50$ | | $p = 100$ | |
|---|---|---|---|---|
| $h$ | ATPR | AFPR | ATPR | AFPR |
| | Proposed nonparametric finite mixture of Gaussian graphical models | | | |
| 0.65 | 0.9258 (0.0113) | 0.1906 (0.0011) | 0.8927 (0.0090) | 0.1072 (0.0006) |
| 0.85 | 0.9390 (0.0116) | 0.1674 (0.0011) | 0.9030 (0.0104) | 0.0920 (0.0005) |
| 1.05 | 0.9481 (0.0112) | 0.1561 (0.0011) | 0.9081 (0.0105) | 0.0845 (0.0005) |
| 1.25 | 0.9523 (0.0107) | 0.1519 (0.0010) | 0.9110 (0.0098) | 0.0810 (0.0006) |
| 1.45 | 0.9550 (0.0103) | 0.1503 (0.0010) | 0.9136 (0.0094) | 0.0804 (0.0005) |
| | Simple mixture of Gaussian graphical models (Ruan, Yuan, and Zou 2011). | | | |
| – | 0.8677 (0.0025) | 0.0751 (0.0007) | 0.6537 (0.0021) | 0.0207 (0.0002) |
| | Time varying Gaussian graphical models (Zhou, Lafferty, and Wasserman 2010) | | | |
| 0.65 | 0.8498 (0.0019) | 0.1055 (0.0007) | 0.6527 (0.0016) | 0.0346 (0.0002) |
| 0.85 | 0.8600 (0.0020) | 0.0883 (0.0007) | 0.6533 (0.0017) | 0.0262 (0.0002) |
| 1.05 | 0.8653 (0.0022) | 0.0799 (0.0007) | 0.6539 (0.0018) | 0.0225 (0.0002) |
| 1.25 | 0.8670 (0.0023) | 0.0769 (0.0007) | 0.6541 (0.0018) | 0.0214 (0.0002) |
| 1.45 | 0.8676 (0.0023) | 0.0759 (0.0007) | 0.6541 (0.0019) | 0.0210 (0.0002) |

NOTE: Ruan, Yuan, and Zou (2011) does not involve the bandwidth selection.

## 5.2. *Mixture of AR and Random Sparse Dependencies*

In the second simulation study, we consider the mixture of AR(1) ($\mathbf{\Sigma}_1^0 = 0.4^{|i-j|}$) and a random sparse structure $\mathbf{\Theta}_2^0$. When $p = 50$ ($p = 100$), we set the diagonal elements of $\mathbf{\Theta}_2^0$ as 0.25 and randomly choose 55 (100) off-diagonal entries to be drawn uniformly from $[-0.25, -0.22]$. We fix the zero mean vector and 11 grid points. Here, we consider the less smooth evolving of $\mathbf{\Theta}_1^z$ and $\mathbf{\Theta}_2^z$: at each of the remaining grid points, we randomly add 5 (20) new edges to both graphs and randomly remove 5 (20) existing edges when $p = 50$ ($p = 100$). If necessary, we increase diagonal elements to keep the precision matrices positive definite.

Table 4 summarizes the performance of choosing the number of mixtures. As shown in Table 4, undersmoothing in $h$ again may cause an underestimated $K$. Combining different

