Biometrics 000, 000–000
000 0000

DOI: 000

Information Enhanced Model Selection for High-Dimensional Gaussian Graphical Model with Application to Metabolomic Data

Jie Zhou[1], Anne Hoen[1,2,*], Susan McRitchie[3], Wimal Pathmasiri[3], Juliette Madan[2], and Weston Viles[4], Erika Dade[2] Margaret Karagas[2], Jiang Gui[1,]**

[1] Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, 1 Medical Center Drive, Hanover, NH, U.S.A
[2] Depatment of Epidemiology, Geisel School of Medicine, Dartmouth College, 1 Medical Center Drive, Hanover, NH, U.S.A
[3]Nutrition Research Institute, University of North Carolina, 500 Laureate Way, Kannapolis, NC, U.S.A.
[4]Department of Mathematics and Statistics, University of Southern Maine, Portland, Maine, U.S.A
*email: anne.g.hoen@dartmouth.edu
**email: jiang.gui@dartmouth.edu

Summary: In light of the low signal-to-noise nature of many large biological data sets, we propose a novel method to identify the structure of association networks using a Gaussian graphical model combined with prior knowledge. Our algorithm includes the following two parts. In the first part we propose a model selection criterion called structural Bayesian information criterion (SBIC) in which the prior structure is modeled and incorporated into the Bayesian information criterion (BIC). It is shown that the popular extended BIC (EBIC) is a special case of SBIC. In second part we propose a two-step algorithm to construct the candidate model pool. The algorithm is data-driven and the prior structure is embedded into the candidate model automatically. Theoretical investigation shows that under some mild conditions SBIC is a consistent model selection criterion for the high-dimensional Gaussian graphical model. Simulation studies validate the superiority of the SBIC over the standard BIC and show the robustness to the model misspecification. Application to relative concentration data from infant feces collected from subjects enrolled in a large molecular epidemiologic cohort study validates that prior knowledge on metabolic pathway involvement is a statistically significant factor for the conditional dependence among metabolites. More importantly, new relationships among metabolites are identified through the proposed algorithm which can not be covered by conventional pathway analysis. Some of them have been widely recognized in the literature.

Key words: Gaussian graphical model; Prior structure; Structural BIC; Two-step algorithm; Pathway analysis.

1. Introduction

Modern 'omics technology can easily generate thousands of measurements in a single run which provides an opportunity for researchers to explore complex relationships in biology. However, It has been widely recognized that biological measurements are usually accompanied by a low ratio of signal-to-noise making detection of effect challenging and final conclusions unreliable. As previously reported in Ideker et al (2011), prior knowledge can play a pivotal role in deciphering this kind of complexity. For example, Segre et al (2010) drew on the prior knowledge on mitochondrial genes sets to investigate whether mitochondrial dysfunction is a cause of the common form of diabetes. Roach et al (2010) identified the gene that causes Miller syndrome based on the human genome reference map. For more work on the application of prior biological knowledge, see Boluki et al (2017); Imoto et al (2004) and Ma (2015). In this paper, our aim is to identify the metabolite network based on pathway analysis.

Biological network, such as microbe-microbe interaction networks, metabolite networks and gene regulation networks have received much attention in recent years. Based on the random graph theory, many algorithms have been proposed in statistics to explore the structure of nework, see Lauritzen (1996). In this respect, Friedman et al (2008); Meinshansen and Bühlmann (2006) investigated the identification problem for high-dimensional Gaussian undirected graphical model, while Cheng et al (2014); Ravikumar et al (2010); Wainwright and Jordan (2003) studied the identification of discrete network modeled by high-dimensional Ising model. In order to deal with the prior structure of network, the Bayesian method is the typical choice in literature. However, finding a realistic prior distribution for the metabolite network is difficult. For the popular choice of conjugate G-Wishart distribution, the complex sampling algorithms from the posterior distribution have hindered its wide use in practice, see Roverato (2002). Ma (2015) considered this problem under the frequentist framework. However Ma (2015) only focused on the deterministic prior structure which in most situations is an unrealistic assumption.

In this paper, we propose a novel method to identify the

structure of a graphical model based on prior information. Contrary to Ma (2015), here uncertainty in prior information is taken into account. Specifically, the algorithm includes the following two parts. In the first part we propose a structural Bayesian information criterion (SBIC) based on Boltzmann distribution which incorporates the prior structure. For high-dimensional models, the extended Bayesian information criterion (EBIC) has been widely used in literature for model selection, see Bogdan et al (2004); Chen and Chen (2008, 2012); Foygel and Drton (2010) for details. Compared to EBIC, SBIC provides a more flexible framework and EBIC can be regarded as a special case of SBIC with null prior structure. In second part, based on the prior structure, we propose a data-driven two-step algorithm to build the model pool. The graph is enriched in the first step and pruned in the second step. This part can be implemented readily by using the R package *glmnet*. Through simulation studies it is shown that the combination of SBIC and two-step algorithm can effectively deal with the prior structure for high-dimensional graphical model and improve the analysis results. As a theoretical basis, for high-dimensional sparse Gaussian graphical models, it is shown that SBIC is consistent for model selection under mild conditions. With the proposed algorithm in hand, we studied ^{1}H NMR-based metabolite data profiled in infant feces collected as part of the New Hampshire Birth Cohort Study, a large prospective cohort study of mothers and their children born in New Hampshire, see Madan et al (2016) for details. The prior structure for these metabolites is constructed based on the related pathway information from the biological database Kyoto Encyclopedia of Genes and Genomes (KEGG). Our results show that pathways have statistically significant effects on the conditional dependence among metabolites. The probability of existence of dependent relationships between two metabolites increases if the proportion of shared pathways increases. Furthermore our approach reveals new relationships among metabolites that can not be identified through standard pathway analysis, though many of which are validated in the literature.

The paper is organized as follows. Section 2 reviews the Gaussian undirected graphical model and related EBIC. A new formulation of EBIC will be introduced. In Section 3, we present our main algorithm. Section 3.1 will elaborate on the definition of structural BIC and its implications. Section 3.2 describes the two-step algorithm for building the candidate model pool. Theoretical results of SBIC will be given in Section 4. In Section 5, the algorithm is evaluated through simulated data. In Section 6 we use the algorithm to investigate the metabolomic data from the New Hampshire Birth Cohort Study. Section 7 concludes with some comments.

2. Gaussian Graphical Model and BIC

2.1 *A brief review of BIC for Gaussian graphical model*

Given p-dimensional normal random vector $\mathbf{X} = (X_1, X_2, \cdots, X_p)^T \sim N(\boldsymbol{\mu}_p, \Sigma_{p\times p})$, an undirected graph is used to depict the conditional dependent relationship among $\mathbf{X}$. If X_i and X_j are independent given all the other components of $\mathbf{X}$, then there is no edge between X_i and X_j otherwise there is an edge between them. The precision matrix is defined as $\Omega_{p\times p} = (\omega_{ij}) = \Sigma^{-1}$. It turns out that the precision matrix completely describes such conditional dependence. Given n i.i.d observations $\tilde{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^T$, our aim is to identify the nonzero components in $\tilde{p} \triangleq p(p-1)/2$ off-diagonal entries in Ω. In its general form, BIC can be stated as follows. Let $\mathcal{E}$ be the model space under consideration with $\pi(E)$ the prior distribution defined on $\mathcal{E}$. Let θ denote the unknown parameter in E with prior distribution $p(\theta)$. With θ in hand, let the density function for $\tilde{X}$ be $f(\tilde{X}|\theta)$ so that the likelihood for observations $\tilde{X}$ can be expressed as $l(\tilde{X}|E) = \int f(\tilde{X};\theta)p(\theta)d\theta$. The posterior distribution of model E can be expressed as

$$p(E|\tilde{X}) = \frac{l(\tilde{X}|E)\pi(E)}{\sum_{E\in\mathcal{E}} l(\tilde{X}|E)\pi(E)}. \tag{1}$$

