HHS Public Access

Author manuscript

J Comput Graph Stat. Author manuscript; available in PMC 2016 October 01.

Published in final edited form as:

J Comput Graph Stat. 2015 October 1; 24(4): 954–974. doi:10.1080/10618600.2014.956876.

Gene regulation network inference with joint sparse Gaussian graphical models

Hyonho Chun, **Xianghua Zhang**, and **Hongyu Zhao**

Abstract

Revealing biological networks is one key objective in systems biology. With microarrays, researchers now routinely measure expression profiles at the genome level under various conditions, and, such data may be utilized to statistically infer gene regulation networks. Gaussian graphical models (GGMs) have proven useful for this purpose by modeling the Markovian dependence among genes. However, a single GGM may not be adequate to describe the potentially differing networks across various conditions, and hence it is more natural to infer multiple GGMs from such data. In the present study, we propose a class of nonconvex penalty functions aiming at the estimation of multiple GGMs with a flexible *joint sparsity* constraint. We illustrate the property of our proposed nonconvex penalty functions by simulation study. We then apply the method to a gene expression data set from the GenCord Project, and show that our method can identify prominent pathways across different conditions.

Keywords

Gaussian graphical models; gene regulation networks; microarrays; gene expression; non-convex penalty; pathways

1 Introduction

Recent advances in high-throughput technology make it possible to simultaneously measure tens of thousands of molecular components. Researchers now routinely collect expression profiles at the genome level under various conditions and infer gene regulation networks by analyzing these datasets with various statistical methods such as Bayesian networks, relevance networks, and Gaussian graphical models (GGMs). Among these methods, we focus on the GGMs, because they have proven among the best in inferring conditional dependence (Markov dependence) networks (Werhli et al., 2006; Soranzo et al., 2007).

When the datasets from multiple conditions are available, it is important to improve the power of the study by modeling all the data so as to effectively accommodate characteristics of the datasets. One characteristic that we incorporate in our approach is *joint sparsity*, which describes the fact that the number of regulations in a biological network is far less than that of a fully connected network, and this sparsity is preserved across multiple conditions. For example, the regulations curated in the KEGG pathway database have tree structures and the established connections among genes only represent a very small fraction of all possible connections.

The joint sparsity principle has been utilized in other multiple GGM approaches with various penalty functions. Chiquet et al. (2009) proposed to use a group lasso penalty, but their approach does not allow the network structure change across conditions. Later, Guo et al. (2011) proposed a group bridge penalization, and it indeed produces network structures that vary across conditions. Hence, the group bridge penalization is preferred in case of estimating multiple gene regulation networks with datasets from multiple tissues/conditions. More recently, Danaher et al. (2012) proposed a joint graphical lasso approach, where they used various ℓ_1 regularization methods for promoting graph similarities. The formulation is convex and can be useful for very high-dimensional problems.

Among these previous approaches, we find that Guo et al. (2011)'s approach can be extended to a wider class of penalty functions, which consists of nonconvex functions. In fact, our proposed class of noncovex penalty functions gives a flexibility in controlling the level of joint sparsity by the choice of the penalty function. Since the level of joint sparsity among multiple biological networks might be much higher (or lower) than what is specified in Guo et al. (2011)'s approach, the broader class of penalty functions in our approach can gain power by putting more (or less) weights on the common network structures.

The rest of this article is organized as follows: In Section 2, we provide a detailed description of our joint estimation procedure with nonconvex penalty functions and show how the level of joint sparsity can be controlled through these penalty functions. We also present the consistency and sparsistency results of the estimate in this section. In Section 3, we show the performance of the methods under various scenarios via simulation, and then in Section 4, we apply our approach to the microarray dataset from the GenCord project (Dimas et al., 2009), which reveals prominent pathways across different cell types in umbilical cords. A brief conclusion follows in Section 5.

2 Estimation of multiple Gaussian graphical models with joint sparsity

In this section, we first review GGMs briefly and then formulate multiple GGMs with joint sparsity that is achieved by using a nonconvex penalty function.

2.1 Brief review of GGMs

A graphical model encodes conditional independence relationships among multiple random variables, $X_1, \ldots, X_p$, by using a graph $\mathcal{G} = (\mathbf{\Gamma}, \mathbf{E})$, where $\mathbf{\Gamma}$ is an index set for vertices and $\mathbf{E}$ is a subset of $\mathbf{\Gamma} \times \mathbf{\Gamma}$ for edges. Under the graphical model, a pair of random variables X_i and X_j are conditionally independent given all the rest if and only if there is no edge between vertices i and j on the graph. Inferring conditional relationships among random variables is not a simple task, because it involves investigation of the joint density factorization. However, if $\mathbf{X} = (X_1, \ldots, X_p)'$ is assumed to follow a multivariate normal distribution $N(0, \mathbf{\Omega}^{-1})$, where $\mathbf{\Omega}$ is the inverse covariance matrix and u' denote the transpose of a vector u, such conditional independence relationships can be directly read from the zero elements of $\mathbf{\Omega}$ (Lauritzen, 1996). Thus, $\omega_{i,j} = 0$ if and only if X_i and X_j are conditionally independent given all the other variables, where $\omega_{i,j}$ is the (i, j)th element of $\mathbf{\Omega}$. Because of this property, the network inference problem is considered as a sparse precision matrix estimation problem under GGMs. The sparse GGM estimation has been extensively studied recently including

Meinshausen and Buhlmann (2006); Yuan and Lin (2007); Peng et al. (2009); Lam and Fan (2009); and Guo et al. (2011).

There are non-likelihood-based approaches including graphical Dantzig selector (Yuan, 2010) and CLIME (Cai et al., 2011). The graphical Dantzig selector improves the pseudo likelihood approach of Meinshausen and Buhlmann (2006). CLIME tries to minimize $|\Omega^{-1} - \mathbf{S}|_\infty$ instead of the negative log likelihood, where $\mathbf{S}$ is the sample covariance matrix and $|\mathbf{A}|_\infty$ is a matrix max norm for a matrix $\mathbf{A}$. It has been shown that both approaches perform well computationally as well as asymptotically. It is possible that one can apply a joint sparsity constraint under various loss functions. However, we do not pursue these approaches in this manuscript, since we are interested in improving the regularization in order to achieve flexible joint sparsity.

2.2 Nonconvex penalty functions for joint sparsity

We consider multiple GGMs across T conditions. Specifically, we assume that a p dimensional random vector $\mathbf{X}^{t,i} \sim N_p(0,(\mathbf{\Omega}^t)^{-1})$ independently, for $i = 1, \ldots, n_t$ and $t = 1, \ldots, T$. The negative log likelihood can be written as

$$L(\{\mathbf{\Omega}^t\}_{t=1}^T) = \sum_{t=1}^{T} \frac{n_t}{2} (\mathrm{tr}(\mathbf{S}^t\mathbf{\Omega}^t) - \log\det(\mathbf{\Omega}^t)),$$

where $\mathbf{S}^t$ and $\mathbf{\Omega}^t$ are sample covariance and precision matrices for the tth condition and tr($\mathbf{A}$) and det($\mathbf{A}$) denote trace and determinant of a matrix $\mathbf{A}$, respectively.

Motivated by the property that the number of edges in a biological network is far less than that of a fully connected network (e.g. a pathway from KEGG database is often represented as a tree which has $p - 1$ edges for p nodes) and that the sparse structure tends to be preserved across multiple conditions, we attempt to improve the accuracy of GGM estimation by employing joint sparsity regularization. Such regularization is achieved by introducing sparsity into the precision matrix through nonconvex penalty functions. The penalized negative log likelihood (PL) is defined as follows:

$$PL_i(\{\mathbf{\Omega}^t\}_{t=1}^T) = L(\{\mathbf{\Omega}^t\}_{t=1}^T) + \lambda \sum_{j \neq j'} f_i\left(\sum_{t=1}^{T} |\omega_{j,j'}^t|\right), \quad (1)$$

where $\omega_{j,j'}^t$ is the (j, j')th element of $\mathbf{\Omega}^t$ and f_i is a nonconvex penalty function. We consider the following three nonconvex penalty functions:

1. $f_1(x) = |x|^{1-\nu}$ for $0 < \nu < 1$
2. $f_2(x) = (\log(|x|) - \log\varepsilon + 1) I(|x| > \varepsilon) + \frac{|x|}{\varepsilon} I(|x| \leq \varepsilon)$
3. $f_3(x) = (-|x|^{1-\nu} + \nu\varepsilon^{1-\nu})\, I(|x| > \varepsilon) - (1 - \nu)|x|\varepsilon^{-\nu} I(|x| \leq \varepsilon)$ for $\nu > 1$.

Here, ε is a small positive constant. The f_1 penalty function has been used in group bridge estimation in a regression context (Huang et al., 2009). The f_2 penalty function is a truncated

log function, and f_3 is a truncated inverse polynomial in which truncation occurs when $|x| \leq \varepsilon$ to avoid infinity. We remark that the log penalty function has been used by others including Sweetkind-Singer (2004) and Mazumder et al. (2011) in different contexts.

The joint estimation of multiple GGMs by using a nonconvex penalty function is not new, as Guo et al. (2011) used the penalty function of $\sqrt{x}$ for the purpose. They showed that the use of $\sqrt{x}$ function is equivalent to the hierarchical penalization of common and condition-specific regularization. In their work, the common structure was introduced to represent an edge set that is the union of all individual edge sets, and it was denoted as a $p \times p$ matrix Θ. They specifically set $\theta_{j,j'}$ to be proportional to $\sqrt{\sum_{t=1}^{T} |\omega_{j,j'}^t|}$, where $\theta_{j,j'}$ is the (j, j')th element of Θ. We found that the joint sparsity regularization can be achieved similarly with functions other than the square root function, which is shown in the following Proposition 1.