**Table 3.** Comparison of graphical model estimation performance in the 1st simulation study.

| $p$ | $h$ | ASL | AFL | AKL | RASE$_\pi$ |
|---|---|---|---|---|---|
| | | Proposed nonparametric finite mixture of Gaussian graphical models | | | |
| 50 | 0.65 | 1.9036 (0.0290) | 6.7950 (0.1320) | 10.7644 (0.8102) | 0.1063 (0.0052) |
| | 0.85 | 1.8572 (0.0290) | 6.5677 (0.1348) | 9.5934 (0.8485) | 0.0972 (0.0050) |
| | 1.05 | 1.8442 (0.0279) | 6.4717 (0.1304) | 9.0155 (0.8427) | 0.0988 (0.0047) |
| | 1.25 | 1.8398 (0.0267) | 6.4272 (0.1246) | 8.7180 (0.8077) | 0.0995 (0.0044) |
| | 1.45 | 1.8383 (0.0259) | 6.4055 (0.1205) | 8.5558 (0.7809) | 0.0993 (0.0041) |
| 100 | 0.65 | 2.0519 (0.0170) | 10.0884 (0.1156) | 20.5775 (1.0949) | 0.0668 (0.0040) |
| | 0.85 | 2.0186 (0.0170) | 9.9191 (0.1260) | 19.3055 (1.3147) | 0.0628 (0.0042) |
| | 1.05 | 2.0129 (0.0170) | 9.8929 (0.1299) | 18.9758 (1.4092) | 0.0648 (0.0039) |
| | 1.25 | 2.0105 (0.0164) | 9.8738 (0.1250) | 18.6405 (1.3602) | 0.0640 (0.0030) |
| | 1.45 | 2.0091 (0.0159) | 9.8648 (0.1224) | 18.5014 (1.3464) | 0.0667 (0.0034) |
| | | Simple mixture of Gaussian graphical models (Ruan, Yuan, and Zou 2011) | | | |
| 50 | – | 2.3119 (0.0021) | 8.6222 (0.0031) | 17.8072 (0.0153) | 0.6250 (0.0031) |
| 100 | – | 2.5324 (0.0024) | 13.1246 (0.0034) | 40.8719 (0.0188) | 0.6336 (0.0027) |
| | | Time varying Gaussian graphical models (Zhou, Lafferty, and Wasserman 2010) | | | |
| 50 | 0.65 | 2.2896 (0.0025) | 8.5715 (0.0032) | 17.7900 (0.0154) | – |
| | 0.85 | 2.2948 (0.0024) | 8.5825 (0.0031) | 17.7274 (0.0154) | – |
| | 1.05 | 2.3007 (0.0023) | 8.5964 (0.0031) | 17.7324 (0.0154) | – |
| | 1.25 | 2.3044 (0.0022) | 8.6054 (0.0031) | 17.7509 (0.0153) | – |
| | 1.45 | 2.3065 (0.0022) | 8.6102 (0.0031) | 17.7647 (0.0153) | – |
| 100 | 0.65 | 2.5298 (0.0025) | 13.0496 (0.0035) | 40.6013 (0.0187) | – |
| | 0.85 | 2.5297 (0.0025) | 13.0743 (0.0035) | 40.6280 (0.0188) | – |
| | 1.05 | 2.5304 (0.0024) | 13.0953 (0.0035) | 40.7040 (0.0188) | – |
| | 1.25 | 2.5311 (0.0024) | 13.1065 (0.0034) | 40.7594 (0.0188) | – |
| | 1.45 | 2.5314 (0.0024) | 13.1121 (0.0034) | 40.7911 (0.0188) | – |

NOTE: Zhou, Lafferty, and Wasserman (2010) does not estimate mixing probabilities.

Table 4. Performance of our proposed BIC criterion in the 2nd simulation study. The true number of mixtures is $K_0 = 2$.

| | $p = 50$ | | | $p = 100$ | | |
|---|---|---|---|---|---|---|
| $h$ | $K = 1$ | $K = 2$ | $K = 3$ | $K = 1$ | $K = 2$ | $K = 3$ |
| 0.65 | 95 | 5 | 0 | 100 | 0 | 0 |
| 0.85 | 23 | 77 | 0 | 5 | 95 | 0 |
| 1.05 | 10 | 89 | 1 | 1 | 99 | 0 |
| 1.25 | 4 | 94 | 2 | 0 | 100 | 0 |
| 1.45 | 0 | 97 | 3 | 0 | 100 | 0 |
| $\min_{\lambda,h}$ BIC | 0 | 97 | 3 | 0 | 100 | 0 |

