bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Jie Zhou[a], Anne G. Hoen[a,b], Susan McRitchie[c], Wimal Pathmasiri[c], Weston D. Viles[d], Quang P. Nguyen[a,b], Juliette C. Madan[b], Erika Dade[b], Margaret R. Karagas[b], Jiang Gui[a,*]

[a] *Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, U.S.A*
[b] *Depatment of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, NH, U.S.A*
[c] *Nutrition Research Institute, University of North Carolina, Kannapolis, NC, U.S.A.*
[d] *Department of Mathematics and Statistics, University of Southern Maine, Portland, Maine, U.S.A*

Abstract

In light of the low signal-to-noise nature of many large biological data sets, we propose a novel method to learn the structure of association networks using Gaussian graphical models combined with prior knowledge. Our strategy includes 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 Bayesian information criterion (BIC). It is shown that the popular extended BIC (EBIC) is a special case of SBIC. In the 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 high-dimensional Gaussian graphical model. Simulation studies validate the superiority of the proposed algorithm over the existing ones and show the robustness to the model misspecification. Application to relative concentration data from infant feces collected from subjects enrolled in a large molecular epidemiological cohort study validates that metabolic pathway involvement is a statistically significant factor for the conditional dependence between metabolites. Furthermore, new relationships among metabolites are discovered which can not be identified by the conventional methods of pathway analysis. Some of them have been widely recognized in biological literature.

[*] Corresponding author
Email address: Jiang.Gui@dartmouth.edu (Jiang Gui)

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 signal-to-noise ratio making detection of effect challenging and final conclusions unreliable. As reported in Ideker et al [31], prior knowledge can play a pivotal role in deciphering this kind of complexity. For example, Segre et al [67] 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 [64] 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 [7], Imoto et al [32], Ma et al [48]. The studies in this paper are motivated by the metabolite pathway information that are available in many public biological databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG). As far as we know, such information have not been well utilized in literature to improve the statistical analysis of metabolites.

Biological network, such as microbe-microbe interaction networks, metabolite association networks and gene regulation networks, have received much attention in recent years. Probabilistic graphical models are typically employed in literature to study such biological networks (Lauritzen and Sheehan [42], Zhang et al [86], Stingo et al [72], Dobra et al [19]). The edges in graphical model represent the conditional dependence among the vertices, which stand for the objects of interest such as microbes, metabolites or genes. In some cases, i.e., causal inference, the directionality of edge, which is indispensable for the computation of casual effects, also has to be considered. Nevertheless, in this paper we only consider association network which can be characterized by undirected graphical model. The identification of the structure of association network is our primary interest. Many algorithms have been proposed in literature in this respect. In particular, for tree and forest, Chow and Liu [17], Edwards et al [21] and Kirshner et al [38] proposed and studied the classic Chow-Liu algorithm and its extensions. Heuristic algorithms such as the hill-climbing algorithm are studied in Hojsgaard et al [30], Jalali et al [36], Lauritzen [41] and Ray et al [62]. Cheng et al [15], Friedman et al [22], Meinshansen and Buhlmann [49], Ravikumar et al [61] and Wainwright and Jordan [79] investigated the L_1-penalized likelihood method for the identification of Gaussian and Ising graphical models.

In order to deal with the prior structure of network, Bayesian methods is the typical choice in literature (Dobra et al [19], Jones et al [35], Scott and Berger [66]). As for Gaussian graphical model (GGM), the most popular Bayesian method is based on the direct modeling of prior distribution of precision matrix, e.g, conjugate G-wishart distribution (Mohammadi and Wit [52]), or spike-slab distribution (Mohammadi [53]) et al. In these situations, though there exist MCMC sampling algorithms for the decomposable graph, the computation becomes challenging for the general non-decomposable graph (Carvalho et al [11], Dobra and Lenkoski [20], Mitsakakis et al [51], Roverato [65], Wang

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[80], Wang and Carvalho [81]). Recently, an efficient sampling-free Bayesian method is proposed in Leday and Richardson [43] for high-dimensional GGM , which employs hypothesis testing to determine the existence of each individual edge based on Bayes factor. There are also several algorithms to deal with the prior information within frequentist framework. For example, the Chow-Liu algorithm can learn the structure when the graph is a tree (Chow and Liu [17]); Chow-Liu algorithm is extended in Edwards et al [21] to deal with the forest; conditional Chow-Liu algorithm is proposed in Kirshner et al [38] to include a given set of edges in the graph. In extended Bayesian information criterion (EBIC) for high-dimensional Gaussian graphical model (Foygel and Drton [26]), as will be shown in Section 3, empty graph is adopted as the prior graph. If a sufficient large weight is assigned to the prior structure, the algorithm will end up with an empty graph. Ma et al [48] focus on the deterministic prior structure; when the random structure is involved in prior graph, they propose an algorithm to construct the model pool.