Proposition 1—If $\{\hat{\mathbf{\Omega}}^t\}_{t=1}^T$ is a local minimizer of $PL_i(\{\mathbf{\Omega}^t\}_{t=1}^T)$, there exists $\hat{\Theta}$ such that $(\{\hat{\mathbf{\Omega}}^t\}_{t=1}^T, \hat{\Theta})$ is a local minimizer of

$$\widetilde{PL}_i(\{\mathbf{\Omega}^t\}_{t=1}^T, \Theta) \; subject\, to\, \theta_{j,j'} \geq 0, \; for\, 1 \leq j, j' \leq p, \quad (2)$$

where $\widetilde{PL}_i(\{\mathbf{\Omega}^t\}_{t=1}^T, \Theta)$ is defined as follows:

$$\widetilde{PL}_i(\{\mathbf{\Omega}^t\}_{t=1}^T, \Theta) = L(\{\mathbf{\Omega}^t\}_{t=1}^T) + \tau \left(\sum_{j \neq j'} g_i(\theta_{j,j'}) \sum_{t=1}^{T} |\omega_{j,j'}^r| + \sum_{j \neq j'} \theta_{j,j'} \right),$$

1. $g_1(x) = x^{\frac{-\nu}{1-\nu}}$; *and* $\lambda = \tau \nu^{-\nu}(1-\nu)^{\nu-1}$
2. $g_2(x) = \begin{cases} \frac{\exp(1-x)}{\varepsilon} & x > 1 \\ \frac{1}{\varepsilon} & 0 \leq x \leq 1 \end{cases}$; *and* $\lambda = \tau$
3. $g_3(x) = \begin{cases} \frac{\nu-1}{\nu}\left(\nu\varepsilon^{1-\nu} - x\right)^{\frac{\nu}{\nu-1}} & x > (\nu-1)\varepsilon^{1-\nu} \\ \frac{\nu-1}{(\nu)\varepsilon^{\nu}} & 0 \leq x \leq (\nu-1)\varepsilon^{1-\nu} \end{cases}$; *and* $\lambda = \frac{\tau}{\nu}$.

Here, $\tau > 0$ is a tuning parameter for $\widetilde{PL}$.

Conversely, if $(\{\hat{\mathbf{\Omega}}^t\}_{t=1}^T, \hat{\Theta})$ is a local minimizer of (2), $\{\hat{\mathbf{\Omega}}^t\}_{t=1}^T$ is a local minimizer of $PL_i(\{\mathbf{\Omega}^t\}_{t=1}^T)$.

The proof is given in Appendix. In the proposition, $\hat{\theta}_{j,j'} (\geq 0)$ is interpreted as a common structure, and defined by the minimizer of the objective function (2). Hence, each nonconvex function yields a different form of $\hat{\theta}_{j,j'}$. In fact, $\hat{\theta}_{j,j'}$ is proportional to $\left(\sum_{t=1}^{T} |\hat{\omega}_{j,j'}^t|\right)^{1-\nu}$, $\log\left(\sum_{t=1}^{T} |\hat{\omega}_{j,j'}^t|\right)$, and $\left(\frac{\nu}{\varepsilon^{\nu-1}} - \frac{1}{\left(\sum_{t=1}^{T} |\hat{\omega}_{j,j'}^t|\right)^{\nu-1}} \right)$, for functions f_1, f_2, and f_3,

respectively, when $\sum_{t=1}^{T}|\hat{\omega}_{j,j'}^{t}| > \varepsilon$. These are increasing functions with respect to $\sum_{t=1}^{T}|\hat{\omega}_{j,j'}^{t}|$ with varying curvatures.

From the proposition, one can find that our proposed approach regularizes the common and condition-specific structures hierarchically with two characteristics. First, the common edge selection is guided by the choice of nonconvex function. As discussed in the previous paragraph, $\hat{\theta_{j,j'}}$ is differently defined depending on the type of the nonconvex function, where f_3 enforces the joint sparsity most strongly, followed by f_2 and f_1. Second, condition-specific edge selection is guided by the weight function $g_i(\hat{\theta_{j,j'}})$. Since g_i is a monotone decreasing function with respect to $\hat{\theta_{j,j'}}$, an edge with a small common structure is penalized more heavily than an edge with a large common structure. Thus, the proposition shows that our approach achieves the joint regularization via the use of a nonconvex penalty function and has flexibility of controlling the balance between common and condition-specific edge selection via the choice of the nonconvex function.

2.3 Algorithm

In this subsection, we describe an algorithm that uses the local linear approximation to find a solution of (1). It has been shown that the minimizer of (1) with the $\sqrt{x}$ penalty function can be found by the local linear approximation (Zou and Li, 2008), which was also used in Guo et al. (2011). The penalty function can be approximated as $f_i(\sum_{j\neq j'}|\omega_{j,j'}^{t}|) \approx f_i(\sum_{j\neq j'}|\hat{\omega}_{j,j'}^{t}|)+\sum_{j\neq j'}f_i'(\sum_{j\neq j'}|\hat{\omega}_{j,j'}^{t}|)(|\omega_{j,j'}^{t}|-|\hat{\omega}_{j,j'}^{t}|)$. By extracting terms related to the $\omega_{j,j'}^{t}$, we get the computational algorithm as follows:

1. Initialize $\hat{\mathbf{\Omega}^t}$ for all $1 \le t \le T$.
2. Update $\hat{\mathbf{\Omega}^t}$ for all $1 \le t \le T$ by solving

$$\operatorname{argmin}_{\mathbf{\Omega}^t}\frac{n_t}{2}(\operatorname{tr}(\mathbf{S}^t\mathbf{\Omega}^t)-\log\{\det(\mathbf{\Omega}^t)\})+\tilde{\lambda}\sum_{j\neq j'}\frac{|\omega_{j,j'}^{t}|}{(\sum_{t=1}^{T}|\hat{\omega}_{j,j'}^{t}|)^{\nu}},$$

   using a **glasso**, where $\hat{\omega}_{j,j'}^{t}$, is the estimate from the previous iteration and $\nu > 0$, $\tilde{\lambda} = \lambda$ for f_2 and $\tilde{\lambda} = |1 - \nu|$ for f_1 and f_3.
3. Repeat step 2 until convergence is achieved.

In the algorithm, ν is the same as the one used in (1) for penalty functions f_1 and f_3, and ν is set to be 1 for penalty function f_2. Specifically, the penalty function f_1, also known as a bridge penalty, considers $0 < \nu < 1$; the penalty function f_2 corresponds to $\nu = 1$; and the penalty function f_3 corresponds $\nu > 1$. Thus, these three penalty functions comprise the continuum of the iteratively reweighted graphical lasso with $\nu > 0$.

Our algorithm only guarantees to yield a local solution, and thus the choice of the initial value is important to get an appropriate estimate. When $n \ge p$, one can use $(\mathbf{S}^t + \delta I)^{-1}$ as an initial estimate, where $\delta \ge 0$ is chosen to be a small constant to avoid singularity. However,

when $n < p$, this form of the initial estimate does not perform well. In this case, one can use the solution of separate GGM approaches with an ℓ_1 regularization, because in high-dimensional estimation, a reasonable estimate can be obtained by using a sparsity regularization.

The tuning parameter, λ, can be selected by minimizing the approximation of Bayesian information criterion (aBIC) as in Yuan and Lin (2007). The aBIC is defined by

$$\mathrm{aBIC}(\lambda)=\sum_{t=1}^{T}\left\{-\log\det(\hat{\mathbf{\Omega}}^t(\lambda))+\mathrm{tr}(\mathbf{S}^t\mathbf{\Omega}^t(\lambda))+\frac{\log(n_t)}{n_t}df_t\right\},$$

where $\{\hat{\mathbf{\Omega}}^t(\lambda)\}_{t=1}^{T}$ are the minimizer of (1) with a tuning parameter λ, and $\mathrm{card}\{(j,j'):j\le j',\hat{\omega}^t_{j,j'}(\lambda)\neq 0\}$ with card representing the cardinality of a finite set. We remark that df_t is a heuristic degrees of freedom and hence the proposed aBIC is an approximation of the original BIC criterion.

2.4 Consistency and Sparsistency

In this subsection, we show that the estimate from the formulation (1) above achieves consistency and sparsistency. The sparsistency, however, is limited in that it only finds a group structure, rather than individual structures.

Denote $\mathbf{E}_t=\{(j,j'):j\neq j',\mathbf{\Omega}^t_{0,j,j'}\neq 0\}$ to be the set of indices of all nonzero off-diagonal elements in $\mathbf{\Omega}^t_0$, $\mathbf{E}=\mathbf{E}_1\cup\ldots\cup\mathbf{E}_T$, $q_t=|\mathbf{E}_t|$ and $q_t=|\mathbf{E}_t|$, where $\mathbf{\Omega}^t_0$ is a true precision matrix.

We assume the following regularity conditions as in Guo et al. (2011):

1. There exist constants ξ_1 and ξ_2 such that for all $p\ge 1$ and $1\le t\le T$,

$$0<\xi_1<\phi_{min}(\mathbf{\Omega}^t_0)\le\phi_{max}(\mathbf{\Omega}^t_0)<\xi_2<\infty,$$

   where $\varphi_{min}(\mathbf{A})$ and $\varphi_{max}(\mathbf{A})$ represent the minimal and maximal eigenvalues of a matrix $\mathbf{A}$.