Through Laplace's method of integration, the following approximation up to a constant can be obtained for $-2\log p(E|\tilde{X})$,

$$-2\log f(\tilde{X}|\theta) + |E|\log n - |E|\log(2\pi) - 2\log p(\theta|E) \tag{2}$$
$$+\log\det(V) - 2\log\pi(E), \tag{3}$$

where V is the expected information matrix for a single observation and $|E|$ is the degree of freedom of model E. By omitting the last four terms which do not involve the sample size n, we get the standard BIC, $\mathrm{BIC}(E) = -2l_n(E) + |E|\log n$ with $l_n(E) = \log f(\tilde{X}|\theta)$. For the high-dimensional regression model, Bogdan et al (2004); Chen and Chen (2008, 2012) proposed extended BIC (EBIC) which puts more weight on sparse model than standard BIC. Foygel and Drton (2010) further generalized EBIC to the Gaussian graphical model which has the following form,

$$\mathrm{EBIC}_\lambda = -2l_n(\Omega(E)) + |E|\log n + 4|E|\lambda\log p, \tag{4}$$

where $\Omega(E)$ is the precision matrix associated with model E. Tuning parameter $0 \leqslant \lambda \leqslant 1$ controls the model complexity. When $\lambda = 0$, EBIC reduces to the standard BIC. As λ becomes larger, (4) will put more weight on the sparse model. The log-likelihood function $l_n(\Omega(E))$ in (4) for the Gaussian graphical model has the following form,

$$l_n(\Omega) = \frac{n}{2}[\log\det(\Omega) - \mathrm{trace}(S\Omega)], \tag{5}$$

where S is the empirical covariance matrix. Foygel and Drton (2010) proved that under the given assumptions, (4) is a consistent model selection criterion for high-dimensional Gaussian graphical model.

Although EBIC has been widely used in the literature for high-dimensional model selection, several limitations have yet to be addressed. For example, EBIC does not take prior information into account. In practice, it is often desired that we can adapt (4) to reflect such prior biological knowledge. Also, the choice of λ has a potentially large impact on the final result. It is helpful to find a proper way to select λ. With these motivations in mind, in Section 3 we propose a new algorithm for the selection of Gaussian graphical model which aims to address these problems. Though we have focused on Gaussian graphical model in this paper, the algorithm can be easily

adapted to accommodate the discrete graphical model such as Ising model.

2.2 *A new formulation of EBIC*

In this section we introduce a different way to formulate EBIC which will facilitate the introduction of prior structure in Section 3. For any given pair of nodes, (X_i, X_j), define the edge variable Z_{ij} equal to one if there exists an edge between X_i and X_j and zero otherwise, i.e., Z_{ij} is the indicator variable for the existence of the edge between nodes X_i and X_j. Due to the symmetry of undirected graph, we have $Z_{ij} = Z_{ji}$ $(1 \leqslant i < j \leqslant p)$. We pool all the Z_{ij} together and define a $p(p-1)/2$-dimensional random vector $\mathbf{Z} = (Z_{12}, Z_{13}, \cdots, Z_{(p-1)p})^T \triangleq (Z_1, \cdots, Z_m)^T$ with $m = p(p-1)/2$. The prior information about the structure of E can be completely described by the probability distribution of $\mathbf{Z}$. Here Boltzmann distribution is employed to model $\mathbf{Z}$. Boltzmann distribution, which originated from statistical physics, has been widely used to model the stochastic phenomenon. Formally Boltzmann distribution can be formulated as,

$$\Pr(\mathbf{Z} = \mathbf{z}) \propto \exp\left(-\frac{\epsilon(\mathbf{z})}{KT}\right), \tag{6}$$

where $\epsilon(\mathbf{z}) \geqslant 0$ is the energy function corresponding to state $\mathbf{z}$, T the temperature parameter and K the Boltzmann constant. Without loss of generality, $K = 2$ will always be assumed in the following. Substitution of (6) into (1) and (2) leads to the following form of BIC for Gaussian graphical model,

$$\mathrm{BIC}_{T,\epsilon}(\mathbf{z}) = -2l_n(\Omega(\mathbf{z})) + |\mathbf{z}| \log n + \epsilon(\mathbf{z})/T, \tag{7}$$

where $|\mathbf{z}|$ denotes the number of nonzero components in $\mathbf{z}$. In order to use (7) in practice, we have to specify the form of $\epsilon(\mathbf{z})$. Among many other possible choices, we consider the following specification,

$$\mathrm{BIC}_{T,W}(\mathbf{z}) = -2l_n(\Omega(\mathbf{z}) + |\mathbf{z}| \log n + \mathbf{z}^T W \mathbf{z}/T, \tag{8}$$

where W is a positive semi-definite matrix. In (8) energy function can be regarded as the squared weighted Euclidean distance between two states, $\mathbf{z}$ and $\mathbf{0}$. It is obvious that (8) includes standard BIC and EBIC as special cases. In fact if $W = 0$, (8) is the standard BIC; if $T = 1/(4\lambda)$, and $W = (\log p)\mathrm{I}_{\tilde{p}}$ with $\mathrm{I}_{\tilde{p}}$ the $\tilde{p} \times \tilde{p}$ identity matrix, then (8) reduces to the EBIC in (4)-(5). With such a specification of W in EBIC, it is straightforward to show that the components of $\mathbf{Z}$ are independent Bernoulli variables with nonzero probability $\frac{1}{1+p^{2\lambda}}$. Such probabilistic explanation can guide us to choose the tuning parameter λ involved in EBIC (4). For example for $\lambda = 0.5$, or equivalently $T = 0.5$, which is often recommended in literature, it implies that the prior mean of total edges is $\tilde{p}/(1+p) \approx (p-1)/2$. More generally it can be seen that for $T > 0$, we have $P(Z_i = 1) < 0.5$ while for $T < 0$, we have $P(Z_i) > 0.5$. So for the graph with $E|\mathbf{Z}| < \tilde{p}/2$, $T > 0$ is a more plausible choice.

In some circumstances, prior information involves not only the mean of the total edges but also its variance which can also be modeled through $\mathrm{BIC}_{T,W}$. Specifically, consider the following form of W for $\mathrm{BIC}_{T,W}$,

$$W(\rho) = D^T R(\rho) D \tag{9}$$

with $D = \mathrm{diag}(\sqrt{\log p}, \cdots, \sqrt{\log p})$, and $R = \rho J_{\tilde{p}} + (1-\rho) I_{\tilde{p}}$ for some $0 \leqslant \rho < 1$. Here $J_{\tilde{p}}$ is the $\tilde{p} \times \tilde{p}$ matrix with all the entries being 1. There is a one-to-one correspondence between (ρ, T) and (μ, σ^2), the mean and variance of total edge. The details about the formulas are given in Appendix. So given prior information about (μ, σ^2), the corresponding parameter (T, ρ) can be easily determined which in turn can be used in $\mathrm{BIC}_{T,W}$ for model selection.