In this paper we propose a novel method to learn the structure of Gaussian graphical models based on a given prior structure. We first propose a model selection criterion called structural Bayesian information criterion (SBIC) to incorporates the prior structure into BIC. Numerous criteria have been proposed in literature for model selection, e.g., Akaike information criterion (AIC, Akaike [1]), Bayesian information criterion (BIC, Schwarz [69]), extended BIC (EBIC, Bogdan et al [4, 5, 6], Chen and Chen [13, 14]), cross-validation method (CV, Stone [75]), generalized CV method (GCV, Geisser [28], Burman [10], Shao [70], Zhang [83]), risk inflation criterion (RIC, Foster and George [25], Zhang and Shen [85]) among many others. In particular, generalized information criterion (GIC) proposed in Shao [71], Kim et al [37] provided a unified framework for many of these criteria in the context of linear regression model, such as AIC, BIC, EBIC and RIC et al. As for the network selection, Foygel and Drton [26] studied the consistency of EBIC for GGM selection. The criterion SBIC proposed in this paper can be regarded as a generalization of the EBIC of Foygel and Drton [26]. Compared with the EBIC, SBIC provides a more flexible framework; in fact, SBIC just reduces to EBIC when the prior graph is an empty graph. As a theoretical basis, for high-dimensional sparse Gaussian graphical models, it is shown that SBIC is a consistent criterion for model selection under mild conditions. Based on the prior structure, we then propose a data-driven two-step algorithm to build the model pool, in which the graph is enriched in the first step and pruned in the second step. Such two-step algorithm can be readily implemented based on the R packages such as *glmnet* (Friedman et al [22]) or *glasso* (Friedman et al [24]). Recall that the well-known greedy equivalence search (GES) algorithm for structure learning of directed acyclic graph also consists of similar edge addition and removal steps aiming to optimize the score function. For decomposable graph, GES algorithm converges to the global optimum in probability as $n \to \infty$ (Chickering et al [16]).

Through simulation studies, it is shown that the combination of the proposed SBIC and two-step algorithm is a robust structure-learning strategy for high-dimensional Gaussian graphical model and outperforms many existing popular

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

structure learning algorithms under the given conditions. As an application, we studied ^{1}H NMR-based metabolite data profiled in infant feces collected as part of the New Hampshire Birth Cohort Study (NHBCS), a large prospective cohort study of mothers and their children born in New Hampshire (Madan et al [46]). The prior structure for these metabolites is constructed based on the related pathway information from KEGG. Our results show that pathway is a statistically significant factor for the conditional dependence between metabolites, i.e., the strength of conditional dependence between two metabolites increases if the proportion of shared pathways increases. Furthermore, the identified network discovers new relationship between metabolites that can not be identified through the conventional methods of pathway analysis, many of which have been validated in biological literature.

The paper is organized as follows. Section 2 briefly reviews the undirected Gaussian graphical model and extended BIC. A new formulation of extended BIC is introduced. In Section 3, we present our main algorithm, in which Section 3.1 introduces structural BIC and its implications; Section 3.2 describes the two-step algorithm for building the candidate model pool. Theoretical investigation of SBIC is given in Section 4. In Section 5, the algorithm is evaluated through simulated data. In Section 6 we use the algorithm to investigate the pathway and metabolomic data from NHBCS. Section 7 concludes with a brief comment.

2. Gaussian Graphical Model and EBIC

2.1. A brief review of EBIC for Gaussian graphical model

For a given p-dimensional multi-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 $G = (V, E)$ is used to represent the conditional dependence relationship between $\mathbf{X}$, where the vertex set V indexes the variables and the edge set E encodes the conditional independence. The precision matrix is defined as $\Omega_{p\times p} = (\omega_{ij}) = \Sigma^{-1}$. It turns out that for multi-normal distribution the precision matrix can completely specify the structure of G. Given n i.i.d. observations, $\tilde{X} = (\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)})^T$, our aim is to learn the structure of G, i.e., to identify the nonzero components in $\tilde{p} \overset{\triangle}{=} p(p-1)/2$ off-diagonal entries in Ω. In its general form, Bayesian information criterion (BIC) can be stated as follows. Let $\mathcal{E}$ be the model space under consideration with $\pi(E)$ the prior probability for $E \in \mathcal{E}$. Let θ denote the unknown parameter in E with prior distribution $p(\theta)$. With θ in hand, let the conditional density function for $\tilde{X}$ be $f(\tilde{X}|\theta)$, then the marginal density function for observations $\tilde{X}$ can be expressed as $f(\tilde{X}|E) = \int f(\tilde{X}|\theta)p(\theta|E)d\theta$. Therefore, the posterior distribution of model E can be expressed as

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

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

$$\begin{aligned} -2\log p(E|\tilde{X}) = -2\log f(\tilde{X}|\theta) + |E|\log n - |E|\log(2\pi) - 2\log p(\theta|E) \quad (2) \\ + \log\det(M) - 2\log\pi(E) + c, \end{aligned}$$

where M is the expected information matrix for single observation, $|E|$ the degree of freedom of model E and c a constant. By omitting the last five terms which do not involve the sample size n, we get the standard Bayesian information criterion, $\text{BIC}(E) = -2l_n(E) + |E|\log n$ with $l_n(E) = \log f(\tilde{X}|\theta)$. For the high-dimensional regression model, the extended BIC (EBIC) is proposed in Bogdan et al [4], Bogdan et al [5], Bogdan et al [6], Chen and Chen [13], Chen and Chen [14], which puts more weight on sparse model than standard BIC. EBIC is further generalized to the Gaussian graphical model in Foygel and Drton [26], which has the following form,