2. There exists a constant $\xi_3>0$ such that

$$\min_{1\le t\le T}\min_{(j,j')\in\mathbf{E}_t}|\mathbf{\Omega}^t_{0,j,j'}|\ge\xi_3.$$

Theorem 1—Under regularity conditions 1 and 2, when $\frac{(p+q)(\log p)}{n}=o(1)$ and $\Lambda_1\sqrt{\frac{\log p}{n}}\le\lambda\le\Lambda_2\sqrt{(1+p/q)\frac{\log p}{n}}$, there exists a local minimizer of the objective function (1), such that $\sum_{t=1}^{T}\|\hat{\mathbf{\Omega}}^t-\mathbf{\Omega}^t_0\|_F=O_p\left(\sqrt{\frac{(p+q)\log p}{n}}\right)$.

Here $\|\mathbf{A}\|_F$ represents the Frobenius norm of a matrix $\mathbf{A}$.

Theorem 2—Under all of the assumptions in Theorem 1, and the assumptions of $\sum_{t=1}^{T}\|\hat{\mathbf{\Omega}}^t-\mathbf{\Omega}_0^t\|^2=O_p(\eta_n)$, where $\eta_n \to 0$ and $\sqrt{\frac{\log p}{n}}+\sqrt{\eta_n}=O(\lambda)$, the local minimizer of the objective function (1) satisfies that (a) $\hat{\omega}^t_{j,j'}=0$ for all $1 \le t \le T$ for any $(j, j') \in \mathbf{E}^c$ and (b) $\hat{\omega}^t_{j,j'} \neq 0$ for some $1 \le t \le T$ for any $(j, j') \in \mathbf{E}$ with probability tending to 1. Here $\|\mathbf{A}\|$ represents the operator norm of a matrix $\mathbf{A}$.

From Theorem 2, we find that the sparsistency holds only at the group level, meaning that it is able to declare edges that do not appear in any condition, but that the sparsistency is not guaranteed to hold at the condition-specific level. In order to achieve sparsistency at the condition-specific level, a separate GGM estimation with a nonconvex penalty function for each condition should be used. We further find that consistency and sparsistency can be achieved simultaneously in very limited scenarios, which was discussed in Guo et al. (2011). It is because λ needs to be bounded both below and above for the consistency, but λ needs to be bounded below only for the sparsistency. These bounds can be matched, when $\sqrt{\frac{\log p}{n}}+\sqrt{\eta_n}=O\left(\sqrt{\frac{(1+p/q)\log p}{n}}\right)$. Due to the norm relationship of a $p \times p$ matrix $\mathbf{A}$ ($\|\mathbf{A}\|_F/\sqrt{p} \le \|\mathbf{A}\| \le \|\mathbf{A}\|_F$), we have that $\eta_n = O((p+q)\log p/n)$ in the worst case scenario and $\eta_n = O((1+q/p)\log p/n)$ in the best case scenario. Hence, q should be $O(1)$ in the worst case and q can be $O(p)$ in the best case. If Theorem 1 is improved to show the operator norm consistency, the inconvenient consistency condition of Theorem 2 can be removed.

3 Simulation Study

3.1 Performance as function of tuning parameter

An incidence matrix with a scale-free network structure is generated using the Barabasi-Albert algorithm (Barabasi and Albert, 1999). We start from six edges, and add one edge at each step. We first generate shared edges and then, for each condition, we add randomly selected $0.1M$ edges as condition-specific edges, where M is the total number of edges in the shared structure. The total number of nodes in a graph, p, is set to be 500, and we consider 5 conditions ($T = 5$). Further, we set the sample size for each condition (n_t) to be 150.

We generate precision matrices by setting the nonzero elements to values that are sampled from Unif($[-1, -0.5] \cup [0.5, 1]$). We then set the diagonal elements to ($1.5\ \Sigma_{j \neq i}\ |\omega_{i,j}|$). In this way, the resulting $\mathbf{\Omega}^t$ may not be positive definite, and thus we repeat this precision matrix generation process until $\mathbf{\Omega}^t$ becomes a positive definite matrix. Due to the scale-free structure, some diagonal elements are much larger than the others. We thus adjust the precision matrices as in Danaher et al. (2012) by replacing the nonzero elements of $\mathbf{\Omega}^t$ with those of $\mathbf{\Omega}^{\tilde{t}}$, where $\mathbf{\Omega}^{\tilde{t}} = (\mathbf{D}^{t^{-1/2}}\mathbf{\Sigma}^{\tilde{t}}\mathbf{D}^{t^{-1/2}})^{-1}$; $\mathbf{\Sigma}^{\tilde{}} = 0.6\mathbf{\Omega}^{t^{-1}} + 0.4\mathbf{D}^t$; and $\mathbf{D}^t$ is the diagonal matrix whose (i, i)th elements is the (i, i)th element of $\mathbf{\Omega}^{t^{-1}}$. Finally, the p dimensional random vectors are simulated from $N(0, \mathbf{\Omega}^{t^{-1}})$. All of the simulation results are based on 100 replicates.

We use $\nu = 0.5$ for f_1; $\nu = 1$ for f_2; and $\nu = 2$ for f_3 penalty function. We compare the performance of our proposed nonconvex penalty approaches to that of a single global GGM and that of multiple GGMs separately, as well as that of Danaher et al. (2012)'s GGL approach. For the GGL approach, we reparametrize the tuning parameters as in the simulation study of Danaher et al. (2012), where $\omega_1 = \lambda_1 + \frac{1}{\sqrt{2}}\lambda_2$ and $\omega_2 = \frac{1}{\sqrt{2}}\lambda_2/(\lambda_1 + \frac{1}{\sqrt{2}}\lambda_2)$. Throughout the simulation study, we set $\varepsilon = 1.0 \times 10^{-6}$, and thr = 0.1 for **glasso** algorithm in our approach. The initial estimate of Ω^t was obtained by applying **glasso** algorithm with $\lambda = \frac{1}{Tp^2}\sum_t\sum_i\sum_j|\mathbf{S}^t_{ij}|$, separately.

When we estimate a single global GGM and multiple separate GGMs, we consider the following objective functions PL_G and PL_S, respectively:

$$PL_G(\mathbf{\Omega}) = \sum_{t=1}^{T}\left(\frac{n_t}{2}\mathrm{tr}(\mathbf{S}^t\mathbf{\Omega}) - \log\det(\mathbf{\Omega})\right) + \lambda\sum_{j\neq j'} f_2\left(|\omega_{j,j'}|\right)$$
$$PL_S(\mathbf{\Omega}^t) = \frac{n_t}{2}(\mathrm{tr}(\mathbf{S}^t\mathbf{\Omega}^t) - \log\det(\mathbf{\Omega}^t)) + \lambda\sum_{j\neq j'} f_2\left(|\omega^t_{j,j'}|\right),$$

for $t = 1, \ldots, T$. The nonconvex penalty function is used for promoting model selection consistency.

A part of criteria to compare the methods are as follows:

C1 # of falsely declared edges at λ:

$$\sum_{t=1}^{T}\mathrm{card}\{(i,j): i>j, \omega^t_{i,j}=0 \text{ and } \hat{\omega}^t(\lambda)_{i,j}\neq 0\},$$

where, *card*(*A*) for a set *A* represents the cardinality of the set *A*.

C2 # of correctly declared edges at λ:

$$\sum_{t=1}^{T}\mathrm{card}\{(i,j): i>j, \omega^t_{i,j}\neq 0 \text{ and } \hat{\omega}^t(\lambda)_{i,j}\neq 0\}.$$

C3 # of falsely declared edges in a combined graph at λ:

$$\mathrm{card}\{(i,j): i>j; \omega^t_{i,j}=0 \text{ for all } t=1,\ldots,T; \text{and } \hat{\omega}^t(\lambda)_{i,j}\neq 0 \text{ for some } t, 1\leq t\leq T\}.$$

C4 # of correctly declared edges in a combined graph at λ:

$$\mathrm{card}\{(i,j): i>j; \omega^t_{i,j}\neq 0 \text{ for some } t, 1\leq t\leq T; \text{and } \hat{\omega}^t(\lambda)_{i,j}\neq 0 \text{ for some } 1\leq t\leq T\}.$$

C5 Relative squared distance (RSD) at λ:

$$\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{\Omega}^t - \hat{\mathbf{\Omega}}^t(\lambda)\|_F^2 / \|\mathbf{\Omega}^t\|_F^2.$$

The simulation study shows that our proposed method performs similarly to the GGL in terms of edge selection accuracy (Figure 1 (a)). Our approach and GGL perform better than the approach of finding a single global GGM or separate multiple GGMs. This trend stays the same, when all methods are compared in terms of finding non-edges across all conditions (common zeros) (Figure 1 (b)). When the methods are compared in terms of timing (Figure 1(c)), our approach starts to show benefit when the estimated graphs become dense. In this simulation, the total number of possible edges was 623,750, and the true number of edges was 8090. When the estimated graph size is greater than 5590, which is a reasonable range of the estimated graph size, our approach is faster than GGL. Finally, when the relative squared distances are compared (Figure 1 (d)), our approach shows the best performance. As shown in subsection 2.4, our approach has consistency in both estimation and model selection, which is reflected in this result. The simulation study suggests that our proposed approach performs very well in terms of model selection, timing, and estimation.

Additionally, the simulation study shows that the use of the extended class of nonconvex function gives the flexibility of controlling the balance of common and condition specific edges. In current scenario, the $f_3,\ \nu = 2$ penalty function performs the best by enforcing common structures.