3. Incorporation of Prior Structure into Model Selection

3.1 *Prior structure enhanced BIC for Gaussian graphical model*

Now let us consider how to adapt $\mathrm{BIC}_{T,W}$ (8) to accommodate the specific structure information. Consider the following common scenario in biology. For $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, suppose that the graph $\tilde{G} = (V, \tilde{E})$ is the prior strucure (e.g., constructed based on some biological theory) and we have to identify the true graph structure based on $\tilde{G}$ and the observations on $\mathbf{X}$. First we introduce the concept of difference graph. For two graphs $\tilde{G}$ and $G = (V, E)$, the difference graph of $\tilde{G}$ and G is defined as the graph which has the same nodes as $\tilde{G}$ and G while the edge set is $\bar{E} = \tilde{E} \triangle E$ and denoted by $\bar{G} = \tilde{G} \triangle G \triangleq (V, \bar{E})$. Here $\triangle$ stands for the symmetrical difference operator between two sets. For a given prior edge set $\tilde{E}$, there is a one-to-one correspondence between $\bar{E}$ and E. Equivalenly, there is a one-to-one correspondence between their edge variable vector, $\bar{\mathbf{Z}} = \mathrm{I}(\tilde{\mathbf{Z}} - \mathbf{Z})$ and $\mathbf{Z}$. Replace $\mathbf{z}$ in the third term in $\mathrm{BIC}_{T,W}$ by $\bar{\mathbf{z}}$, we obtain the following structural Bayesian information criterion (SBIC),

$$\mathrm{SBIC}_{T,W}(\mathbf{z}) = -2l_n(\Omega(\mathbf{z}) + |\mathbf{z}| \log n + \bar{\mathbf{z}}^T W \bar{\mathbf{z}}/T, \tag{10}$$

in which the first term measures the fitness between model and data, the second term measures the model complexity and the third term measures the deviation of the model from the prior structure. Minimization of (10) will lead to solutions that achieve balance between these terms. Essentially we have assumed that $\bar{\mathbf{Z}}$ in (10) has Boltzmann distribution,

$$P(\bar{\mathbf{Z}} = \bar{\mathbf{z}}) \propto \exp\left(-\frac{\bar{\mathbf{z}}^T W \bar{\mathbf{z}}}{2T}\right). \tag{11}$$

If we set $W = \mathrm{diag}(\log p, \cdots, \log p)$ just like EBIC, then (10) reduces to

$$\mathrm{SBIC}_T(\mathbf{z}) = -2l_n(\Omega(\mathbf{z}) + |\mathbf{z}| \log n + |\bar{\mathbf{z}}| \log p/T, \tag{12}$$

which will be used in the numerical studies in Section 5 and 6.

Remark. (i) If $\tilde{\mathbf{z}} = 0$, i.e., the prior structure is a graph with no edges, then SBIC in (12) reduces to EBIC in (4). So EBIC is a special case of SBIC. (ii) If T is large enough, then the model selected by SBIC is the same as that from standard BIC. If T is small enough, then the model selected by SBIC is the prior structure. For other T, the model selected by SBIC will be a compromise of these two extreme cases. (iii) The choice of T in (12) relies on the expected error rate of prior structure. The expected error rate is defined as $r = \frac{m_1+m_2}{\tilde{p}}$, where m_1 is the number of true edges that have been missed by prior structure while m_2 is the number of edges that have

been mistakenly added to the prior structure. Note we can always assume $0 \leqslant r \leqslant 0.5$ and $r = 0.5$ will lead to the standard BIC. There is an one-to-one correspondence between T and r. The more intuitive explanation of r can guide us to find the appropriate value for T.

The generalization of (12) is possible. For example in (12) it has been implicitly assumed that the probability of adding an edge to the prior graph, p_1, and the probability of deleting an edge from the prior graph, p_2, is equal. In some cases compared to pruning edge, we may be more inclined to add edges to the prior graph, i.e., $p_1 > p_2$. The following simple generalization of (12) can accommodate such situation,

$$\begin{aligned}\mathrm{SBIC}_{T_1,T_2}(\bar{\mathbf{z}}) &= -2l_n(\Omega(\mathbf{z}) + |\mathbf{z}|\log n + |\bar{\mathbf{z}}_\mathbf{1}|\log p/T_1 \\ &\quad +|\bar{\mathbf{z}}_\mathbf{2}|\log p/T_2 \end{aligned} \tag{13}$$

where $\bar{\mathbf{z}}_1$ is the indicator vector of whether the entries of $(\tilde{\mathbf{z}} - \mathbf{z})$ are 1 while $\bar{\mathbf{z}}_2$ is the indicator vector for -1. If $T_1 < T_2$, then SBIC_{T_1,T_2} favor the graphs which share more edges with prior structure. The cost for such flexibility is that we have to specify the values for both T_1 and T_2.

3.2 *Construction of candidate model pool based on prior structure*

From Example 1 in Section 5, we can see that with the aid of prior structure, structural BIC can outperform the standard BIC. Note that there are only six variables involved in Example 1 and consequently the exhaustive search in the model space is possible. As the number of variables gets larger, it becomes unrealistic to carry out a exhaustive search in the model space and we have to choose a subset of the model space as the candidate model pool. A common practice for the construction of candidate model pool for high-dimensional model is to use the solution path of lasso. The disadvantage of such a practice is that the models in the model pool have nothing to do with the prior structure. Even with SBIC in hand, we still have a high probability to end up with a poor model. It is necessary to incorporate the prior structure into the construction of model pool. There are multiple methods to get this done. For example in addition to the solution path of lasso, we may simply include random samples from the Boltzmann distribution corresponding to the prior structure as a part of the model pool. However this method turns out to be very inefficient for high-dimensional model. An alternative way is to adapt the penalty term in lasso using the prior structure so that the resulted solution path can automatically be related to the prior structure. Similar idea has been investigated under the name of generalized lasso, e.g., Tibshirani and Taylor (2011). For present situation, without loss of generality, let us consider the node X_i and its neighborhood. Given the prior structure, let $\tilde{\mathbf{z}}_0^{(i)}$ be a $(p-1)$-dimensional vector with components 0 or 1, in which 0 indicates no association while 1 means association with X_i in prior structure. Then the model pool may be constructed by solving a series of the following optimization problems,

$$\min_{\boldsymbol{\beta}^{(i)}} \|\mathbf{x}_i - \sum_{j\neq i} \beta_j^{(i)}\mathbf{x}_j\|_2^2 + \lambda\|\mathbf{z}^{(i)} - \tilde{\mathbf{z}}_0^{(i)}\|_0, \tag{14}$$

where vector $\mathbf{z}^{(i)}$ is the indicator vector of $\boldsymbol{\beta}^{(i)} = (\beta_1^{(i)}, \cdots, \beta_{i-1}^{(i)}, \beta_{i+1}^{(i)}, \cdots, \beta_p^{(i)})^T$ for a given model as $\tilde{\mathbf{z}}_0^{(i)}$. For a large λ, the nonzero components of the resulting solution to (14) will be the same as the prior structure. As λ deceases, the solution will include more edges that have not appeared in the prior structure. In the extreme case of $\lambda = 0$, as in standard lasso, all the edges will be selected.

Note (14) is not a convex optimization problem and there is no existing software to solve (14). In the following we propose a two-step algorithm to build the model pool. The algorithm can also incorporate the prior structure into the candidate model in the meanwhile can be easily implemented based on the existing R package such as *glmnet*. Simulation results show that the model pool constructed by two-step algorithm has a big advantage over standard lasso. Specifically, given ith node, the algorithm consists the following two steps.