$$\text{EBIC}_\lambda = -2l_n(\Omega(E)) + |E|\log n + 4|E|\lambda\log p, \tag{3}$$

where $\Omega(E)$ is the precision matrix associated with model E. Tuning parameter $0 \le \lambda \le 1$ controls the model complexity. When $\lambda = 0$, EBIC reduces to the standard BIC. As λ increases, (3) will put more weight on the sparse model. The log-likelihood function $l_n(\Omega(E))$ in (3) for the Gaussian graphical model has the following form,

$$l_n(\Omega) = \frac{n}{2}[\log\det(\Omega) - \text{trace}(S\Omega)], \tag{4}$$

where S is the empirical covariance matrix. It is proved in Foygel and Drton [26] that EBIC_λ (3) is a consistent model selection criterion for high-dimensional GGM under some mild conditions.

Although EBIC has been widely used in literature for high-dimensional model selection, several limitations have not been addressed adequately. In particular, EBIC does not take prior information into consideration. Typically, prior information is integrated with data through Bayesian method; however, as we have mentioned in previous section, the complicated forms of posterior distribution for non-decomposable graphs have undermined the popularity of Bayesian method in practice. In this context, incorporating prior information into BIC is a natural choice. On the other hand, given such model selection criterion, an appropriate candidate model pool is indispensable since the exhaustive search within the whole model space is impossible for high-dimensional graphical model. With these motivations in mind, in the following section we propose a new strategy for the selection of Gaussian graphical model when the prior structure is available. Though we focus on Gaussian graphical model in this paper, the algorithm can be easily adapted to accommodate more general undirected graphical model, e.g., Ising model.

2.2. A new interpretation of EBIC

In this section, we introduce a different way to interpret EBIC which will facilitate the introduction of prior structure in Section 3. For any given pair of

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

vertices, (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 $Z_{ij} = Z_{ji}$ $(1 \leq i < j \leq p)$. Then we define a $\tilde{p}$-dimensional random vector $\mathbf{Z} = (Z_{12}, Z_{13}, \cdots, Z_{(p-1)p/2})^T \triangleq (Z_1, \cdots, Z_{\tilde{p}})^T$. The prior information about the structure of E can be completely specified by the probability distribution of $\mathbf{Z}$. The following Boltzmann distribution is employed to model the distribution of $\mathbf{Z}$,

$$\Pr(\mathbf{Z} = \mathbf{z}) \propto \exp\left(-\frac{\epsilon(\mathbf{z})}{KT}\right), \tag{5}$$

where $\epsilon(\cdot) \geq 0$ is called energy function, T the temperature parameter, and K the Boltzmann constant. Note that there is a one-to-one correspondence between $\mathbf{Z} = \mathbf{z}$ and model E so that (5) amounts to defining a prior distribution $\pi(E)$ for E. Without loss of generality, $K = 2$ is always assumed in the following discussion. Substitution of (5) into (2) yields the following approximation to (2) for Gaussian graphical model,

$$\mathrm{BIC}_{T,\epsilon}(\mathbf{z}) = -2l_n(\Omega(\mathbf{z})) + |\mathbf{z}| \log n + \epsilon(\mathbf{z})/T, \tag{6}$$

where $|\mathbf{z}|$ denotes the number of nonzero components in $\mathbf{z}$. The third to fifth terms and the constant c in (2) have been omitted here since neither sample size n nor model dimensionality p is involved in these terms. In order to use (6) in practice, we consider the following simple yet flexible quadratic specification of $\epsilon(\cdot)$,

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

for some given positive semi-definite matrix W. In (7), the energy function can be regarded as the squared weighted Euclidean distance between the given state, $\mathbf{z}$, and the origin state, $\mathbf{0}$. It is obvious that both BIC and EBIC are the special cases of (7). In fact, if $W = 0$, (7) is the standard BIC; if $T = 1/(4\lambda)$ and $W = (\log p)\mathrm{I}_{\tilde{\mathrm{p}}}$ with $\mathrm{I}_{\tilde{\mathrm{p}}}$ the $\tilde{p} \times \tilde{p}$ identity matrix, then (7) reduces to the EBIC (3)-(4). 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 interpretation can guide us to choose a proper tuning parameter λ involved in EBIC (3). For example, for $\lambda = 0.5$, or equivalently $T = 0.5$, which is often recommended in literature (Foygel and Drton [26]), it implies that the prior mean number of total edges is $\tilde{p}/(1+p) \approx (p-1)/2$. More generally, we have $P(Z_i = 1) < 0.5$ whenever $T > 0$ and $P(Z_i) > 0.5$ whenever $T < 0$. So for the sparse graph with $E|\mathbf{Z}| < \tilde{p}/2$, $T > 0$ is a plausible choice.