Table 5. Comparison of graphical model selection performance in the 2nd simulation study.

| | $p = 50$ | | $p = 100$ | |
|---|---|---|---|---|
| $h$ | ATPR | AFPR | ATPR | AFPR |
| Proposed nonparametric finite mixture of Gaussian graphical models | | | | |
| 0.65 | 0.8732 (0.0160) | 0.1769 (0.0011) | 0.8890 (0.0113) | 0.1095 (0.0005) |
| 0.85 | 0.8859 (0.0177) | 0.1555 (0.0013) | 0.9042 (0.0118) | 0.0923 (0.0005) |
| 1.05 | 0.8889 (0.0183) | 0.1442 (0.0014) | 0.9098 (0.0118) | 0.0836 (0.0006) |
| 1.25 | 0.8888 (0.0184) | 0.1398 (0.0014) | 0.9117 (0.0115) | 0.0803 (0.0006) |
| 1.45 | 0.8894 (0.0185) | 0.1382 (0.0014) | 0.9124 (0.0113) | 0.0793 (0.0006) |
| Simple mixture of Gaussian graphical models (Ruan, Yuan, and Zou 2011). | | | | |
| – | 0.8504 (0.0023) | 0.0738 (0.0007) | 0.6926 (0.0019) | 0.0250 (0.0002) |
| Time varying Gaussian graphical models (Zhou, Lafferty, and Wasserman 2010) | | | | |
| 0.65 | 0.8389 (0.0020) | 0.1020 (0.0007) | 0.6919 (0.0015) | 0.0406 (0.0002) |
| 0.85 | 0.8457 (0.0021) | 0.0863 (0.0007) | 0.6935 (0.0016) | 0.0312 (0.0002) |
| 1.05 | 0.8493 (0.0022) | 0.0784 (0.0006) | 0.6935 (0.0017) | 0.0271 (0.0002) |
| 1.25 | 0.8503 (0.0022) | 0.0757 (0.0006) | 0.6936 (0.0018) | 0.0257 (0.0002) |
| 1.45 | 0.8506 (0.0022) | 0.0746 (0.0006) | 0.6933 (0.0019) | 0.0253 (0.0002) |

NOTE: Ruan, Yuan, and Zou (2011) does not involve the bandwidth selection.

bandwidths, our proposed BIC criterion has a consistent performance in choosing the correct number of mixtures. The graphical model selection and the estimation performance are summarized in Tables 5 and 6. Overall, our proposed method clearly outperform two competitors. Here, both estimation and selection performances are relatively more sensitive to bandwidths. Oversmoothing in $h$ may lead to biased estimation of both precision matrices and mixing proportions.

Based on simulation studies, our proposed method clearly outperforms Zhou, Lafferty, and Wasserman (2010) and Ruan, Yuan, and Zou (2011) in the presence of nonparametric finite mixtures. The numerical evidence is consistent with theoretical properties of our proposed method.

## 6. A Real Application To ADHD-200 Imaging Data

This section applies our proposed method to estimate time-varying brain functional connectivity from ADHD-200 Global Competition data (Biswal et al. 2010). The ADHD-200 training dataset has fMRI images, diagnosis information, and other demographic variables (e.g., age, IQ, gender, and handedness) of 776 subjects from eight different sites. We focus on $N = 284$ subjects whose ages range from 9 to 13, since observed ages are not uniformly distributed outside the range 9–13. Each fMRI image has measurements for $p = 264$ seed voxels. Among these subjects, 186 subjects are typically developing controls and 98 subjects are diagnosed with ADHD. We choose age as the covariate $Z$, and normalize it to [0, 1]. We consider three interested ages: 9, 11, and 13.

We first determine the number of mixtures by using our proposed BIC in Section 3.2. BIC is minimized when $K = 2$, which implies two heterogenous groups. Then, we estimated heterogeneous graphical models at the aforementioned interested ages, respectively. Table 7 shows the estimated mixing proportions.