Forward Step (Enrichment). In this step the prior structure of the graph is fixed and we consider how to select the nodes from the rest nodes and add them into the neighborhood of ith node. Let $A_i = (A_{i1}, A_{i2})$ in which A_{i1} is the indices of the nodes that have appeared in prior neighborhood of X_i while A_{i2} is the indices of the rest nodes. For a given increasing sequence, $0 \leqslant \lambda_1^{(1)} < \cdots < \lambda_1^{(m_1)}$, this step can be accomplished by solving the following pm_1 optimization problems,

$$\hat{\boldsymbol{\beta}}^{(i)} \triangleq \arg\min_{\boldsymbol{\beta}^{(i)}} \|\mathbf{x}_i - \sum_{j\neq i} \beta_j^{(i)}\mathbf{x}_j\|_2^2 + \lambda_1^{(k)}\|\boldsymbol{\beta}^{(i)}(A_{i2})\|_1 \tag{15}$$

for $i = 1, \cdots, p$, $k = 1, \cdots, m_1$. Through (15) we aim to pick up the nodes that have been omitted by the prior structure. Denote by A_{i3} the nodes that appear in the solution $\hat{\boldsymbol{\beta}}^{(i)}$. Combination of $A_{13}, \cdots, A_{p3}$ leaves us m_1 graphs denoted by $G^{(k)}$ for $k = 1, \cdots, m_1$ respectively.

Backward Step (Pruning). Note each $G^{(k)}$ $(k = 1, \cdots, m_1)$ from first step contains the prior structure. In order to prune the redundant edges in prior structure, for a given increasing sequence, $0 \leqslant \lambda_2^{(1)} < \cdots < \lambda_2^{(m_2)}$, we solve the following pm_1m_2 optimization problems,

$$\hat{\boldsymbol{\beta}}^{(i)}(A_{i3}) \triangleq \arg\min_{\boldsymbol{\beta}^{(i)}(A_{i3})} \|\mathbf{x}_i - \sum_{j\neq i} \beta_j^{(i)}\mathbf{x}_j\|_2^2 + \lambda_2^{(h)}\|\boldsymbol{\beta}^{(i)}(A_{i3})\|_1 \tag{16}$$

for $i = 1, \cdots, p$, $h = 1, \cdots, m_2$ and $k = 1, \cdots, m_1$. Here A_{i3} is a given neighborhood in $G^{(k)}$. The final index set form (16) is denoted by A_{i4}. Combination of $A_{14}, \cdots, A_{p4}$ leaves us a graph $G^{(kh)}$ for $k = 1, \cdots, m_1$ and $h = 1, \cdots, m_2$. Thus there are total m_1m_2 candidate models in the final model pool.

Remark. If the prior structure is a graph with no edge, then only *Forward step* is involved to build the model pool. If the prior structure is a complete graph, then only *Backward step* is involved. The model pools for these two extreme cases turn out to be the same as that from the standard lasso algorithm.

4. Theoretical Results

In this section we investigate the theoretical properties of SBIC. It is shown that under the given assumptions, SBIC can consistently select the underlying model for high-dimensional Gaussian graphical model where the number of nodes may increase as sample size increases.

First let us introduce some notations for the ease of exposition. Recall $\mathbf{z}$ is the $p(p-1)/2$-dimension vector indicating whether there is an edge between given two vertices. Define $|\mathbf{z}| = \sum_{i=1}^{p(p-1)/2} z_i$ and let $\mathbf{z_0}$ be the vector corresponding to the true graph E_0 under consideration. We confine ourselves to the graphs with no more than q edges and let $\mathcal{E}_q$ denote such graph set with $\mathcal{Z}_q \subset R^{p(p-1)/2}$ the corresponding indicator vector set. Let σ^2_{max} be the largest diagonal component of the true covariance matrix Σ_0, λ_{max} be the largest eigenvalue of true precision matrix Θ_0 and $\tau_{\max}$ and $\tau_{\min}$ are the the largest and smallest eigenvalue of W respectively. With these notations in hand, the consistency for $\mathrm{BIC}_{T,W}$ (8) and SBIC (12) are proved in Theorem 1 and 2 respectively. For $\mathrm{BIC}_{T,W}$, the following assumptions are involved.

Assumption 1. $E_0 \in \mathcal{E}_q$ is decomposable;

Assumption 2. $p = O(n^\kappa)$ for some $0 < \kappa < 1$;

Assumption 3. $\exists$ constant $C > 0$ such that $\sigma^2_{max}\lambda_{max} \leqslant C$ and $\theta_0 = \min_{e\in E_0} |(\Theta_0)_e| > 0$

Assumption 4. $\exists \epsilon > 0$ such that $0 < 2T(4+\epsilon-\frac{1}{2\kappa})\log p \leqslant \tau_{\min} \leqslant \tau_{\max} = o(p)$.

Theorem 1. Under Assumptions 1-4, the model selection procedure based on $\mathrm{BIC}_{T,W}$ given in (8) is consistent, i.e., as $n \to \infty$ we have

$$\mathbf{z_0} = \arg\min_{\mathbf{z}\in\mathcal{Z}_q} \mathrm{BIC}_{T,W}(\mathbf{z}) \tag{17}$$

in probability.

Now let's consider SBIC (12) in which prior structure is available for the underlying graphical model. Recall $\tilde{G} = (V, \tilde{E})$ is the prior structure, $G_0 = (V, E_0)$ is the true graph and $\bar{G} = (V, \bar{E})$ is the difference graph of $\tilde{G}$ and G. Particularly $\bar{G}_0$ is the difference graph of $\tilde{G}$ and G_0. Here we have assumed $\tilde{G}$ and G_0 have the same nodes.

Assumption 1$'$ $\tilde{E} \in \mathcal{E}_{q_1}$, $\bar{E}_0 \in \mathcal{E}_{q_2}$ for some integers q_1 and q_2 and E_0 is decomposable.

Assumption 4$'$ For $\kappa_0 = \frac{1}{\kappa} - \gamma > 0$, $\exists\ \epsilon > 0, 0 < \tau < 1$ such that $\tau\kappa_0 > 4 + \epsilon$.

Assumption 1$'$ says that $\bar{\mathbf{z}}_\mathbf{0}$ has at most q_2 nonzero components which means that we can reach the true model E_0 by adding or deleting at most q_2 edges from the prior model $\tilde{E}$ and so $E_0 \in \mathcal{E}_{q_1+q_2}$. Given the observations $\tilde{\mathbf{X}} = (\mathbf{X_1}, \cdots, \mathbf{X_n})$, we have the following result hold.

Theorem 2. Given Assumption 1$'$ and 2, 3 and 4$'$, SBIC (12) can consistently select the true graph structure G_0, i.e., as $n \to \infty$, we have

$$\mathbf{z}_0 = \arg\min_{\mathbf{z}\in\mathcal{Z}_{q_1+q_2}} \mathrm{SBIC}_T(\mathbf{z}) \tag{18}$$

in probability.

The detailed proofs of Theorem 1 and 2 are provided in Appendix B.

5. Simulation Studies

Two examples will be considered in this section. The first example considers a low-dimensional graph in which only six nodes are involved. In such case all the candidate models can be investigated. It is shown that structural BIC can uniformly outperform the standard BIC. In the second example, a graph with 40 nodes is considered. First it is shown that the candidate model pool constructed by two-step algorithm is superior to the model pool constructed by standard lasso in which the same model selection criterion SBIC is used. Then we combine the model pool and model selection criterion together and show that structural BIC outperforms standard BIC and exhibits the robustness to the specification of temperature parameter.