3.2 Performance as a function of *n*, *p* and *T*

In this subsection, we compare three different nonconvex functions ($f_1,\ \nu = 0.5$; $f_2,\ \nu = 1$; $f_3,\ \nu = 2$) under various settings of n, p and T. The datasets are simulated as in the previous subsection 3.1, and the results are based on 100 replicates.

The tuning parameter λ is chosen by using the aBIC criterion. Table 3 shows that when the number of condition is large ($T = 5$), the $f_3,\ \nu = 2$ penalty function performs the best. When $T = 2$, the $f_2,\ \nu = 1$ penalty performs the best in any combination of n and p. Across all simulations, the $f_2,\ \nu = 1$ penalty function performs better than the $f_1,\ \nu = 0.5$ penalty function.

3.3 Discussion on penalty function selection

One can select an appropriate type of penalty function by adopting a tuning criterion. This requires two-way tuning for the regularization parameter λ and the type of nonconvex function (equivalently, the choice of ν), which has been adopted in the adaptive lasso (Zou, 2006). However, it is quite challenging to find an optimal tuning parameter in high-dimensional problems with low sample sizes, which often occur in many genomic data analyses. Based on our simulation studies, we find that imposing a stronger level of sparsity generally improves the performance under this setting with more benefit of using f_2 over f_1 than that of using f_3 over f_2. We thus recommend to use the f_2 penalty function, when it is challenging to use the two-way tuning in case of high-dimensional and low sample size problems.

4 Real data analysis

We apply the proposed joint estimation of multiple GGMs with a nonconvex penalty function to a gene expression dataset. It has been suggested that gene regulations differ among conditions due to differential use of the regulatory elements of genes, and understanding this differential regulation is an interesting scientific problem. One such study was done by Dimas et al. (2009), in which gene expressions were measured from three cell types, primary fibroblasts, Epstein-Barr virus (EBV)-immortalized B-cells, and T-cells of 85 individuals participating in the GenCord project. We remark that these three cell types were extracted from umbilical cords, where the fibroblasts were obtained by culturing finely cut cord tissue, and B-cells from cord blood with EBV-immortalization; and T-cells from cord blood with PHA stimulation (Dimas et al., 2009).

Since each individual contributed the three cell types, and thus the three sets of datasets are not independent to each other. This aspect is not properly addressed in the current analysis, which is our future work. However, the possible similarity of graphs due to the dependence can be reflected with the joint sparsity regularization. The dataset contains mRNA levels that are quantified with 48,804 probes with the Illumina WG-6 v3 expression array. We convert the probe level data to the gene level data by taking the average of the probes mapped to a gene. We then take the log transformation to make the data more normally distributed, and then use a total of 17,945 autosomal RefSeq genes' expression for inferring network of genes.

Due to the limitation of the sample size ($n_1 = n_2 = n_3 = 85$), we partition genes into smaller groups by using pathway information. We extract the information of 528 pathways from the KEGG database. Among these, we separately analyze 277 pathways that contain at least 3 and at most 29 genes in our dataset. We apply our approach with the f_2, $\nu = 1$ penalty function and select $2\lambda/\tilde{n}_t$ from a total of 300 possible values equally spaced in $\log_{10}$ scale between 10^{-5} and 1. When the ratio of the largest to the smallest eigen value of $\mathbf{S}^t$ is smaller than or equal to 1000, we use $(\mathbf{S}^t)^{-1}$ as the initial estimate. Otherwise, we use $(\mathbf{S}^t + \delta I_p)^{-1}$ as the initial estimate, where $\delta = \max(\frac{1}{1000}\|\mathbf{S}^t\|^2, \frac{1}{10p^2}\sum_{i,j}|\mathbf{S}^t_{i,j}|)$. This choice was made in order to reflect the differing size and signal strength of the networks.

In our analysis, we extracted edge information from the database, and treated this as a true network structure. We considered this true structure as a collection of condition-specific edges without the condition information, because the database is curated by collecting the parts of gene regulations over various conditions. If an approach can preserve the tissue-specificity, it would capture the true edges more accurately than the approaches that capture only the common ones.

Rather than handpicking one of the 277 pathways for the presentation, we summarize our results using the true network structure. We compute the area under the curve (AUC) of each receiver operating curve (ROC) for each pathway. The AUC is used in order to avoid tuning parameter selection. Therefore, we evaluate three AUC values for each pathway corresponding to the three cell types. We then sort the pathways based on their AUC values, and list the pathways that have AUC values larger than 0.8 in Table 1. We compared our

approach to the global and the separate approaches. The detailed comparison of the JGGM and the separate approaches is in Supplementary Materials. The pathways that have AUC values of 1 tend to have small graphs. However, even for these small graphs, the other methods do not always have the AUC values of 1, suggesting that the AUCs of 1 are not just automatic results from the nature of tiny graphs.

Our results show that the identified pathways from these three distinct cell types are very similar to each other. Notably, the identified pathways are mostly relevant to immune responses, where the role of PPARgamma in immune response regulation via dendritic cell control and lipid metabolism was demonstrated in the literature (Szatmari et al., 2007; Wieser et al., 2008). Considering that the fibroblast cells are quite different from the other cells, this result might need some more validation. The possible explanation could be the fact that these cells were all taken from the umbilical cords. Also, a similar conclusion was reached by Flutre et al. (2013), when they analyzed the same dataset for finding expression quantitative loci (eQTL). They found that most of genetical genomic controls are not cell-type specific, suggesting that these three cell types might function similarly in the umbilical cords.

There are few pathways that were selected in a specific cell type; *retinol metabolism pathway* appears in the top list only in EBV-transformed B-cells, where it is known that retinol is essential for growth of activated human B-cells (Buck et al., 1990); *Alzheimer's disease pathway* appears only in the fibroblasts cells; and *Cycling of Ran in nucleocytoplasmic pathway* appears only in the activated T-cells, where *Ras* gene is an oncogene that is related to abnormal cell proliferation (Xia et al., 2008).

Figure 2 shows the estimated GGMs with genes consisting of an *embryonic stem cell pathway*. We set $2\lambda/\tilde{n}_t$ to be 2.4477×10^{-4} for the separate approach and 1.8738×10^{-5} for our joint analysis by using the aBIC criterion. We can confirm that a joint approach is able to produce different graphs for the three different cells. The graphs of the EBV-transformed B-cells and the T-cells are the same, but the graph of fibroblasts contains few more edges, which is consistent with the fact that the B-cells and T-cells are more close to each other than to the fibroblasts cells.

5 Conclusion

With the advancement of biological network annotations and network theory developments, researchers now attempt to use network information to decipher biological processes. Although useful, the network annotation itself is not complete, and much can be learned through inferring a network structure from multiple sources of data. In this paper, we propose to infer GGMs from gene expression data with a nonconvex penalty function for joint sparsity in order to effectively estimate multiple network structures.

We have shown the consistency and sparsistency of the proposed approach and have found that the sparsistency holds only for group-level selection. Nonetheless, this limitation is not critical because the sample size is often not large enough to invoke such theoretical results in a real application. Our simulation study showed that the proposed nonconvex function performs well by capitalizing the shared sparsity across different conditions.

We have applied the proposed approach to analyze a gene expression dataset from the GenCord project. We utilized the KEGG pathway information to facilitate our analysis and interpretation. We found that the pathways related to immune responses were pronounced in our study of umbilical cord tissues. We remark that the conclusion is based only on AUC values that do not account for associated random errors, and finding a measure that leads to a proper graph enrichment study will be our future work.

Although we suggested that the proposed nonconvex objective function can be optimized via an iteratively reweighted adaptive lasso algorithm, we did not prove that this solution is a global one. This is a general problem in most regularization approaches that use nonconvex penalty functions, and we leave it as an important problem for future research. Our future work further includes characterizing uncertainties in inferred network structures. We could use the bootstrap approach, but we would like to find the approximate variance of our proposed estimator for computational efficiency.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

H. Chun's research was supported by NSF grant DMS-1107025 and H. Zhao's research was supported in part by NSF grant DMS-1106738 and NIH grants R01-GM59507 and P01-CA154295.

References

Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999; 286:509–512. [PubMed: 10521342]

Buck J, Ritter G, Dannecker L, Katta V, Cohen SL, Chait BT, Hammerling U. Retinol is essential for growth of activated human B cells. J Exp Med. 1990; 171:1613–1624. [PubMed: 2332732]

Cai T, Liu W, Luo X. A constrained l_1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association. 2011; 106:594–697.

Chiquet J, Grandvalet Y, Ambroise C. Inferring Multiple Graphical Structures. 2009 unpublished.

Danaher P, Wang P, Witten D. The joint graphical lasso for inverse covariance estimation across multiple classes. 2012 arXiv:1111.0324.

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Arcelus MG, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermizakis ET, Antonarakis SE. Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. Science. 2009; 325:1246–1250. [PubMed: 19644074]

Flutre T, Wen X, Pritchard J, Stephens M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. PLOS Genetics. 2013

Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. Biometrika. 2011; 98:1–15. [PubMed: 23049124]

Huang J, Ma S, Xie H, Zhang C-H. A group bridge approach for variable selection. Biometrika. 2009; 96:339–335. [PubMed: 20037673]

Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. Annals of Statistics. 2009; 37:4254–4278. [PubMed: 21132082]

Lauritzen, SL. Graphical Models. Oxford: Clarendon Press; 1996.