Table 7. Estimated mixing proportions at age 9, 11, and 13.

| | Age 9 | Age 11 | Age 13 |
|---|---|---|---|
| $\hat{\pi}_1$ | 0.6742 | 0.6870 | 0.7011 |
| $\hat{\pi}_2$ | 0.3258 | 0.3130 | 0.2989 |

Table 6. Comparison of graphical model estimation performance in the 2nd simulation study.

| $p$ | $h$ | ASL | AFL | AKL | $\text{RASE}_\pi$ |
|---|---|---|---|---|---|
| Proposed nonparametric finite mixture of Gaussian graphical models | | | | | |
| 50 | 0.65 | 2.1992 (0.0352) | 7.4927 (0.1548) | 12.7527 (0.9811) | 0.1290 (0.0058) |
| | 0.85 | 2.1312 (0.0362) | 7.2900 (0.1646) | 11.8158 (1.0693) | 0.1270 (0.0060) |
| | 1.05 | 2.1031 (0.0360) | 7.2195 (0.1680) | 11.5378 (1.1045) | 0.1260 (0.0058) |
| | 1.25 | 2.0973 (0.0359) | 7.2041 (0.1683) | 11.4536 (1.1069) | 0.1261 (0.0055) |
| | 1.45 | 2.0965 (0.0357) | 7.2018 (0.1682) | 11.4282 (1.1064) | 0.1264 (0.0054) |
| 100 | 0.65 | 2.2518 (0.0201) | 10.0592 (0.1270) | 20.0166 (1.1550) | 0.0695 (0.0047) |
| | 0.85 | 2.2053 (0.0186) | 9.8330 (0.1261) | 18.2240 (1.2288) | 0.0595 (0.0046) |
| | 1.05 | 2.1981 (0.0174) | 9.7937 (0.1237) | 17.7696 (1.2573) | 0.0599 (0.0038) |
| | 1.25 | 2.2014 (0.0169) | 9.7985 (0.1221) | 17.6933 (1.2571) | 0.0621 (0.0034) |
| | 1.45 | 2.2039 (0.0166) | 9.8037 (0.1209) | 17.6861 (1.2501) | 0.0647 (0.0034) |
| Simple mixture of Gaussian graphical models (Ruan, Yuan, and Zou 2011) | | | | | |
| 50 | – | 2.6246 (0.0028) | 8.8763 (0.0045) | 16.0568 (0.0146) | 0.6264 (0.0029) |
| 100 | – | 2.7834 (0.0022) | 13.0387 (0.0038) | 36.7745 (0.0173) | 0.6287 (0.0028) |
| Time varying Gaussian graphical models (Zhou, Lafferty, and Wasserman 2010) | | | | | |
| 50 | 0.65 | 2.6187 (0.0028) | 8.8399 (0.0046) | 16.1026 (0.0148) | - |
| | 0.85 | 2.6203 (0.0028) | 8.8517 (0.0046) | 16.0414 (0.0149) | - |
| | 1.05 | 2.6220 (0.0028) | 8.8619 (0.0046) | 16.0298 (0.0149) | - |
| | 1.25 | 2.6230 (0.0028) | 8.8674 (0.0046) | 16.0337 (0.0148) | - |
| | 1.45 | 2.6235 (0.0028) | 8.8702 (0.0046) | 16.0386 (0.0148) | - |
| 100 | 0.65 | 2.7776 (0.0023) | 12.9454 (0.0037) | 36.3898 (0.0169) | - |
| | 0.85 | 2.7781 (0.0022) | 12.9779 (0.0038) | 36.4454 (0.0172) | - |
| | 1.05 | 2.7798 (0.0022) | 13.0040 (0.0038) | 36.5506 (0.0172) | - |
| | 1.25 | 2.7810 (0.0022) | 13.0178 (0.0038) | 36.6256 (0.0172) | - |
| | 1.45 | 2.7817 (0.0022) | 13.0244 (0.0038) | 36.6678 (0.0173) | - |

NOTE: Zhou, Lafferty, and Wasserman (2010) did not estimate mixing probabilities.