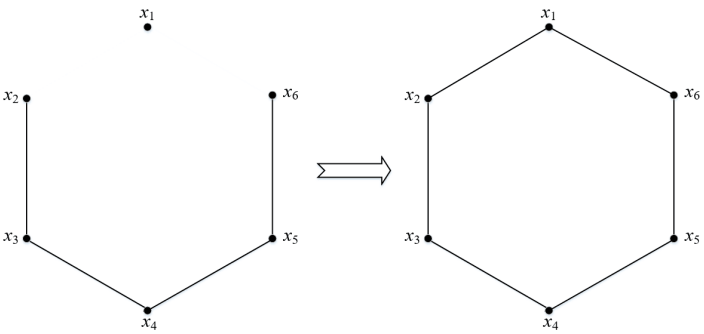

Figure 1. Graphs involved in Example 1. The left one is used as the prior structure while the right one is the true structure.

Example 1. Let us consider a circle with six nodes as shown in Figure 1. Specifically we have $X_i = \alpha X_{i-1} + \epsilon_i$ for $i = 2, \cdots, 5$ and $X_1 = \alpha X_6 + \epsilon_1$. Sample size are set to be $n = 40, 80$. For coefficient we consider the cases of $\alpha = 0.3, 0.4$ and 0.5 respectively with $\epsilon_i \overset{i.i.d.}{\sim} N(0, 1)$. The left graph in Figure 1 is used as prior graph while the right one is the true graph. We use the error rate to determine the temperature parameter T. From Figure 1 the true error rate is $r = 2/15$. In order to evaluate the consequence of misspecification of error rate, we also consider five other choices, $r = 1/15, 3/15, \cdots, 6/15$ from which temperature parameter T can be determined respectively. Two criteria, True positive rate (TPR) and False positive rate (FPR) are employed to compare the performance of SBIC and BIC which are defined as the following,

$$\mathrm{TPR} = \frac{\#\{\text{identified true edges}\}}{\#\{\text{all true edges}\}}, \tag{19}$$

$$\mathrm{FPR} = \frac{\#\{\text{falsely identified edges}\}}{\#\{\text{all none edges}\}}. \tag{20}$$

For each scenario, the replication is set to be $N = 100$ and the resulted TPR and FPR are listed in Table 1. The first number in parentheses is TPR and the second is FPR. It can be seen that, when the error rate is specified correctly, i.e., $r = 2/15$, SBIC outperforms BIC for all the cases considered, either in terms of TPR or FPR. Even for the misspecification cases, in most scenarios considered, SBIC still outperforms standard BIC, especially in terms of FPR. SBIC shows robustness with respect to the misspecification of the expected error rate.

Example 2. Consider a Gaussian graphical model with a tree structure. Specifically, let $X = (X_1, \cdots, X_{40})$ be a random vector with $X_1 \sim N(0, 1)$. For $i = 2, 3, 4$, we have $X_i = \alpha X_1 + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$. For $i = 5, 6, 7$, we have $X_i = \alpha X_2 + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$. For $i = 8, 9, 10$, we have $X_i = \alpha X_3 + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$. In this manner, all the variables can be generated. The structure of X is shown in

Table 1

Performance comparison for BIC and SBIC for the low-dimensional Gaussian graphical model with prior structure information.

		n=40			n=80		
		$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
SBIC	$r = 1/15$	(0.718, 0.002)	(0.785, 0.020)	(0.897, 0.010)	(0.780, 0.004)	(0.912, 0.006)	(0.982, 0.002)
	2/15	(0.742, 0.001)	(0.842, 0.009)	(0.925, 0.018)	(0.823, 0.002)	(0.940, 0.005)	(0.993, 0.004)
	3/15	(0.653, 0.023)	(0.802, 0.033)	(0.915, 0.025)	(0.792, 0.016)	(0.943, 0.011)	(0.993, 0.009)
	4/15	(0.600, 0.024)	(0.817, 0.026)	(0.930, 0.025)	(0.800, 0.014)	(0.943, 0.012)	(0.998, 0.010)
	5/15	(0.540, 0.032)	(0.778, 0.042)	(0.913, 0.044)	(0.778, 0.030)	(0.947, 0.026)	(0.995, 0.013)
	6/15	(0.532, 0.039)	(0.765, 0.063)	(0.901, 0.072)	(0.772, 0.021)	(0.952, 0.026)	(0.990, 0.025)
BIC		(0.493, 0.076)	(0.767, 0.090)	(0.915, 0.075)	(0.735, 0.054)	(0.930, 0.050)	(0.997, 0.032)

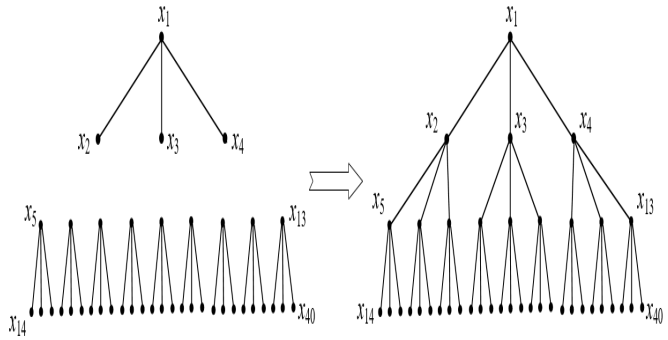

Figure 2. The graphs involved in Example 2. The left one is used as the prior while the right one is the true graphical structure.

Figure 2. The left graph in Figure 2 is used as the prior structure and the right graph is the real structure.

Figure 3 presents the plots for TPR and FPR as a function of α respectively. In each plot two curves are drawn in which the solid one corresponds to model pool constructed from standard lasso and the dashed one corresponds to model pool constructed from two-step algorithm. For both cases structural BIC is employed to select the model in which temperature parameter is set based on $r = 9/780$. Replication is $N = 100$. Sample size is $n = 60$. From the plots it is obvious that TPR from two-step algorithm is higher than TPR from standard lasso while FPR from two-step algorithm is lower than FPR from standard lasso. In particular the difference becomes more prominent when the association among the nodes is weak.

Table 2 lists the results for SBIC and BIC under different scenarios. Specifically, the sample sizes are $n = 50, 100$ and replication is $N = 100$. Three choices of association strength are $\alpha = 0.3, 0.4, 0.5$. As for temperature parameter T, six choices for expected error rate, $r = 3/780, 9/780, 18/780, 27/780, 36/780, 390/780$, are considered. As in Example 1 temperature parameter can be derived from the error rate. Two-step algorithm is used to construct the candidate model pool for these five cases while standard lasso is used for the last row.

From Table 2 it can be seen that the worst cases occur at the combination of BIC and lasso. The best cases occur at the combination of SBIC and two-step algorithm. For the rows with the error rate other than the true value $r = 9/780$, if it is not too far from $r = 9/780$, the results are comparable with the results from $r = 9/780$. For the row of $r = 390/780$ which corresponds to the combination of BIC and two-step algorithm, the results are similar to the last row.

In summary, if prior structure is available for high-dimensional graphical model, then both model selection criterion and candidate model pool should incorporate such information. The results from the proposed procedure demonstrate robustness to the misspecification of the expected error rate.