Mazumder R, Friedman JH, Hastie T. SparseNet: Coordinate Descent With Nonconvex Penalites. Journal of the American Statistical Association. 2011; 106:1125–1138. [PubMed: 25580042]

Meinshausen N, Buhlmann P. High-dimensional graphs with the lasso. Annals of Statistics. 2006; 34:1436–1462.

Peng J, Wang P, Zhou N, Zhu J. Partial Correlation Estimation by Joint Sparse Regression Model. Journal of American Statistical Association. 2009; 104:735–746.

Rothman A, Bickel P, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics. 2008; 2:494–515.

Soranzo N, Bianconi G, Altafini C. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. Bioinformatics. 2007; 23:1640–1647. [PubMed: 17485431]

Sweetkind-Singer, JA. Log-Penalized Linear Regression. Standford University; 2004.

Szatmari I, Tsik D, Agostini M, Nagy T, Gurnell M, Barta E, Chatterjee K, Nagy L. PPARgamma regulates the function of human dendritic cells primarily by altering lipid metabolism. Blood. 2007:3271–3280. [PubMed: 17664351]

Werhli AV, Grzegorczyk M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and Bayesian networks. Bioinformatics. 2006; 22:2523–2531. [PubMed: 16844710]

Wieser F, Waite L, Depoix C, Taylor R. PPAR Action in Human Placental Development and Pregnancy and Its Complications. PPAR Research. 2008

Xia F, Canovas P, Guadagno T, Altieri D. A survivin-ran complex regulates spindle formation in tumor cells. Mol Cell Biol. 2008; 28:5299–5311. [PubMed: 18591255]

Yuan M. High dimensional inverse covariance matrix estimation via linear programming. Journal of Machine Learning Research. 2010; 11:2261–2286.

Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika. 2007; 94:19–35.

Zou H. The adaptive lasso and its oracle properties. Journal of American Statistical Association. 2006; 101:1418–1429.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics. 2008; 36:1108–1126.

Appendix

Proof of Proposition 1

Proof 1

The proof for the f_1 penalty function is provided in Huang et al. (2009).

We start to prove the proposition for the case of the f_2 penalty function. When $\sum_{t=1}^{T}|\omega_{j,j'}^{t}|>\varepsilon$, one can find that the solution of the derivative equation, $\frac{\partial}{\partial\theta_{j,j'}}\widetilde{PL}(\{\mathbf{\Omega}^t\}_{t=1}^{T},\Theta)=0$, is $\theta_{j,j'}=1-\log(\varepsilon)+\log(\sum_{t=1}^{T}|\omega_{j,j'}^{t}|)$. Hence, $\sum_{t=1}^{T}|\omega_{j,j'}^{t}|>\varepsilon$ is equivalent to $\hat{\theta_{j,j'}}>1$. Pluggin this into $\widetilde{PL}(\{\mathbf{\Omega}^t\}_{t=1}^{T},\Theta)$ yields a profiled penalized likelihood of $\widetilde{pPL}(\{\mathbf{\Omega}^t\}_{t=1}^{T})=L(\{\mathbf{\Omega}^t\}_{t=1}^{T})+\tau\sum_{j\neq j'}(\log(\sum_{t=1}^{T}|\omega_{j,j'}^{t}|)-\log\varepsilon+1)$. By taking $\lambda=\tau$, one can find that $\widetilde{pPL}(\{\mathbf{\Omega}^t\}_{t=1}^{T})=PL(\{\mathbf{\Omega}^t\}_{t=1}^{T})$.

When $\sum_{t=1}^{T}|\omega_{j,j'}^{t}|\leq\varepsilon$, the penalty form of $PL(\{\mathbf{\Omega}^t\}_{t=1}^{T})$ becomes $\frac{1}{\varepsilon}\sum_{t=1}^{T}|\omega_{j,j'}^{t}|$. This is equivalent to not assuming a common structure, which can be achieved by setting $g(\hat{\theta_{j,j'}})$ to be a constant function $\frac{1}{\varepsilon}$ when $0\leq\hat{\theta_{j,j'}}\leq 1$.

We then prove the proposition for the case of the f_3 penalty function by using the same principle. We find that the solution $\theta_{j,j'}=\nu\varepsilon^{1-\nu}-\left(\sum_{t=1}^{T}|\omega_{j,j'}\right)^{1-\nu}$. Hence, $\sum_{t=1}^{T}|\omega_{j,j'}^{t}|>\varepsilon$ is equivalent to $\hat{\theta_{j,j'}}>(\nu-1)\varepsilon^{1-\nu}$. This yields a profiled likelihood of $\widetilde{pPL}(\{\mathbf{\Omega}^t\}_{t=1}^{T},)=L(\{\mathbf{\Omega}^t\}_{t=1}^{T})+\frac{\tau}{\nu}\sum_{j\neq j'}(\nu\varepsilon^{1-\nu}-\sum_{t=1}^{T}|\omega_{j,j'}|^{1-\nu})=PL(\{\mathbf{\Omega}^t\}_{t=1}^{T})$ by taking $\lambda=\frac{\tau}{\nu}$.

Lemma 1

If either x or y is greater than $\tau(>0)$, then $|x^{\alpha}-y^{\alpha}|\tau^{1-\alpha}\leq|x-y|$, for $0<\alpha<1$.

Proof 2

Without loss of generality, we can assume that $x\geq y$.

When $x>\tau>y$,

$$(x^{\alpha}-y^{\alpha})\tau^{1-\alpha}\leq x-y^{\alpha}\tau^{1-\alpha}\leq x-y.$$

When $x>y>\tau$,

$$x^{\alpha}-y^{\alpha}\leq\frac{1}{y^{1-\alpha}}(x,y)\leq\frac{1}{\tau^{1-\alpha}}(x-y).$$

Proof of Theorem 1

Proof 3

Theorem 1 can be proved with a slight extension to the proof of Guo et al. (2011), which is similar to the proof of Theorem 1 of Rothman et al. (2008).

Denote the objective function 1 as $Q(\mathbf{\Omega})$, where $\mathbf{\Omega}=\{\mathbf{\Omega}^t\}_{t=1}^{T}$ and we write the true precision matrices as $\mathbf{\Omega}_0=\{\mathbf{\Omega}_0^t\}$. We would like to show $Q(\Omega)$ has the local minimum near $\mathbf{\Omega}_0$.

Specifically, we would like to show that $P(Q(\tilde{\Delta})=Q(\mathbf{\Omega}_0+\Delta)-Q(\mathbf{\Omega}_0)>0)$ converges to 1, when $\Delta\in\partial\mathcal{A}$, where $\partial\mathcal{A}=\{\Delta:\sum_{t=1}^{T}\|\Delta^t\|_F=Mr_n\}$, and $\Delta^t=\hat{\mathbf{\Omega}}^t-\mathbf{\Omega}_0^t$, and M is a positive constant and $r_n=\sqrt{\frac{(p+q)\log p}{n}}$.

We will use the following notation: for a matrix $\mathbf{M}=[m_{j,j'}]_{p\times p}$, $|\mathbf{M}|_1=\Sigma_{j,j'}|m_{j,j'}|$, $\mathbf{M}^{+}$ is a diagonal matrix with the same diagonal as $\mathbf{M}$, $\mathbf{M}^{-}=\mathbf{M}-\mathbf{M}^{+}$, and M_S is M with all elements outside an index set $\mathbf{S}$ replaced by zeros. Also, vec($\mathbf{M}$) for the vectorized form of $\mathbf{M}$, and $\otimes$ for the Kronecker product of two matrices.

As in Guo et al. (2011), Q is the sum of the following components:

$$I_1 = \sum_{t=1}^{T} trace((\mathbf{S}^t - \sum_0^t)\Delta^k)$$
$$I_2 = \sum_{t=1}^{T} \tilde{\Delta}^{t'} \int_0^1 (1-v)(\mathbf{\Omega}_0^t + v\Delta^t)^{-1} \otimes (\mathbf{\Omega}_0^t + v\Delta^t)^{-1} dv \tilde{\Delta}^t$$
$$I_3 = \lambda \sum_{(j,j')\in \mathbf{E}^c} f_i(\sum_{t=1}^{T} (|\delta_{j,j'}^t|))$$
$$I_4 = \lambda \sum_{(j,j')\in \mathbf{E}} (f_i(\sum_{t=1}^{T} |\omega_{j,j'}^t|) - f_i(\sum_{t=1}^{T} |\omega_{0_{j,j'}}^t|))$$

The bound for the likelihood part can be found in Guo et al. (2011), where

$$|I_1| \leq C_1 \sqrt{\frac{\log p}{n}} \sum_{t=1}^{T} |\Delta^{t-}|_1 + C_2 \sqrt{\frac{p \log p}{n}} \sum_{t=1}^{R} \|\Delta^{t+}\|_F,$$
$$I_2 \geq \frac{1}{4\xi_2^2} \sum_{t=1}^{T} \|\Delta^t\|_F^2,$$

for some constants C_1 and C_2 with probability tending to 1.

When $(p + q)(\log p)/n$ is small,

$$I_3 \geq \lambda \sum_{t=1}^{T} |\Delta_{\mathbf{E}^c}^{t-}|_1,$$

due to the concavity of the penalty functions.