In some circumstances, prior information comprise both the mean and variance of the total edges, which can also be modeled through $\mathrm{BIC}_{T,W}$. Specifically, consider the following form of W for $\mathrm{BIC}_{T,W}$,

$$W(\rho) = D^T R(\rho) D \tag{8}$$

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in which $D = \text{diag}(\sqrt{\log p}, \cdots, \sqrt{\log p})$ and $R = \rho J_{\tilde{p}} + (1-\rho) I_{\tilde{p}}$ for some $0 \leq \rho < 1$. 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 edges. The details about the formulas are given in Supplementary Materials. So for any specification of (μ, σ^2), the corresponding parameter (T, ρ) can be easily determined which in turn can be used in $\text{BIC}_{T,W}$ for model selection.

3. Incorporation of Prior Structure into Model Selection

3.1. Prior structure enhanced BIC for Gaussian graphical model

In Section 2.2, it is shown that the third term in EBIC is the squared distance between a given state $\mathbf{z}$ and the origin state $\mathbf{0}$, or in other words, the squared distance between the given graph and the empty graph. Now let us adapt $\text{BIC}_{T,W}$ (7) to accommodate the prior structure of graph. For $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, suppose that graph $\tilde{G} = (V, \tilde{E})$ represents the prior structure (e.g., constructed based on some biological theory), and our aim is to learn the underlying true structure based on $\tilde{G}$ and the observations on $\mathbf{X}$. First, we intruduce the concept of difference graph. For two graphs $\tilde{G} = (V, \tilde{V})$ and $G = (V, E)$, the difference graph of $\tilde{G}$ and G is defined as the graph which has the same vertex set V as $\tilde{G}$ and G while the edge set is $\bar{E} = \tilde{E} \triangle E$. Here $\triangle$ stands for the symmetrical difference operator between two sets. Let the difference graph denoted by $\bar{G} = \tilde{G} \triangle G \overset{\triangle}{=} (V, \bar{E})$. For a given prior edge set $\tilde{E}$, there is a one-to-one correspondence between $\bar{E}$ and E. Equivalently, there is a one-to-one correspondence between their edge vector, $\bar{\mathbf{Z}} = \text{I}(\tilde{\mathbf{Z}} - \mathbf{Z})$ and $\mathbf{Z}$. Replacing $\mathbf{z}$ in the third term of $\text{BIC}_{T,W}$ with $\bar{\mathbf{z}}$, we obtain the following structural Bayesian information criterion (SBIC),

$$\text{SBIC}_{T,W}(\mathbf{z}) = -2 l_n(\Omega(\mathbf{z})) + |\mathbf{z}| \log n + \bar{\mathbf{z}}^T W \bar{\mathbf{z}} / T, \tag{9}$$

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 (9) will lead to solutions that achieve balance between these terms. Essentially, we have assumed that $\bar{\mathbf{Z}}$ in (9) has Boltzmann distribution,

$$P(\bar{\mathbf{Z}} = \bar{\mathbf{z}}) \propto \exp\left(-\frac{\bar{\mathbf{z}}^T W \bar{\mathbf{z}}}{2T}\right). \tag{10}$$

If we set $W = \text{diag}(\log p, \cdots, \log p)$ just like EBIC, then (9) reduces to

$$\text{SBIC}_T(\mathbf{z}) = -2 l_n(\Omega(\mathbf{z})) + |\mathbf{z}| \log n + |\bar{\mathbf{z}}| \log p / T, \tag{11}$$

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

Remarks. (i) If $\tilde{\mathbf{z}} = 0$, i.e., the prior structure is empty graph, then SBIC in (11) reduces to EBIC in (3), i.e., EBIC is a special case of SBIC. (ii) If T is large enough, then the model selected by SBIC is the same as that selected

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

by standard BIC; if T is small enough, then the model selected by SBIC is just the prior graph. For other T, the model selected by SBIC will be a compromise of these two extreme cases. (iii) The choice of T in (11) relies on the expected error rate of prior structure. The expected error rate is defined as $r = E|\bar{\mathbf{Z}}_0|/\tilde{p}$, where $\bar{\mathbf{Z}}_0$ is the edge vector of the difference graph between true and prior structure. It is straightforward to show that $r = \frac{E(m_1)+E(m_2)}{\tilde{p}}$, where m_1 is the number of true edges that have been omitted by prior structure while m_2 is the number of edges that have been mistakenly included in the prior structure. In many cases, r is more intuitive than T. On the other hand, we have from (10), $E|\bar{\mathbf{Z}}|_0 = \frac{\tilde{p}}{1+p^{1/(2T)}}$, from which we have $T = \frac{\log p}{2\log(1/r-1)}$. Given such one-to-one correspondence between T and r, the intuitive interpretation of r can guide us to determine the appropriate value for T. Particularly, if $r = 0.5$, then $T = \infty$, which yields the standard BIC. (iv) The knowledge about the expected error rate r may be derived from domain knowledge, as we did for the metabolite data in Section 6. In case that such information is not available, Bogdan et al [4, 5, 6] recommended a tuning parameter in the context of regression model with which the family wise error rate (FWER) is shown to be approximately 8% for the dataset with the sample size $n \geq 200$ and the number of variables $p \geq 30$. Similar strategy can also be employed in the context of GGM model to control the FWER when no prior information about r is available.