Table 2
Performance comparison for BIC and SBIC for the high-dimensional Gaussian graphical model with prior structure information.

		n=50			n=100		
		$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$
SBIC	$r = 3/780$	(0.794, 0.0000)	(0.838, 0.0004)	(0.876, 0.0010)	(0.894, 0.0006)	(0.941, 0.0008)	(0.971, 0.0010)
	9/780	(0.809, 0.0004)	(0.844, 0.0009)	(0.885, 0.0013)	(0.912, 0.0009)	(0.953, 0.0008)	(0.974, 0.0010)
	18/780	(0.816, 0.0009)	(0.848, 0.0010)	(0.894, 0.0016)	(0.902, 0.0009)	(0.952, 0.0009)	(0.982, 0.0014)
	27/780	(0.828, 0.0009)	(0.870, 0.0015)	(0.899, 0.0020)	(0.904, 0.0009)	(0.949, 0.0008)	(0.975, 0.0013)
	36/780	(0.827, 0.0012)	(0.870, 0.0016)	(0.902, 0.0027)	(0.902, 0.0008)	(0.951, 0.0011)	(0.979, 0.0013)
	390/780	(0.916, 0.0324)	(0.948, 0.0284)	(0.958, 0.0199)	(0.972, 0.0172)	(0.991, 0.0147)	(0.996, 0.0114)
BIC		(0.664, 0.0276)	(0.836, 0.0262)	(0.907, 0.0262)	(0.932, 0.0283)	(0.981, 0.0241)	(0.993, 0.0135)

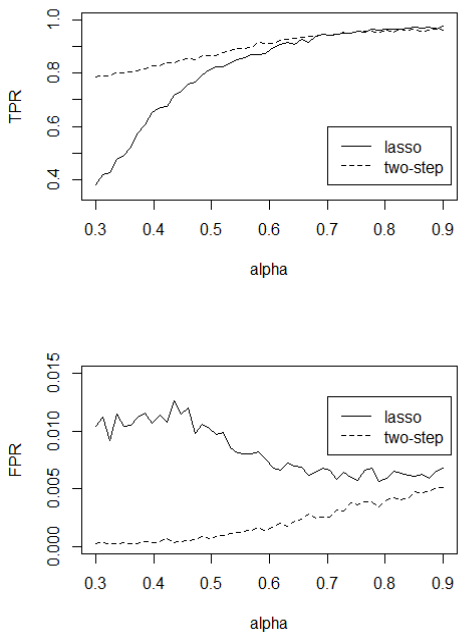

Figure 3. The top plot is the true positive rate versus association strength α and the bottom plot is the false positive rate versus α. The solid lines correspond to the model pool constructed by lasso while the dashed lines correspond to the model pool constructed by two-step algorithm.

6. Metabolite Network for Infant Feces

Metabolites in human body are intrinsically related with different diseases. Understanding the relationship among metabolites are helpful to design appropriate treatment. To this end, multiple methods have been proposed in literature to identify the structure of metabolite network . For example, Gao et al (2015); Karnovsky et al (2012) used the biochemical domain knowledge to construct the metabolite network. Barupal et al (2012); Grapov et al (2015) constructed the network based on structural similarity and mass spectral similarity of metabolites. The metabolite prior network in this paper is constructed based on the method in Gao et al (2015); Karnovsky et al (2012).

The dataset considered here comes from the New Hampshire Birth Cohort Study, an ongoing prospective cohort study of women and their young children, to demonstrate the efficiency of our algorithm. see Madan et al (2016). The dataset was obtained from metabolomics characterizations of stool samples collected from infants at approximately six weeks to one year of age. Sample preparation (with some modifications), ^{1}H NMR data acquisition, and metabolites profiling procedures have been previously described in Brim et al (2017); Sumner et al (2009, 2015); Banerjee et al (2012); Pathmasiri et al (2012). Chenomx NMR Suite 8.4 Professional software (Edmonton, Alberta, Canada) was used to determine relative concentration (Weljie et al (2006)) of selected metabolites from a curation of list of metabolites that are associated with host-microbiome metabolism, see Li et al (2008); Paul

et al (2016). This resulted in a total of 882 observations for 36 metabolites in this data set. All the observations for metabolites were standardized so that they have zero mean and unit standard error, see van den Berg (2006). In the following we consider to identify the network among these metabolites using the algorithms proposed in Section 3.

We use pathway analysis to construct the prior structure. These pathway data are obtained from biological database Kyoto Encyclopedia of Genes and Genomes (KEGG) which provides state-of-the-art information about the metabolites and their pathways. Specifically, each of the targeted metabolites is listed with its associated KEGG Compound ID. Compound information for small molecules in the KEGG database can be retrieved using KEGGREST, a client API written for R (Dan Tenenbaum (2018). KEGGREST: Client-side REST access to KEGG. R package version 1.22.0). Using functions in the KEGGREST library, the database resource was queried in the R language to retrieve the list of one or more pathways associated with each metabolite. With the pathway information in hand, for two given metabolites X_i and X_j, let the pathways associated with X_i and X_j are respectively $Z_i = \{Z_{i1}, \cdots, Z_{im_i}\}$ and $Z_j = \{Z_{j1}, \cdots, Z_{jm_j}\}$. Denote the common pathways of X_i and X_j by $Z_{ij} = Z_i \cap Z_j$ and define

$$s_{ij} = \frac{|Z_{ij}|}{\min\{|Z_i|, |Z_j|\}}.$$

If $s_{ij} \geqslant 0.8$, then X_i and X_j are regarded as associated and there is an edge between them. With threshold equal to 0.8, there are 27 edges among these metabolites. With threshold equal to 0.6, there are 117 edges among these metabolites. We use the difference of the two number as the expected number of edges in difference graph between the prior network and true network which in turn implies that the value of temperature parameter involved in SBIC is $T = 1$. As for the construction of model pool, we set $m_1 = m_2 = 200$ with $\lambda_{\max}/\lambda_{\min} = 0.01$ in (15) and (16), where $\lambda_{\max}$ repsents the minimal λ at which the neighborhood is an empty set. Then based on SBIC (12) and two-step algorithm, we can get the final network. Comparison of the prior network to the final network reveals that there are 153 edges added and 3 edges deleted from the prior network. Figure 4 shows the added edges. The three deleted edges are between (Methionine, Tryptophan), (Glutamate, Histidine), (Asparagine, Valine) respectively.

A primary question here is that whether the edges that are defined by pathway reflect the association between metabolites. If pathway does not contain any information about metabolites, then such prior network can be regarded as built just randomly. Then the probability p_1 that an edge is deleted from and the probability p_2 that an edge is added to the prior network should be equal. Thus we can consider the following hypothesis testing problem, $\mathrm{H}_0 : p_1 = p_2$. The test statistic involved is $U = \frac{\hat{p}_1 - \hat{p}_2}{(\mathrm{var}(\hat{p}_1) + \mathrm{var}(\hat{p}_2))^{1/2}}$ where $\hat{p}_1$ and $\hat{p}_2$ are the maximum likelihood of p_1 and p_2 respectively. In light of central limit theorem, it can be shown that the p-value for the hypothesis above is 0.0234. With such a p-value, we can tentatively assert that pathway have statistically significant effect on the association between metabolites.