Also,

$$I_4 \leq \lambda \sum_{j\neq j:(j,j')\in \mathbf{E}} \left| f_i(\sum_{t=1}^{T} |\omega_{j,j'}^t|) - f_i(\sum_{t=1}^{T} \omega_{0_{j,j'}}^t) \right|.$$

For the f_1 function, by using Lemma 1,

$$I_4 \leq \frac{\lambda}{\xi_3^\nu} \sum_{j\neq j':(j,j')\in \mathbf{E}} |\sum_{t=1}^{T} |\omega_{j,j'}^t| - \sum_{t=1}^{T} |\omega_{0_{j,j'}}^t||$$
$$\leq \frac{\lambda}{\xi_3^\nu} \sum_{t=1}^{T} \sum_{j\neq j':(j,j')\in \mathbf{E}} |\omega_{j,j'}^t - \omega_{0_{j,j'}}^t|$$
$$\leq \frac{\lambda}{\xi_3^\nu} \sqrt{q} \sum_{t=1}^{T} \|\Delta^t\|_F$$
$$\leq \frac{\Lambda_2}{\xi_3^\nu} \sqrt{\frac{(p+q)\log p}{n}} \sum_{t=1}^{T} \|\Delta^t\|_F.$$

For the functions f_2 and f_3,

$$\begin{aligned}
|I_4| &\le \lambda \sum_{j\neq j':(j,j')\in \mathbf{E}} |f_i(\sum_{t=1}^{T}|\omega_{j,j'}^t|) - f_i(\sum_{t=1}^{T}|\omega 0_{j,j'}^t|)| \\
&\le \lambda \sum_{j\neq j':(j,j')\in \mathbf{E}} f_i'(\xi_3 - \sum_{t=1}^{T}\|\Delta^t\|_F)\sum_{t=1}^{T}|\omega_{j,j'}^t - \omega 0_{j,j'}^t| \\
&\le \lambda \sum_{j\neq j':(j,j')\in \mathbf{E}} f'(\xi_3/2)\sum_{t=1}^{T}|\omega_{j,j'}^t - \omega 0_{j,j'}^t| \; for\; sufficiently\; small\; r_n, \\
&\le \lambda \sum_{j\neq j'(j,j')\in \mathbf{E}} \frac{\nu-1}{(\xi_3/2)^\nu}\sum_{t=1}^{T}|\omega_{j,j'}^t - \omega 0_{j,j'}^t| \\
&\le \frac{(\nu-1)\lambda}{(\xi_3/2)^\nu}\sqrt{q}\sum_{t=1}^{T}\|\Delta^t\|_F \\
&\le \frac{(\nu-1)\Lambda_2}{(\xi_3/2)^\nu}\sqrt{\frac{(p+q)\log p}{n}}\sum_{t=1}^{T}\|\Delta^t\|_F,
\end{aligned}$$

where $\nu \ge 1$ and $f_i'(a)$ denotes $\frac{\partial f_i(x)}{\partial x}|x=a$

The second inequality comes from the application of the mean value theorem and the fact that f' is decreasing function as well as $|\omega 0_{j,j'}^t| > \xi_3$.

Combining all the results,

$$\begin{aligned}
\tilde{Q}(\Delta) &\ge -|I_1| + I_2 + I_3 - |I_4| \\
&\ge -C_1\sqrt{\frac{\log p}{n}}\sum_{t=1}^{T}(|\Delta_{\mathbf{E}}^{t-}|_1 + |\Delta_{\mathbf{E}^c}^{t-}|_1) - C_2\sqrt{\frac{p\log p}{n}}\|\Delta^{t+}\|_F + \frac{1}{4\xi_2^2}\sum_{t=1}^{T}\|\Delta^t\|_F^2 + \Lambda_1\sqrt{\frac{\log p}{n}}\sum_{t=1}^{T}|\Delta_{\mathbf{E}^c}^{t-}|_1 - \frac{\Lambda_2}{g(\xi_3)}\sqrt{\frac{(p+q)\log p}{n}}\sum_{t=1}^{T}\|\Delta^t\|_F \\
&\ge (\Lambda_1 - C_1)\sqrt{\frac{\log p}{n}}\sum_{t=1}^{T}|\Delta_{\mathbf{E}^c}^{t-}|_1 - (C_1+C_2)\sqrt{\frac{(p+q)\log p}{n}}\sum_{t=1}^{T}\|\Delta^t\|_F + \frac{1}{4\xi_2^2}\sum_{t=1}^{T}\|\Delta^t\|_F^2 - \frac{\Lambda_2}{g(\xi_3)}\sqrt{\frac{(p+q)\log p}{n}}\sum_{t=1}^{T}\|\Delta^t\|_F \\
&\ge \frac{1}{4\xi_2^2 T}\left(\sum_{t=1}^{T}\|\Delta^t\|_F\right)^2 - (C_1+C_2)\sqrt{\frac{(p+q)\log p}{n}}\sum_{t=1}^{T}\|\Delta^t\|_F - \frac{\Lambda_2}{g(\xi_3)}\sqrt{\frac{(p+q)(\log p)}{n}}\sum_{t=1}^{T}\|\Delta^t\|_F \; for\; \Lambda_1 > C_1 \\
&= \left(\sum_{t=1}^{T}\|\Delta^t\|_F\right)^2\left(\frac{1}{4T\xi_2^2} - \frac{C_1+C_2+\Lambda/g(\xi_3)}{\sum_{t=1}^{T}\|\Delta^t\|_F/\sqrt{\frac{(p+q)(\log p)}{n}}}\right),
\end{aligned}$$

where $g(\xi) = \xi^\nu$ for f_1 penalty function, and $g(\xi) = (\nu-1)(\xi/2)^{-\nu}$ for f_2 and f_3 penalty functions.

Thus, for sufficiently large M, we have $Q(\tilde{\Delta}) > 0$ for any $\Delta \in \partial\, \mathcal{A}$.

Proof of Theorem 2

Proof 4

Define $\mathbf{E}_n = \mathbf{E}_{n,1} \cup \ldots \cup \mathbf{E}_{n,T}$, where $\mathbf{E}_{n,t} = \{(j,j'): j \neq j', \hat{\omega}_{j,j'}^t \neq 0\}$.

We first show that $P(\mathbf{E} \subseteq \mathbf{E}_n)$ converges to 1.

$P(\mathbf{E} \subseteq \mathbf{E}_n) = P(|\hat{\omega}_{j,j'}^t| > 0\; for\; some\; t \in 1,\ldots,T\; for\; all\; (j,j') \in \mathbf{E})$. Since

$\sum_{t=1}^{T}\|\hat{\mathbf{\Omega}}^t_{j,j'}-\mathbf{\Omega}^t_{0,j,j'}\|_F=O_p(\sqrt{\frac{(p+q)\log p}{n}})$ by Theorem 1, one can see that $P(|\hat{\omega}^t_{j,j'}|>0 \text{ for some } t\in 1,\ldots,T \text{ for all } (j,j')\in\mathbf{E})\to P(|\omega^t_{0,j,j'}|>0 \text{ for some } t\in 1,\ldots,T \text{ for all } (j,j')\in\mathbf{E})$ which should be 1 due to the fact that $|\omega^t_{0,j,j'}|>\xi_3>0$ for some t for all $(j, j')\in\mathbf{E}$.

In order to show that $P(\mathbf{E}_n\in\mathbf{E})$ converges to 1, we will show $P(\mathbf{E}^c\subseteq\mathbf{E}^c_n)$ converges to 1.

For this, we need to show that for any $(j, j)\in\mathbf{E}^c$, the derivative $\frac{\partial Q}{\partial\omega^t_{j,j'}}$ has the same sign as $\hat{\omega}^t_{j,j'}$ for all $1\le t\le T$ with probability tending to 1.

We first discuss the f_1 penalty function. The derivative of the objective function can be written as

$$\frac{\partial Q}{\partial\omega^t_{j,j'}}=W_1(t,j,j')+W_2 sign(\omega^t_{j,j'}),$$

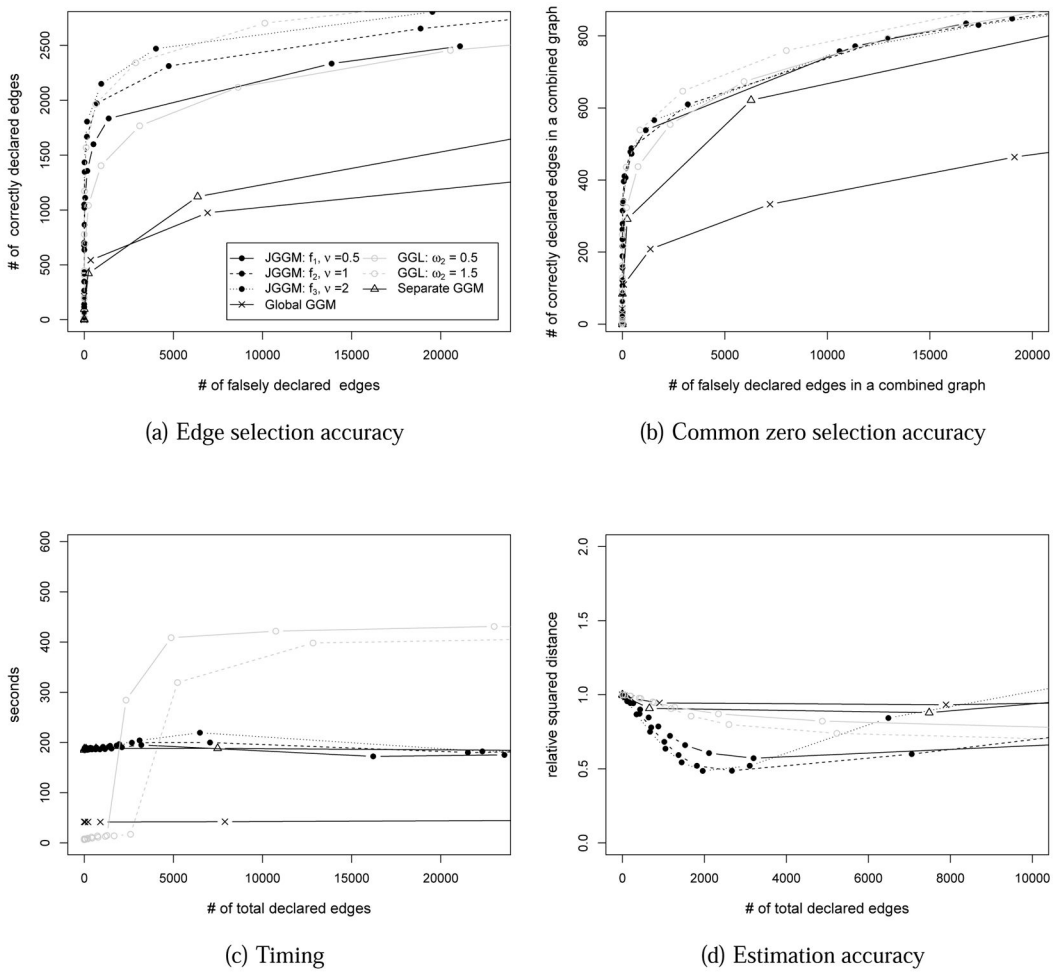
where $W_1(t,j,j')=\mathbf{S}^t_{j,j'}-\sum^t_{j,j'}$ and $W_2=\lambda(1-\nu)(\sum_{t=1}^{T}|\omega^t_{j,j'}|)^{-\nu}$, where $0<\nu 1$.