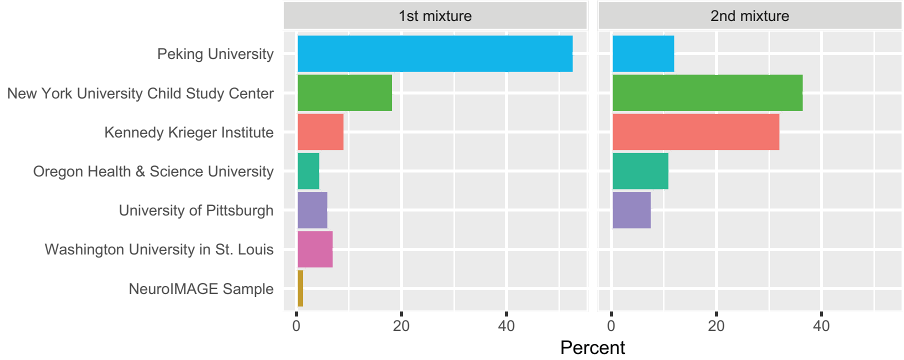

**Figure 2.** Site information: 1st mixture (left panel) and 2nd mixture (right panel).

We can see a slightly increasing trend in first mixture proportion as age increases.

If the covariate $Z$ is ignored, we may fail to account for heterogeneity in the ADHD-200 data. To illustrate this, we also estimate the simple mixture of Gaussian mixture models with $K = 2$. Then, we obtain that $\hat{\pi}_1 = 0.9509$ and $\hat{\pi}_2 = 0.0491$, which do not reveal any meaningful heterogeneous subpopulation. In what follows, we will investigate two heterogenous dependencies identified by our method through profiling their demographic variables, and explain their differences from three different perspectives: site information, impact of gender, and impact of IQ. We shall show that our results are supported by existing scientific findings.

First, we explore how subpopulations are composed with subjects from various sites. As shown in Figure 2, both mixtures are formed with subjects from heterogeneous locations. This confirms the previous study that geographic locations are not sufficient to explain variability of ADHD ( Polanczyk et al. 2007). Hence, we need to further investigate their individual commonality beyond geographic locations.

Second, we study the relationship between gender and ADHD status (i.e., whether the subject has ADHD or not). In the existing literature, gender difference in ADHD is an important research topic (Arnold 1996). There are two main arguments about the reason why the boys are more likely to be diagnosed with ADHD than girls. On one hand, some researchers insist that there is actually gender difference in ADHD. On the other hand, other studies show that there are other issues for example, other demographic covariates or the bias of the diagnose test that affects diagnose results. Here, we test the independence between gender and ADHD status in two mixture, respectively. Their corresponding contingency tables are given in Table 8. For the first mixture, we compute the chi-square test statistic, $\chi^2 = 7.3106$ with associated $p$ value $= 0.0069$. It implies that there is an indeed relationship between gender and ADHD status in this mixture. While for the second mixture, we have chi-square test statistic, $\chi^2 = 0.5237$ with associated $p$ value $= 0.4693$, which indicates that gender and ADHD status could be independent in the second mixture. Therefore, we successfully identify two heterogeneous subpopulations: the first one is consistent with the previous study that ADHD is diagnosed at a significantly higher rate in boys than in girls (Sauver et al. 2004); the second one is consistent with another study that there may exist other covariates strongly related to the ADHD, but which are not different for boys and girls so that ADHD are not systematically different for boys and girls (Bauermeister et al. 2007). Both subpopulations support their corresponding scientific findings, respectively.