The generalization of (11) is possible. For example, it has been implicitly assumed that the probability of adding an edge to the prior graph, p_1, and the probability of removing an edge from the prior graph, p_2, is equal. In some cases, compared with pruning edges, we may be more inclined to add edges to the prior graph, i.e., $p_1 > p_2$. The following simple generalization of (11) can accommodate such situation,

$$\begin{aligned} \mathrm{SBIC}_{T_1,T_2}(\bar{\mathbf{z}}) &= -2l_n(\Omega(\mathbf{z})) + |\mathbf{z}|\log n + |\bar{\mathbf{z}}_{\mathbf{1}}|\log p/T_1 \\ &\quad +|\bar{\mathbf{z}}_{\mathbf{2}}|\log p/T_2, \end{aligned} \tag{12}$$

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. Different combination of T_1 and T_2 reflects our different belief about the prior structure. For example, a small T_1 and a large T_2 indicate that we have higher confidence about the edges than the non-edges in prior structure. The cost of such flexibility is that we have to specify the values for both T_1 and T_2 which may be challenging in some circumstances.

As has been mentioned in Section 1, there are already several different ways to model the prior structure of network in Bayesian methods, see, e.g., Mohammadi and Wit [52], Mohammadi [53], Mukherjee and Speed [56]. In particular, the concordance function proposed in Mukherjee and Speed [56] plays the similar role as the energy function (5) in this paper. They demonstrated how different type of prior structures can be integrated into the prior distribution of the network. The network structure can then be inferred based on the samples from the posterior distribution. Note that the size of the model space for network grows super-exponentially as the number of vertices increases. Given this fact, Leday and Richardson [43] pointed out that the MCMC-based strategy

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for model selection of network can hardly get sufficient qualified samples that can represent the real posterior distribution, which will eventually compromise the reliability of the final estimates of network. The proposed model selection criterion (9) combined with the candidate model pool detailed in the next section provides an alternative yet feasible way to deal with these problems and demonstrates its superiority in the simulation studies compared with Bayesian network selection.

3.2. Construction of model pool based on prior structure

With p variables, the size of the model space for graphical model is $2^{\tilde{p}}$. If $p = 10$, then the size of model space is 2^{45}. So as p increases, it quickly becomes intractable to find the optimal model by searching the whole model space. There are two ways in literature to deal with this problem. One is to use the heuristic algorithm, including greedy or stepwise forward/backward search, to find the optimum of score function (Chickering et al [16]). The other is to select a subset of the model space as the candidate model pool and then use the model in this model pool which minimize the score function as our selected model (Friedman et al [22]). In this paper, we focus on the second method. A common practice for the construction of candidate model pool for graphical model is to use the solution path of graphical lasso (Friedman et al [22]). The disadvantage of graphical lasso is that it does not integrate prior structure when building the model pool. Even with SBIC in hand, we may still end up with a poor model choice. It is necessary to incorporate the prior structure into the construction of model pool. For example, in addition to the solution path of graphical lasso, we may simply include random samples from the Boltzmann distribution (10) corresponding to the prior structure as a part of the model pool. This method turns out to be inefficient for high-dimensional model given the huge size of model space. Alternatively, we can carefully devise the penalty term in graphical lasso so that the solution path can relate to the prior structure automatically. This method usually involves complex optimization algorithm that can not be easily solved based on the existing software. In the following, we propose an intuitive algorithm to build the model pool based on the prior structure, which can be easily implemented using the popular R package such as *glasso* (Friedman et al [22]) or *glmnet* (Friedman et al [24]). This algorithm bears some similarities to the well-known greedy equivalence search (GES) algorithm while the latter aims to learn the structure of directed acyclic graphical model (Chickering et al [16], Ramsey et al [59]). Recall that GES algorithm optimizes the given score function by an edge addition or removal in each step until the algorithm converges. If the true model is decomposable, it is proved that GES algorithm can consistently select the true model as sample size tends to infinity (Chickering et al [16]). Though our algorithm also involves edge addition and removal, the present objective is to construct the model pool instead of searching the optimum of score function. Specifically, the algorithm consists of the following two steps.

Step 1 (Edge enrichment): Since some edges may have been omitted by the prior graph, in this step we consider how to pick up the omitted edges. To this

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

end, for a given increasing sequence of λ, $0 \leq \lambda_1^{(1)} < \cdots < \lambda_{m_1}^{(1)}$, we solve the following optimization problems for $i = 1, \cdots, m_1$,

$$\arg\min_{\Omega:\tilde{E}\subseteq E(\Omega)} \left\{ \text{trace}(S\Omega) - \log\det(\Omega) + \lambda_i^{(1)} \|\Omega\|_1 \right\}, \qquad (13)$$

where $E(\Omega)$ is the edge set of graph corresponding to Ω, $\|\Omega\|_1 = \sum_{1\leq i<j\leq p} |\omega_{ij}|$. In (13) we fix the edges in prior structure and consider how to enrich it by selecting the edges from the rest edges. This step leaves us m_1 graphs denoted by $G^{(i)}$ for $i = 1, \cdots, m_1$ respectively.