Arguing as in Theorem 2 of Lam and Fan (2009), one can show that $\max_{t,j,j'}W_1(t,j,j')=O_p((\frac{\log p}{n})^{1/2}+\eta_n^{1/2})$.

For $(j, j')\in\mathbf{E}^c$, $\sum_{t=1}^{T}|\hat{\omega}^t_{j,j'}|=O_p(\eta_n)$ and $\eta_n^{-\nu}$ goes to ∞, and $(\frac{\log p}{n})^{1/2}+\eta_n^{1/2}=O(\lambda)$, W_2 dominates $\max_{(t,j,j')} W_1(t,j,j')$.

For the functions f_2 and f_3, $W_2=\frac{\lambda}{\max((\sum_{t=1}^{T}|\omega^t_{j,j'}|),\varepsilon)}$ and $W_2=\frac{\lambda(\nu-1)}{\max((\sum_{t=1}^{T}|\omega^t_{j,j'}|)^{\nu},\varepsilon^{\nu})}$, respectively, and $\nu>1$.

For $(j, j')\in\mathbf{E}^c$, $\sum_{t=1}^{T}|\hat{\omega}^t_{j,j'}|=O_p(\eta_n)$. Then, $W_2=O_p(\lambda\min(\eta_n^{-\nu},\varepsilon^{-\nu}))$, $\nu\ge 1$ and W_2 dominates $\max_{t,j,j'} W_1(t, j, j')$, by taking sufficiently small ε.



(a) Edge selection accuracy

(b) Common zero selection accuracy

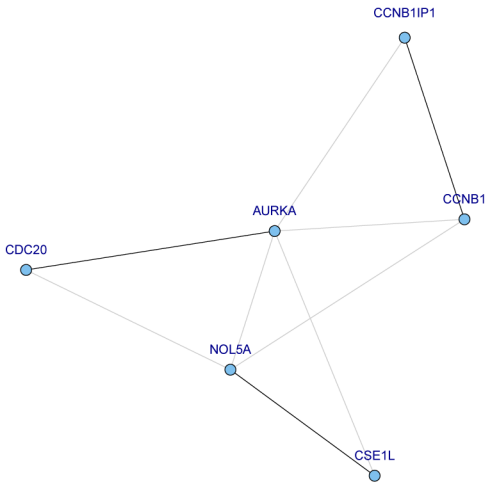
(c) Timing

(d) Estimation accuracy

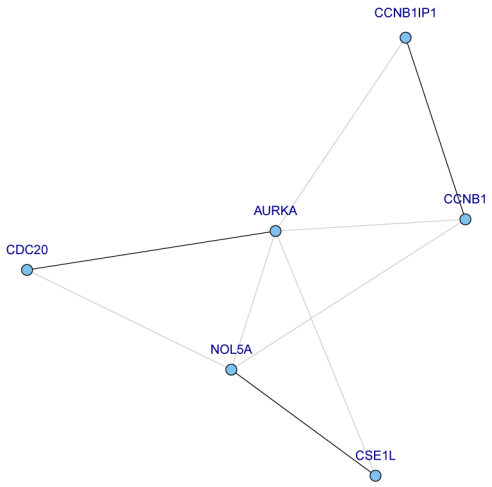
Figure 1.

Performance comparison on simulated data of $n_t = 150$, $p = 500$ and $T = 5$. The performances of joint sparsity GGMs (JGGM) with f_1, $\nu = 0.5$; f_2, $\nu = 1$; and f_3, $\nu = 2$ are compared to the performances of the single global GGM approach and the separate GGM approaches, as well as group graphical lasso (GGL) (Danaher et al., 2012) with $\omega_2 = 0.5$ and $\omega_2 = 1$. (a): The number of correctly declared edges is plotted against the number of falsely declared edges. (b): The number of correctly declared edges in a combined graph is plotted against the number of falsely declared edges in a combined graph. (c): Running time (in seconds) is plotted against the number of total declared edges. (d): The relative squared distance (RSD) of the estimated models from the true models is plotted.

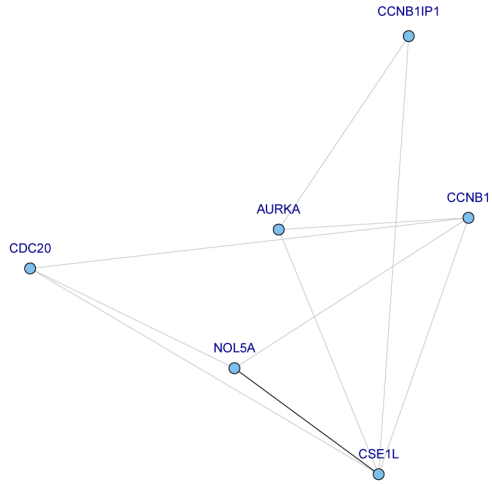
(a) EBV-trans. B-cells and T-cells: Separate approach



(b) EBV-trans. B cells and T-cells: Joint approach



(c) Fibroblasts: Separate approach



(d) Fibroblasts: Joint approach

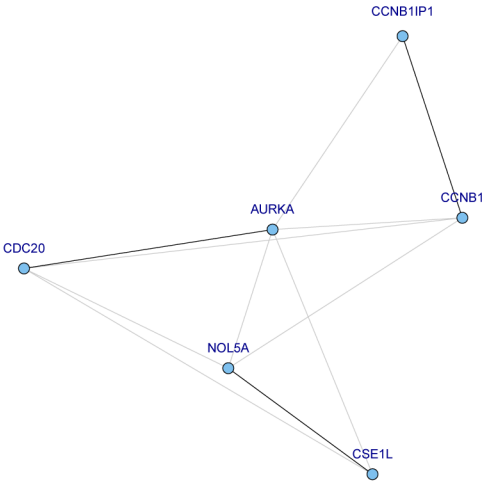


Figure 2.

The estimated GGMs for Embryonic Stem Cell pathway

Edges from the separate approach and joint approach (f_2, $\nu = 1$) are depicted with gray lines, and the edges that match with the KEGG database are colored with black. The graphs of the EBV-transformed B-cells and T-cells are the same in both separate and joint approaches, but that of the fibroblasts has few more edges, which is consistent with the fact that the B-cells and T-cells are more similar to each other.