**Table 8.** Contingency tables with gender and whether they have ADHD or not. 1st mixture (left table) and 2nd mixture (right table).

| | Male | Female | Total | | Male | Female | Total |
|---|---|---|---|---|---|---|---|
| Control | 77 | 57 | 134 | Control | 30 | 22 | 52 |
| ADHD | 48 | 13 | 61 | ADHD | 25 | 12 | 37 |
| Total | 125 | 70 | 195 | Total | 55 | 34 | 89 |

**Table 9.** Average full scale IQ scores for both mixtures.

| | 1st mixture | 2nd mixture |
|---|---|---|
| Male control | 118.85 | 112.30 |
| Female control | 115.54 | 111.77 |
| Male ADHD | 106.53 | 108.00 |
| Female ADHD | 99.54 | 104.00 |

Third, we investigate the full scale IQ scores for both mixtures. Table 9 summarizes the average full scale IQ for each mixture. For the first mixture, we can clearly see typically developing children have a higher average IQ score than the children with ADHD: female control's average IQ is 16 higher than female ADHDs average IQ, and male control's average IQ is 12.32 higher than male ADHDs average IQ. This result shows that full scale IQ scores are reliably different between individuals with ADHD and typically developing controls in the first mixture, which is consistent with Frazier, Demaree, and Youngstrom (2004); García-Sánchez et al. (1997). However, the second mixture exhibits a very different aspect: typically developing children in the second mixture have a lower average IQ score compared to those in the first mixture, while ADHD children in the second mixture have a higher average IQ score than those in the first mixture. Therefore, there is only no significant difference in terms of average IQ between typically growing children and children with ADHD in the second mixture: female control's average IQ is 7.77 higher than female ADHDs average IQ, and male control's average IQ is 4.30 higher than male ADHDs average IQ. Based on the full scale IQ, these two heterogeneous mixtures further justifies the power of our proposed method.

## Supplementary Materials

Algorithm Details and Proofs: This supplement consists of the technical details of Algorithm 1 and the proofs of Theorems 1-3. (UTCH_S_1408497.pdf)

## Acknowledgments

The authors thank the editor, an associate editor, and two referees for their constructive comments and suggestions. Lingzhou Xue's research is partially supported by the National Science Foundation grant DMS-1505256.

## Funding

Funding was provided by the Directorate for Mathematical and Physical Sciences [1505256] .

## ORCID

Lingzhou Xue http://orcid.org/0000-0002-8252-0637

## References

Ahmed, A., and Xing, E. P. (2009), "Recovering Time-Varying Networks of Dependencies in Social and Biological Studies," *Proceedings of the National Academy of Sciences*, 106(29), 11878–11883. [512,513]

Arnold, L. E. (1996), "Sex Differences in ADHD: Conference Summary," *Journal of Abnormal Child Psychology*, 24(5), 555–569. [519]

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017), "Statistical Guarantees for the em Algorithm: From Population to Sample-Based Analysis." *The Annals of Statistics*, 45(1), 77–120. [513]

Bauermeister, J. J., Shrout, P. E., Chávez, L., RubioStipec, M., Ramírez, R., Padilla, L., Anderson, A., García, P., and Canino, G. (2007), "ADHD and Gender: Are Risks and Sequela of ADHD the same for Boys and Girls?" *Journal of Child Psychology and Psychiatry*, 48(8), 831–839. [511,519]

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., Dogonowski, A. M. (2010), "Toward Discovery Science of Human Brain Function," *Proceedings of the National Academy of Sciences*, 107(10), 4734–4739. [511,518]

Brillinger, D. R. (1977), Discussion of "Consistent Nonparametric Regression" by C. J. Stone, *The Annals of Statistics*, 5(4), 622–623. [512]

Cai, T., Liu, W., and Luo, X. (2011), "A Constrained $\ell_1$ Minimization Approach to Sparse Precision Matrix Estimation,." *Journal of the American Statistical Association*, 106(494), 594–607. [511]