One potential concern about the previous analysis is that

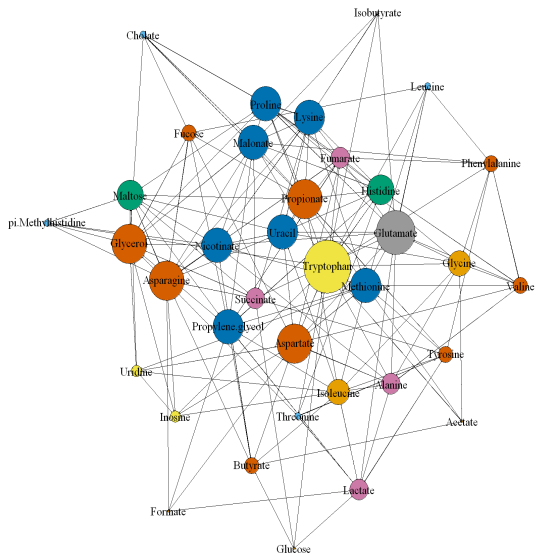

Figure 4. The edges that appear in the final network while were not included in prior network.

the conclusion may be biased by the prior structure. However, we still can use the following method to validate this conclusion. Specifically, we just consider the added edges in Figure 4 which are not involved in prior structure. For any given $0 < s < 0.8$, we construct the prior network E_s by using the same procedure as above, i.e, add an edge for (X_i, X_j) if $s_{ij} \geqslant s$ otherwise not. Note for $s = 0.8$ there are 153 added edge among total 603 edges, apart from the 27 prior edges. Imagine that if pathways have no impact on the association of metabolites, then the proportion of 153 added edges in E_s should be the same as for $s = 0.8$, i.e., $p_0 = \frac{153}{603} = 0.2537$. Define p_s the probability of the edges in Figure 4 falling into E_s, then the null hypothesis is $\mathrm{H}_0 : p_s = p_0$. For $s = 0.1, 0.2, \cdots, 0.6$, the estimate $\hat{p}_s$ can be shown to be $(0.2578, 0.2596, 0.2800, 0.3300, 0.3630, 0.4111)$ and the corresponding p-value for the hypothesis H_0 are $(0.3959, 0.3658, 0.1287, 0.0059, 0.0011, 0.0003)$. Based these results, we can say that pathway is statistically significant factor on the association of metabolite. The possibility of association will increase as the threshold s increases. Figure 5 depicts the empirical probability of association as a function of threshold.

It should be stressed that the discussion above does not mean that prior network must have to share some common information with the data. If a prior network is theoretically sound, such prior network is also feasible. However, if a prior network can find the support from both the theory and data, in our view, it is more advantageous than the one with support just from theory or subjective belief.

We have confirmed that part of the association among metabolites can be attributed to pathway. The next question

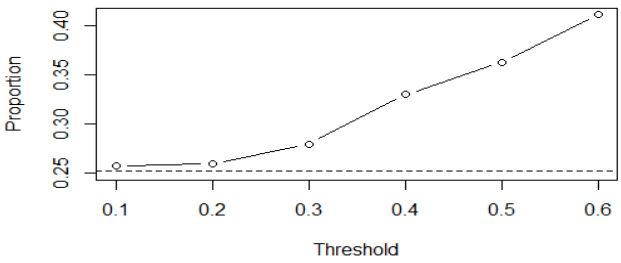

Figure 5. The proportion of the added edges in prior network E_s as a function of threshold s. The bottom dashed line corresponds to the line under the null hypothesis.

Table 3
Edges in Figure 4 that can not be covered by standard pathway analysis

Malonate	Asparagine, Cholate, Isobutyrate, Tryptophan, Phenylalanine, Propionate, Succinate, Lysine
Propylene glycol	Butyrate, Formate, Methionine, Fumarate, Histidine, Isoleucine, Maltose, Fucose
π.Methylhistidine	Asparagine, Maltose, Nicotinate, Tryptophan

we aim to address is whether all associations among metabolites can be explained solely by pathway? To try to answer this question, first we define a more inclusive prior structure among metabolites based on pathway. Specifically, whenever two metabolites have any pathway in common, then there is an edge between them and no edge otherwise. By comparing the network in Figure 4 to this prior structure, we found that there are 20 edges which are not covered by the prior structure. In other words, pathway analysis cannot cover all the relationships among metabolites.

These 20 edges are listed in Table 3. Among these 20 edges, 8 edges are related with malonate, 8 edges are related with propylene glycol and 4 with π-Methylhistidine. Malonate is a well-known competitive inhibitor of succinate dehydrogenase (SDH) while SDH is a complex of four polypeptides (SDH A–D) that catalyzes the conversion of succinate to fumarate and functions in mitochondrial energy generation, oxygen sensing and tumor suppression. Propylene glycol is a widely used drug vehicle with serious side effects reported in clinical studies and recognized toxicity, see Morshed et al (1998, 1994). In light of these existing studies, it is not surprising to find their wide connection with other metabolites even they do not share any pathway.

In summary, metabolic pathways can explain most of the connections among the metabolites but not completely. This may be explained by the fact that conventional metabolic pathway datasets only focus on the endogenous reactions occurring within the cell. It is possible that some important reactions may be omitted by conventional pathway analysis. However, by appropriately combining prior knowledge with empirical data analysis, the proposed method can discovered these reactions in a more comprehensive way.

7. Conclusion

We have developed a novel method to select the high-dimensional Gaussian graphical model with the aid of prior structure. Such prior structure is often the result of biological knowledge. The algorithm consists of two parts. In the first part we proposed a model selection criterion called structural BIC which can be regarded as a generalization of the widely used extended BIC. In second part, we propose a two-step algorithm to construct the candidate model pool which incorporates the prior structure during the construction. It is proved that under the given assumptions the structural BIC is a consistent model selection criterion. Simulation results validate the efficacy and robustness of the algorithm.

We applied the proposed algorithm to the metabolite data from infant feces for which the prior network is constructed through the pathways shared by metabolites. It is shown that pathway is a statistically significant factor for the association of metabolites. As the network based on the pathway analysis have been widely used in many fields, these findings provide statistical basis for such practice. We also found new relationships among metabolites that have been omitted by conventional pathway analysis in which most of them is related two well-known important metabolites.

It is possible to use the proposed algorithm analyzing other types of prior network for metabolites, e.g, the structural similarity based prior network. More generally, other biological network such as gene regulation network or microbial interaction network if the related prior structure is available. The algorithm can be easily adapted for the binary data such as Ising model. It is known that model selection with prior structure for Ising model is complex and little work has been done in this respect. Our method provides a possible solution to this issue and deserves further investigation in the future.

Acknowledgements

We are grateful to the participants and staff of the New Hampshire Birth Cohort Study for providing the processed metabolomics data. This work is supported in part by US National Institutes of Health grants (R01LM012723; P20ES018175; P01ES022832; UG3OD023275), US Environmental Protection Agency grant (RD83459901), the Children's Center grant, the Superfund grant and COBRE.

Supplementary Materials

The supplementary materials for this article are available in Appendix A-C, which include the marginal distribution of a homogeneous Boltzmann distribution, the proofs for theorem 1 and 2 and the prior structure for metabolomic data in section 5.

References

Banerjee R, Pathmasiri W, Snyder R, McRitchie S, Sumner S. Metabolomics of brain and reproductive organs: characterizing the impact of gestational exposure to butylbenzyl phthalate on dams and resultant offspring Metabolomics. 2012. doi: 10.1007/s11306-011-0396-y.

Barupal K D, Haldiya K P, Wohlgemuth G, Kind T, Kothari S L, Pinkerton E K, Fiehn O (2012). MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. Bioinformatics, 13(99), 1–15.

Bogdan M., Ghosh J K., Doerge R W. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics, 167(2): 989–99.

Boluki S., Esfahani M S., Qian X., Dougherty E R. Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors. Bioinformatics, 18(14): 61–80.

Brim, H., S. Yooseph, E. Lee, Z. A. Sherif, M. Abbas, A. O. Laiyemo, S. Varma, M. Torralba, S. E. Dowd, K. E. Nelson, W. Pathmasiri, S. Sumner, W. de Vos, Q. Liang, J. Yu, E. Zoetendal and H. Ashktorab (2017). A Microbiomic Analysis in African Americans with Colonic Lesions Reveals Streptococcus sp.VT162 as a Marker of Neoplastic Transformation, Genes (Basel) 8(11).