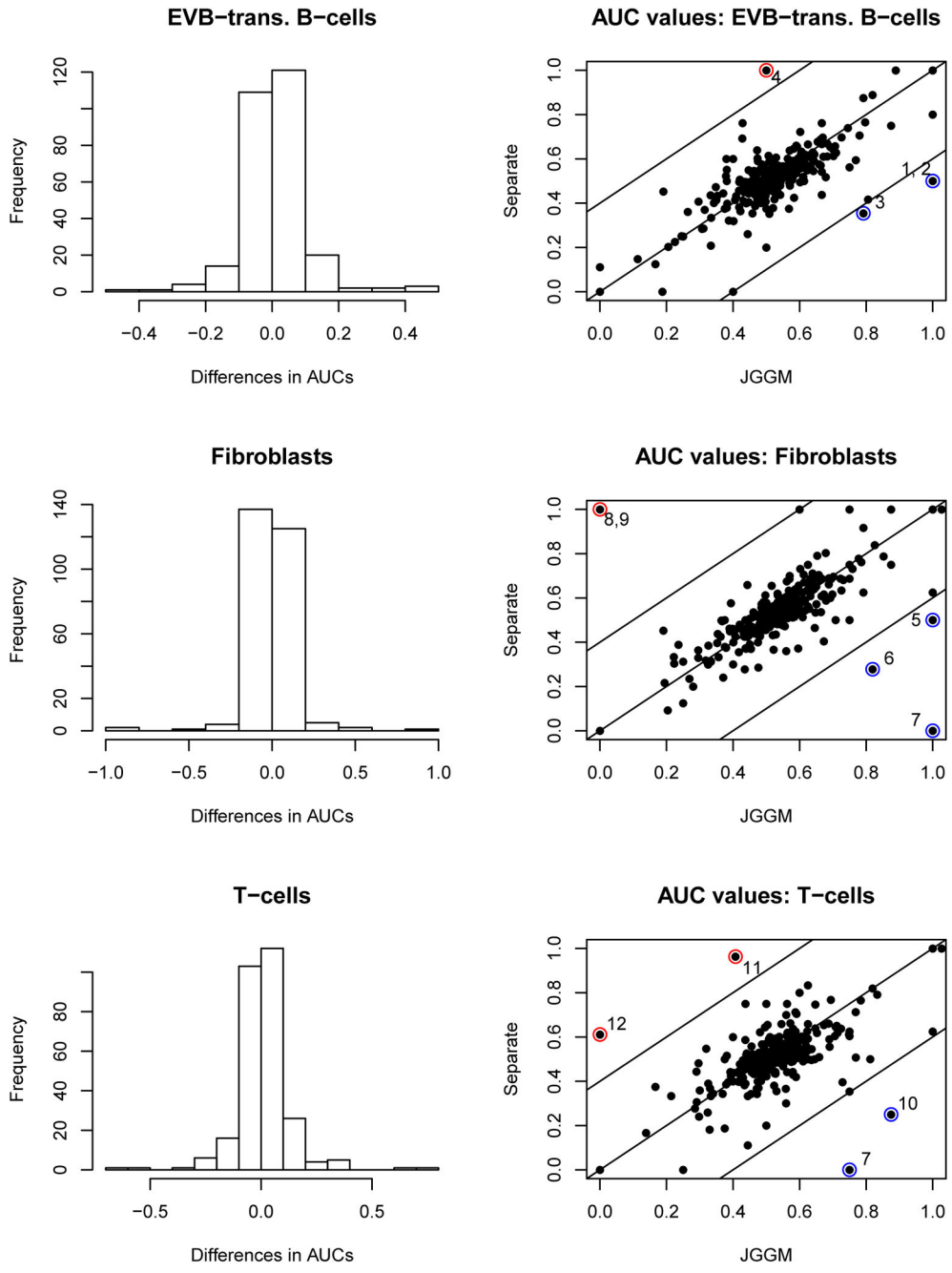


Figure 3.

The histograms of the differences in AUC values (JGGM - Separate) and the scatter plots of the AUC values are presented. 150, 134 and 149 pathways (out of 277) show higher AUC values with the JGGM approach for B-cells, fibroblasts and T-cells, respectively. The pathways with the AUC difference greater than 0.4 are marked and the names are given in Table 2.

Table 1

The selected pathways by using the JGGM approach. For the JGGM and the separate GGM approaches, three AUC values corresponding to three cell types were evaluated for each pathway. For each cell type, the pathways that have AUC values from JGGM higher than 0.8 are listed and the corresponding AUC values from global and separate approaches are shown for comparison. The number of genes and the number of edges from the KEGG database are presented.

Pathway name	AUC			# genes	# edges
	JGGM	Global	Separate		
EBV-transformed B cells					
Basic Mechanisms of SUMOylation pathway	1.00	1.00	0.80	4	5
Basic mechanism of action of PPARa, PPARb(d) and PPARg and effects on gene expression pathway	1.00	1.00	0.50	3	2
Dendritic cells in regulating TH1 and TH2 Development pathway	1.00	0.75	1.00	4	2
IL 18 Signaling Pathway pathway	1.00	1.00	1.00	3	1
Acetylcholine Synthesis	1.00	0.63	1.00	4	2
Beta Oxidation of Unsaturated Fatty Acids	1.00	0.00	1.00	3	2
Cytokines and Inflammatory Response	1.00	1.00	0.50	3	2
Steroid Biosynthesis	1.00	0.00	1.00	3	2
hsa00830 (Retinol metabolism)	0.89	0.56	1.00	4	3
Cytokine Network pathway	0.88	1.00	0.75	4	2
Embryonic Stem Cell	0.82	0.79	0.89	6	3
Free Radical Induced Apoptosis pathway	0.81	0.74	0.42	6	6
Primary Fibroblasts					
Basic Mechanisms of SUMOylation pathway	1.00	1.00	0.80	4	5
Basic mechanism of action of PPARa, PPARb(d) and PPARg and effects on gene expression pathway	1.00	1.00	1.00	3	2
Dendritic cells in regulating TH1 and TH2 Development pathway	1.00	0.75	0.63	4	2
IL 18 Signaling Pathway pathway	1.00	1.00	0.50	3	1
Cytokines and Inflammatory Response	1.00	1.00	1.00	3	2
Steroid Biosynthesis	1.00	0.00	0.00	3	2
hsa00750 (Vitamin B6 metabolism)	1.00	0.60	1.00	4	5
Cytokine Network pathway	0.88	1.00	0.75	4	2
Acetylcholine Synthesis	0.88	0.63	1.00	4	2
hsa05010 (Alzheimer's disease)	0.85	0.58	0.79	8	6

Pathway name	AUC			# genes	# edges
	JGGM	Global	Separate		
EBV-transformed B cells					
Th1 Th2 Differentiation pathway	0.83	0.53	0.84	18	11
Embryonic Stem Cell	0.82	0.79	0.28	6	3
Primary T-cells					
Basic Mechanisms of SUMOylation pathway	1.00	1.00	1.00	4	5
Basic mechanism of action of PPARa, PPARb(d) and PPARg and effects on gene expression pathway	1.00	1.00	1.00	3	2
Dendritic cells in regulating TH1 and TH2 Development pathway	1.00	0.75	0.63	4	2
IL 18 Signaling Pathway pathway	1.00	1.00	1.00	3	1
Beta Oxidation of Unsaturated Fatty Acids	1.00	0.00	1.00	3	2
Cytokines and Inflammatory Response	1.00	1.00	1.00	3	2
Cytokine Network pathway	0.88	1.00	0.25	4	2
Cycling of Ran in nucleocytoplasmic pathway	0.83	0.71	0.79	5	6
Embryonic Stem Cell	0.82	0.79	0.82	6	3
Acetylcholine Synthesis	0.81	0.63	0.50	4	2

Table 2

The names of pathways that show the differences of AUC values greater than 0.4.

1	Basic mechanism of action of PPARa, PPARb(d) and PPARg and effects on gene expression pathway
2	Cytokines and Inflammatory Response
3	hsa04940 (Type I diabetes mellitus)
4	Degradation of the RAR and RXR by the proteasome pathway
5	IL 18 Signaling Pathway pathway
6	Embryonic Stem Cell
7	Steroid Biosynthesis
8	Oxidative reactions of the pentose phosphate pathway pathway
9	hsa00471(D-Glutamine and D-glutamate metabolism)
10	Cytokine Network pathway
11	Inhibition of Huntington’s disease neurodegeneration by histone deacetylase inhibitors pathway
12	Rho-Selective Guanine Exchange Factor AKAP13 Mediates Stress Fiber Formation pathway

Table 3

Performance as a function of n, p and T. Means (standard deviations) over 100 replicates are shown for five criteria defined in Section 3.1. The tuning parameter λ was selected by using aBIC criterion.

Setting			Tuning parameter		Comparison criteria				
p	n_t	T	Penalty type	$\log_{10} 2\,\lambda/n_t$	Criterion C1	Criterion C2	Criterion C3	Criterion C4	Criterion C5
500	150	5	f_1, $\nu = 0.5$	−1.0	173.98 (20.58)	1355.98 (52.67)	155.88 (16.92)	406.22 (15.79)	0.66 (0.01)
500	150	5	f_2, $\nu = 1$	−1.4	699.42 (47.57)	1970.92 (60.20)	439.58 (26.00)	488.00 (15.13)	0.49 (0.01)
500	150	5	f_3, $\nu = 2$	−1.9	404.38 (44.84)	1981.66 (67.37)	168.32 (16.62)	437.74 (15.77)	0.48 (0.01)
500	150	2	f_1, $\nu = 0.5$	−1.1	313.90 (26.22)	393.72 (25.70)	299.38 (24.33)	257.20 (16.58)	0.78 (0.02)
500	150	2	f_2, $\nu = 1$	−1.5	343.98 (30.90)	448.36 (28.54)	293.58 (24.61)	264.82 (17.02)	0.72 (0.02)
500	150	2	f_3, $\nu = 2$	−2.3	659.60 (181.48)	527.50 (33.94)	476.40 (156.66)	291.80 (20.66)	0.85 (0.04)
1000	150	2	f_1, $\nu = 0.5$	−0.8	132.76 (17.06)	383.10 (24.56)	131.78 (16.84)	285.26 (16.42)	0.91 (0.01)
1000	150	2	f_2, $\nu = 1$	−1.2	527.52 (35.33)	625.80 (29.85)	502.08 (33.42)	408.60 (19.34)	0.81 (0.01)
1000	150	2	f_3, $\nu = 2$	−2.2	4195.08 (97.95)	1079.98 (39.09)	3648.30 (82.54)	642.00 (25.28)	0.97 (0.02)
1000	500	2	f_1, $\nu = 0.5$	−1.5	1774.14 (65.72)	2481.38 (32.03)	1674.50 (60.00)	1445.36 (20.29)	0.29 (0.01)
1000	500	2	f_2, $\nu = 1$	−2.0	1301.30 (60.64)	2519.94 (34.97)	1072.66 (46.47)	1417.20 (21.53)	0.24 (0.01)
1000	500	2	f_3, $\nu = 2$	−3.0	1920.64 (74.52)	2611.12 (35.09)	1273.74 (47.60)	1434.72 (20.99)	0.26 (0.01)