Step 2 (Edge pruning): Each $G^{(i)}$ $(i = 1, \cdots, m_1)$ from the first step contains the prior structure. Since some redundant edges may have been mistakenly included in the prior structure, we aim to prune these edges from these graphs in this step. To this end, for a given $G^{(i)}$ $(1 \leq i \leq m_1)$, we solve the following m_2 optimization problems for an increasing sequence $0 \leq \lambda_1^{(2)} < \cdots < \lambda_{m_2}^{(2)}$,

$$\arg\min_{\Omega:E(\Omega)\subseteq E_i} \left\{ \text{trace}(S\Omega) - \log\det(\Omega) + \lambda_j^{(2)} \|\Omega\|_1 \right\}, \qquad (14)$$

where E_i is the edge set of graph $G^{(i)}$. In (14), for a given graphical model $G^{(i)}$, we consider how to prune the false edges that have been included in the prior structure. This step leaves us graphs $G^{(ij)}$ for $i = 1, \cdots, m_1$ and $j = 1, \cdots, m_2$. Thus there are total $m_1 m_2$ candidate models in the final model pool.

Remarks. (1) If the prior structure is an empty graph, then only *edge enrichment* step is involved to build the model pool; if the prior structure is a complete graph, then only *edge pruning* step is involved. The model pools for these two extreme cases turn out to be the same as that from the standard lasso. (2) The two-step algorithm above is implemented through the graphical lasso (Friedman et al [23]); nevertheless, the algorithm can also be equivalently implemented through the neighborhood method (Meinshansen and Buhlmann [49]). The consistency of neighborhood method is guaranteed by the property of lasso for high-dimensional regression model (Zhao and Yu [87]). When neighborhood method is used, the R package *glmnet* (Friedman et al [24]) can be employed to facilitate the computation.

A concern raised by the reviewers is about the performance of the proposed algorithm when there exists big discrepancy between the prior and true network. Firstly, we note that the proposed two-step algorithm always includes the solution path of lasso as a part of the model pool; consequently, the model pool can cover the true model if the sample size n is reasonably large with respect to the number of vertices p. As far as SBIC is concerned, this issue essentially is about how to select the temperature parameter T. If we have a good estimate of the discrepancy between prior and true network, or equivalently the average error rate r, then an appropriate T can be selected. In that case, a bad prior network will make the last term in SBIC relatively small and the model selection will be mainly determined by other terms in SBIC. However, if we mistakenly assume a small r for an actually large discrepancy, that will have a negative effect on the model selection. In that situation, sensitivity analysis of the result with respect to the tuning parameter T is recommended, based on which tuning parameter can be chosen to ensure a robust network selection, see section 5.1 for details.

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4. Consistency of SBIC

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

First let us introduce some notations for ease of exposition. Recall that $\mathbf{z}$ is the $\tilde{p}$-dimensional edge vector indicating whether or not there is an edge between two given vertices. Define $|\mathbf{z}| = \sum_{i=1}^{\tilde{p}} z_i$ and let $\mathbf{z_0}$ be the edge vector corresponding to the true graph E_0 under consideration. Let $\mathcal{E}_q$ denote the graph set with no more than q edges and $\mathcal{Z}_q \subset R^{\tilde{p}}$ the corresponding edge vector set. Let σ^2_{max} denote the largest diagonal component of the true covariance matrix Σ_0 while λ_{max} denote the largest eigenvalue of true precision matrix Θ_0. For a given positive semi-definite matrix W, let $\tau_{\max}$ and $\tau_{\min}$ be the largest and smallest eigenvalue of W respectively. With these notations in hand, the consistency for $\text{BIC}_{T,W}$ (7) and SBIC (11) are proved in Theorem 1 and 2 respectively. For $\text{BIC}_{T,W}$, 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 $\sigma^2_{max}\lambda_{max} \le C$ and $\theta_0 = \min_{e\in E_0} |(\Theta_0)_e| > 0$

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

Theorem 1. Under Assumptions 1-4, the model selection procedure based on $\text{BIC}_{T,W}$ given in (7) is consistent, i.e., as $n \to \infty$ we have

$$\mathbf{z_0} = \arg\min_{\mathbf{z}\in\mathcal{Z}_q} \text{BIC}_{T,W}(\mathbf{z}) \tag{15}$$

in probability.

Now let us consider SBIC (11) in which prior structure is available for the underlying graphical model. Recall that $\tilde{G} = (V, \tilde{E})$ is the prior structure, $\bar{G} = (V, \bar{E})$ is the difference graph between $\tilde{G}$ and G and $\bar{G}_0$ is the difference graph of prior graph $\tilde{G}$ and true graph G_0. Here we have assumed $\tilde{G}$ and G_0 have the same vertex set. In order to prove the consistency of SBIC (11) , Assumptions 1 and 4 have to be replaced by the following Assumptions $1^{'}$ and $4^{'}$.

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 edge set E_0 by adding or deleting at most q_2 edges from the prior edge set $\tilde{E}$ and so $E_0 \in \mathcal{E}_{q_1+q_2}$. Given the observations $\tilde{X} = (\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)})$, we have the following result.

Theorem 2. Under Assumption $1^{'}$ and 2, 3 and $4^{'}$, SBIC (11) can consistently select the true graph structure G_0, i.e., as $n \to \infty$, we have

$$\mathbf{z}_0 = \arg\min\nolimits_{\mathbf{z}\in\mathcal{Z}_{q_1+q_2}} \text{SBIC}_T(\mathbf{z}) \tag{16}$$

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in probability.