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012), "Latent Variable Graphical Model Selection via Convex Optimization" (with discussion), *The Annals of Statistics*, 40(4), 1935–1967. [511]

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, 76(2), 373–397. [511]

Dempster, A. (1972), "Covariance Selection," *Biometrics*, 28, 157–175. [512]

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004), "Sparse Graphical Models for Exploring Gene Expression Data," *Journal of Multivariate Analysis*, 90(1), 196–212. [511]

Fan, J., Feng, Y., and Wu, Y. (2009), "Network Exploration via the Adaptive Lasso and Scad Penalties," *The Annals of Applied Statistics*, 3(2), 521–541. [513]

Fan, J., Xue, L., and Zou, H. (2014), "Strong Oracle Optimality of Folded Concave Penalized Estimation," *The Annals of Statistics*, 42(3), 819–849. [513]

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics* 29, 153–193. [515]

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97(458), 611–631. [516]

Frazier, T. W., Demaree, H. A., and Youngstrom, E. A. (2004), "Meta-Analysis of Intellectual and Neuropsychological Test Performance in Attention-deficit/Hyperactivity Disorder," *Neuropsychology*, 18(3), 543. [519]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9(3), 432–441. [511,514]

García-Sánchez, C., Estévez-González, A., Suárez-Romero, E., and Junqué, C. (1997), "Right Hemisphere Dysfunction in Subjects with Attention-Deficit Disorder with and Without Hyperactivity," *Journal of Child Neurology*, 12(2), 107–115. [519]

Goldfarb, D., Ma, S., and Scheinberg, K. (2013), "Fast Alternating Linearization Methods for Minimizing the Sum of Two Convex Functions," *Mathematical Programming* 141(1–2), 349–382. [514]

Huang, M., Li, R., and Wang, S. (2013), "Nonparametric Mixture of Regression Models," *Journal of the American Statistical Association*, 108(503), 929–941. [512,515,516]

International HapMap 3 Consortium. (2010), "Integrating Common and Rare Genetic Variation in Diverse Human Populations," *Nature*, 467(7311), 52–58. [511]

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010), "Estimating Time-Varying Networks," *The Annals of Applied Statistics*, 4(1), 94–123. [512,513]

Lindsay, B. G. (1995), "Mixture Models: Theory, Geometry and Applications," Presented at the NSF-CBMS Regional Conference Series in Probability and Statistics. [512]

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012), "High-Dimensional Semiparametric Gaussian Copula Graphical Models," *The Annals of Statistics*, 40(4), 2293–2326. [511]

Loader, C. R. (1996), "Local Likelihood Density Estimation," *The Annals of Statistics*, 24(4), 1602–1618. [512,516]

——— (2006), *Local Regression and Likelihood*, New York, NY: Springer. [512,516]

Ma, S., Xue, L., and Zou, H. (2013), "Alternating Direction Methods for Latent Variable Gaussian Graphical Model Selection," *Neural Computation*, 25(8), 2172–2198. [511]

Mack, Y. P., and Silverman, B. W. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61(3), 405–415. [516]

Mallapragada, P. K., Jin, R., and Jain, A. (2010), "Non-Parametric Mixture Models for Clustering," In: Hancock, E. R., Wilson, R. C., Windeatt, T., Ulusoy, I., Escolano, F. (eds), *Structural, Syntactic, and Statistical Pattern Recognition*, 334–343. SSPR/SPR 2010. Lecture Notes in Computer Science, vol 6218. 334–343, Springer-Verlag. [512]

McLachlan, G., and Peel, D. (2004), *Finite Mixture Models*, Wiley series in probabilities and statistics. New York, NY: Wiley-Interscience. [512]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 34(3), 1436–1462. [511]