Chen I., Yogeshwar D. Kelkar., Yu Gu., Jie Zhou., Xing Qiu., Hulin Wu. (2017) High-dimensional linear state space models for dynamic microbial interaction networks. PlOS ONE, 15: 1-20.

Chen J and Chen Z. (2008). Extended Bayesian information criterion for model selection with larger model space. Biometrika, 94, 759-771.

Chen J and Chen Z. (2012). Extended BIC for small-n-large-p sparse GLM. Statistics Sinica, 22, 555-574.

Cheng J., Levina E., Wang P., Zhu J. (2014) Sparse Ising model with covariates. Biometrics, 70, 943-953.

Christine Peterson, Francesco Stingo., Marina Vannucci (2015) Bayesian Inference of Multiple Gaussian Graphical Models, Journal of the American Statistical Association, 110(509), 159–174.

Friedman J., Hastie T., Tibshirani R. (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistcs, 9, 432-441.

Friedman J., Hastie T., Tibshirani R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw, 33(1): 1–22.

Foygel Rina., Drton Mathias. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models, NIPS.

Gao J, Tarcea V G, Karnovsky A, Mirel B R, Weymouth T E, Beecher C W, Cavalcoli J D, Athey B D, Omenn G S, Burant C F, Jagadish H V (2010). Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics, 26(7): 971-3. doi: 10.1093/bioinformatics/ btq048.

Grapov D, Wanichthanarak K, Fiehn O. MetaMapR: pathway independent metabolomic network analysis incorporating unknowns. Bioinformatics, 31(16):2757-60. doi: 10.1093/bioinformatics/btv194.

Ideker T., Dutkowski J., Hood L. (2011) Boosting signal-to-noise in complex biology, prior knowledge is power. Cell, 144(6): 860–863.

Imoto S., Higuchi T., Goto T., Tashiro K., Kuhara S., Miyano S. (2004) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003.

Ippolito, Joseph E., Matthew E. Merritt, Fredrik Bäckhed, Krista L. Moulder, Steven Mennerick, Jill K. Manchester, Seth T. Gammon, David Piwnica-Worms, and Jeffrey I. Gordon. Linkage between cellular communications, energy utilization, and proliferation in metastatic neuroendocrine cancers. Proceedings of the National Academy of Sciences 103, no. 33 (2006): 12505-12510.

Karnovsky A, Weymouth T, Hull T, Tarcea V G, Scardoni G, Laudanna C, Sartor M A, Stringer K A, Jagadish H V, Burant C, Athey B, Omenn G S. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics, 28(3): 373–80. doi: 10.1093/bioinformatics/btr661.

Lauritzen S L. (1996). Graphical Models. Oxfor University Press.

Li M, Wang B, Zhang M, Rantalainen M, Wang S, Zhou H, Zhang Y, Shen J, Pang X, Zhang M, Wei H, Chen Y, Lu H, Zuo J, Su M, Qiu Y, Jia W, Xiao C, Smith LM, Yang S, Holmes E, Tang H, Zhao G, Nicholson JK, Li L, Zhao L. (2008). Symbiotic gut microbes modulate human metabolic phenotypes. Proc Natl Acad Sci, 105(6):2117-22.

Li Fan and Zhang R (2010) Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics, Journal of the American Statistical Association, 105(491), 1202–1214.

Madan JC, Hoen AG, Lundgren SN, Farzan SF, Cottingham KL, Morrison HG, Sogin ML, Li H, Moore JH, Karagas MR. (2016) Association of Cesarean Delivery and Formula Supplementation With the Intestinal Microbiome of 6-Week-Old Infants. JAMA Pediatr, 170(3):212-9.

Marino S, Baxter NT, Huffnagle GB, Petrosino JF, Schloss PD. (2014) Mathematical modeling of primary succession of murine intestinal microbiota. Proceedings of the National Academy of Sciences, 111 (1): 439–444.

Ma J. (2015) Estimation and Inference for High-Dimension Gaussian Graphical Models with Structural Constraints. PhD Dissertation, University of Michigan.

Meinshansen N., P Bühlmann. (2006). High dimensional graphs and variable selection with lasso. The annals of statistics, 34(3), 1436–1462.

Meier L., Geer S and Bühlmann P. (2008). The group lasso for logistic regression. J. R. Statist. Soc. B., 70, 53-71.

Morshed K M., Jain K S., McMartin E K. (1998) Propylene Glycol-Mediated Cell Injury in a Primary Culture of Human Proximal Tubule Cells. Toxicological Sciences, 46, 410–417.

Morshed K M., Jain K S., McMartin E K. (1994) . Acute toxicity of propylene glycol: an assessment using cultured proximal tubule cells of human origin. Fundam. Appl. Toxicol. 23(1), 38-43.

Pathmasiri W, Pratt K J, Collier D N, Lutes L D, McRitchie S, Sumner S C J. (2012) Integrating metabolomic signatures and psychosocial parameters in responsivity to an immersion treatment model for adolescent obesity. Metabolomics. 2012;8(6):1037-51. doi: 10.1007/s11306-012-0404-x.

Paul, H. A., M. R. Bomhof, H. J. Vogel and R. A. Reimer (2016). Diet-induced changes in maternal gut microbiota and metabolomic profiles influence programming of offspring obesity risk in rats. Sci Rep 6: 20683.

Ravikumar P., Wainwright M J. and Lafferty J D. (2010). High-dimensional Ising model selection using L_1 regularized logistic regression. Annals of Statistics, 38, 1287-1319.

Roach J C, Glusman G, Smit A F, Huff C D, Hubley R, Shannon P T, Rowen L, Pant K P, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde L B, Hood L., Galas D J. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing, Science, 328(5978):636-9.

Roverato A. (2002). Hyper Inverse Wishart Distribution for Non- Decomposable Graphs and its Application to Bayesian Inference for Gaussian Graphical Models, Scandinavian Journal of Statistics, 29, 391–411.

Segre A., Groop L., Mootha V., Daly M., Altshuler D. (2010) Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet 6 .

Siegmund D. (2004). Model selection in irregular problems: Application to mapping quantitative trait loci. Biometrika, 91, 785–800.

Sumner S, Snyder R, Wingard C, Mortensen N, Holland N, Shannahan J H, et al. (2015) Distribution and biomarkers of carbon-14-labeled fullerene C ([C(U)]C) in female rats and mice for up to 30 days after intravenous exposure. Journal of applied toxicology : JAT. 2015. Epub 2015/03/03. doi: 10.1002/jat.3110. PubMed PMID: 25727383.

Sumner S, Snyder R, Burgess J, Myers C, Tyl R, Sloan C, et al. (2009) Metabolomics in the assessment of chemical-induced reproductive and developmental outcomes using non-invasive biological fluids: application to the study of butylbenzyl phthalate. Journal of applied toxicology : JAT. 2009;29(8):703-14. Epub 2009/09/05. doi: 10.1002/jat.1462. PubMed PMID: 19731247.

Tibshirani, Ryan J., Taylor, Jonathan. (2011) The solution path of the generalized lasso. Ann. Statist. 39 (3): 1335–1371.

van den Berg, R. A., H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde and M. J. van der Werf (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7: 142.

Wainwright M J. and Jordan M I. (2003). Graphical models, exponential families and variational inference. Technical Report 649, Dept. Statistics, Univ. California, Berkeley. MR2082153.

Weljie, A. M., J. Newton, P. Mercier, E. Carlson and C. M. Slupsky (2006). Targeted profiling: quantitative analysis of 1H NMR metabolomics data. Anal Chem 78(13): 4430-4442.