The details of the proof for Theorem 1 and 2 are provided in Supplementary Materials. It should be noted that, in order to facilitate the proof, we have imposed strong assumptions on the dimensionality and the underlying graphical structure such as decomposability. It is possible that the results still hold if some of these assumptions are relaxed. In particular, both the evaluation of SBIC and implementation of two-step algorithm do not depend on the decomposability of the underlying graph.

It also should be pointed out that we only discussed the problem of model selection for GGM in this study and did not investigate the problem of its statistical inference. It is implicitly assumed that the statistical inference can be reasonably conducted after the model is selected, though this is not always the case in practice. Additional assumptions including irrepresentability and beta-min condition have been suggested in literature to ensure the consistency of such estimate (B$\ddot{u}$hlmann et al [9]). Recently, several novel methods that integrated the model selection and the statistical inference have been proposed for GGM. For example, Ren et al [63] employed a two-dimensional regression model to estimate each entry of precision matrix and the asymptotical distribution is shown to be normal. Jankova and van de Geer [33] adapted the neighborhood method of Meinshansen and Buhlmann [49] based on the Karush-Kuhn-Tucker (KKT) conditions and proposed an estimator for each entry of precision matrix which is shown to converge asymptotically to the normal distribution. R package *SILGGM* has been developed to implement the statistical inference for GGM using these new methods (Zhang et al [84]).

5. Simulation Studies

In this section the proposed algorithm is evaluated based on simulated data. In the first example, we consider a tree graph. For tree graphs, the well-known Chow-Liu algorithm can optimally and efficiently learn the structure (Chow and Liu [17], Edwards et al [21], Kirshner et al [38]). We will compare the performance of our method with Chow-Liu algorithm. In addition, there is a temperature parameter T involved in SBIC. Though ideally T should be determined based on the prior information, e.g., expected error rate, in practice prior information are often biased to some extent. So we also perform sensitivity analyses to evaluate the robustness of our method with respect to the misspecification of T. In the second example, we go beyond the tree model and consider the randomly generated graphs which may be non-decomposable. Based on the simulated data from these graphs, we compare the proposed algorithm with other popular model selection methods in literature. It is demonstrated that the proposed algorithm can outperform these existing algorithms under the given scenarios in terms of two indices, true positive rate (TPR) and false positive rate (FPR), which are defined as

$$\text{TPR} = \frac{\#\{\text{identified true edges}\}}{\#\{\text{all true edges}\}}, \quad \text{FPR} = \frac{\#\{\text{falsely identified edges}\}}{\#\{\text{all null edges}\}}, \quad (17)$$

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

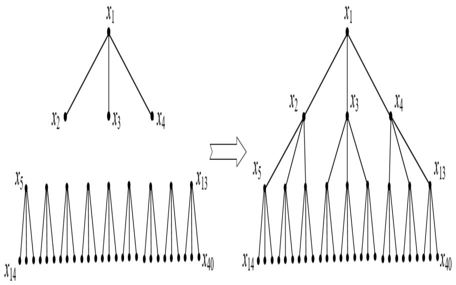

Figure 1: The graphs involved in Section 5.1. The left one is used as the prior graph while the right one is the true graph.

for which a higher TPR and lower FPR indicate a better model selection.

5.1. Sensitivity analysis based on tree model

Consider a Gaussian graphical model with fixed 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 the same fashion, all the variables can be generated. The structure of X is shown in the right plot in Figure 1. The left graph in Figure 1 is used as the prior structure.

Two plots in Figure 2 present TPR and FPR as the function of α respectively. In each plot two curves are drawn in which the solid curve corresponds to the model pool constructed from standard lasso while the dashed curve corresponds to the model pool constructed from two-step algorithm. In both cases SBIC is employed to select the model in which temperature parameter is set based on true error rate $r = 9/780$. Here the replication is $N = 100$; the sample size is $n = 60$. The difference in each plot reflects the difference between the two model pools. From Figure 2 it is obvious that TPR from two-step algorithm is higher than that from standard lasso while FPR from to-step algorithm is lower than that from standard lasso. In particular, the difference becomes more prominent when the association among the variables is weak.

Table 1 lists the results for multiple specifications of T and Chow-Liu algorithm under different scenarios. Specifically, the sample sizes are $n = 50, 100$; the number of replication is $N = 100$. Three choices of association strength are $\alpha = 0.3, 0.4, 0.5$. As for temperature parameter T, five choices for expected error rate, $r = 3/780, 9/780, 18/780, 27/780, 36/780$, are considered, from which T can be derived based on the formula in Section 3.1. For example, the value of T in SBIC$_1$ corresponding to $r = 3/780$ can be shown to be 0.332.