Ng, B., Varoquaux, G., Poline, J. B., and Thirion, B. (2013), "A Novel Sparse Group Gaussian Graphical Model for Functional Connectivity Estimation," In: Gee, J. C., Joshi, S., Pohl, K. M., Wells, W. M., Zöllei, L. (eds) *Information Processing in Medical Imaging*, 256–267. IPMI 2013. Lecture Notes in Computer Science, vol 7917. Springer, Berlin, Heidelberg. [511]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of the American Statistical Association*, 104(486), 735–746. [511]

Polanczyk, G., de Lima, M. S., Horta, B. L., Biederman, J., and Rohde, L. A. (2007), "The Worldwide Prevalence of ADHD: A Systematic Review and Metaregression Analysis," *The American Journal of Psychiatry*, 164(6), 942–948. [519]

Rodríguez, A., Lenkoski, A., and Dobra, A. (2011), "Sparse Covariance Estimation in Heterogeneous Samples," *Electronic Journal of Statistics*, 5, 981–1014. [512]

Ruan, L., Yuan, M., and Zou, H. (2011), "Regularized Parameter Estimation in High-Dimensional Gaussian Mixture Models," *Neural Computation*, 23(6), 1605–1622. [512,513,516,517,518]

Ryali, S., Chen, T., Supekar, K., and Menon, V. (2012), "Estimation of Functional Connectivity in fmri Data using Stability Selection-Based Sparse Partial Correlation with Elastic Net Penalty," *Neuroimage*, 59(4), 3852–3861. [511]

Sauver, J. L. S., Barbaresi, W. J., Katusic, S. K., Colligan, R. C., Weaver, A. L., and Jacobsen, S. J. (2004), "Early Life Risk Factors for Attention-Deficit/Hyperactivity Disorder: A Population-Based Cohort Study," *Mayo Clinic Proceedings*, 79(9), 1124–1131. [511,519]

Schäfer, J., and Strimmer, K. (2005), “An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks,” *Bioinformatics*, 21(6), 754–764. [511]

Städler, N., Bühlmann, P., and Van De Geer, S. (2010), “$\ell_1$-penalization for Mixture Regression Models” (with discussion), *Test*, 19(2), 209–256. [512,515]

Tibshirani, R., and Hastie, T. (1987), “Local Likelihood Estimation.” *Journal of the American Statistical Association*, 82(398), 559–567. [512]

Tseng, P. (2001), “Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization,” *Journal of Optimization Theory and Applications*, 109(3), 475–494. [515]

Varoquaux, G., Gramfort, A., Poline, J. B., and Thirion, B. (2010), “Brain Covariance Selection: Better Individual Functional Connectivity Models using Population Prior.” In: Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds) *Advances in Neural Information Processing Systems*, 23, 2334–2342. Curran Associates, Inc, New York. [511]

Voorman, A., Shojaie, A., and Witten, D. (2014), “Graph Estimation with Joint Additive Models,” *Biometrika*, 101(1), 85–101. [511]

Wang, P., Chao, D. L., and Hsu, L. (2011), “Learning Oncogenic Pathways from Binary Genomic Instability Data,” *Biometrics*, 67(1), 164–173. [511]

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(4), 892–900. [511,514]

Wu, C. J. (1983), “On the Convergence Properties of the em Algorithm,” *The Annals of Statistics*, 11, 95–103. [513]

Xue, L., and Zou, H. (2012), “Regularized Rank-Based Estimation of High-Dimensional Nonparanormal Graphical Models,” *The Annals of Statistics*, 40(5), 2541–2571. [511]

Xue, L., Zou, H., and Cai, T. (2012), “Nonconcave Penalized Composite Conditional Likelihood Estimation of Sparse Ising Models,” *The Annals of Statistics*, 40(3), 1403–1429. [511]

Yuan, M., and Lin, Y. (2007), “Model Selection and Estimation in the Gaussian Graphical Model,” *Biometrika*, 94(1), 19–35. [511]

Zhou, S., Lafferty, J., and Wasserman, L. (2010), “Time Varying Undirected Graphs,” *Machine Learning*, 80(2), 295–319. [512,513,516,517,518]