From Table 1 it can be seen that: (1) For the combination of SBIC and two-step algorithm, in most cases both TPR and FPR increase if T increases. For the rows with the error rate other than the true value $r = 9/780$, the

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

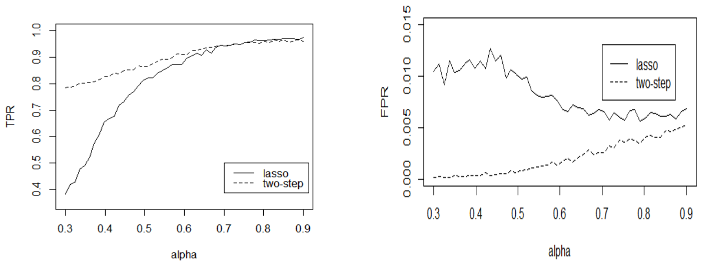

Figure 2: The left plot is the TPR versus association strength α and the right plot is the FPR versus α. The solid lines correspond to the model pool constructed based on lasso while the dashed lines correspond to the model pool constructed based on two-step algorithm.

moderate deviation of r from the true value does not have big impact on the final results; (2) The combination of BIC and lasso yields large false positive rates which explains the popularity of EBIC; the combination of BIC and two-step algorithm yields better results, i.e., higher TPR and lower FPR, especially when the association strength α is low and sample size is small; (3) In most cases the combination of SBIC and two-steep algorithm yields higher TPR and lower FPR than Chow-Liu algorithm. Nevertheless, compared to the combination of BIC and lasso, Chow-Liu algorithm yields comparable TPR and lower FPR.

In summary, if prior structure is available for graphical model, then both model selection criterion and candidate model should incorporate such information. The results from the proposed procedure also demonstrate robustness to the misspecification of temperature parameter.

5.2. Comparison with other model selection strategies based on general graphical model

In this section, we extend the tree graph considered in Section 5.1 to the randomly generated graph with up to 200 vertices. The proposed model selection algorithm is compared with four popular model selection methods in literature, including (1) Bayesian method; (2) CV(cross validation)+Lasso; (3) EBIC+Lasso; (4) BIC+rLasso. For Bayesian method, independent edge inclusion indicator variables are assumed for the edges set. The same Bernoulli prior is assumed for all indicator variables. With a given network structure, we then use the G-Wishart distribution as the prior distribution of precision matrix, which is known to be conjugate distribution for the normally distributed data. For second method CV+Lasso, model selection criterion is CV while model pool is built by lasso. No prior information is involved here. For EBIC+Lasso, model selection criterion is EBIC while model pool is built by lasso. Prior structure is used in EBIC in the same way as SBIC; however, the tuning parameter in (3) is set to be fixed at $\lambda = 0.5$ as commonly suggested in literature. Method 4 is proposed in Ma et al [48] where rLasso stands for residual lasso. Prior

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: Sensitivity analysis of structural Bayesian information criterion (SBIC) with respect to the temperature parameter. The first number in the parentheses is true positive rate (TPR), and the second number is false positive rate (FPR). The results for BIC and Chow-Liu algorithm are also listed. The number of replication for each result is $N = 100$.

	n=50			n=100		
	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$
SBIC_1+TS	(0.794, 0.0000)	(0.838, 0.0003)	(0.876, 0.0009)	(0.894, 0.0005)	(0.941, 0.0007)	(0.971, 0.0009)
SBIC_2+TS	(0.809, 0.0003)	(0.844, 0.0008)	(0.885, 0.0012)	(0.912, 0.0008)	(0.953, 0.0007)	(0.974, 0.0009)
SBIC_3+TS	(0.816, 0.0008)	(0.848, 0.0009)	(0.894, 0.0015)	(0.902, 0.0008)	(0.952, 0.0008)	(0.982, 0.0013)
SBIC_4+TS	(0.828, 0.0008)	(0.870, 0.0014)	(0.899, 0.0019)	(0.904, 0.0008)	(0.949, 0.0007)	(0.975, 0.0012)
SBIC_5+TS	(0.827, 0.0011)	(0.870, 0.0015)	(0.902, 0.0026)	(0.902, 0.0007)	(0.951, 0.0010)	(0.979, 0.0012)
BIC+TS	(0.916, 0.0323)	(0.948, 0.0283)	(0.958, 0.0198)	(0.972, 0.0171)	(0.991, 0.0146)	(0.996, 0.0113)
BIC+Lasso	(0.664, 0.0275)	(0.836, 0.0261)	(0.907, 0.0261)	(0.932, 0.0282)	(0.981, 0.0240)	(0.993, 0.0134)
Chow-Liu	(0.653, 0.0179)	(0.839, 0.0079)	(0.917, 0.0039)	(0.9107, 0.0050)	(0.979, 0.0009)	(0.991, 0.0003)

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

information is used in rLasso to construct the model pool. Specifically, when a part of the structure is known a priori with certainty, Ma et al [48] proposed to use lasso to construct the model pool based on the residuals from the linear regression of each variable on its known neighbors. Given such model pool, they then employed BIC to select graphical model. Such method to build the model pool, however, will be biased when the prior structure involves randomness. The two-step algorithm proposed in this paper takes all the vertices into consideration in the enrichment step which theoretically leads to a less biased model pool than that from rLasso. Furthermore, since the model pool can not fully reflect the randomness information in prior structure, the combination of SBIC and the proposed model pool should have better performance if the prior information have been reasonably specified .

Specifically, we first randomly generate a $p \times p$ adjacency matrix M_1 as the true structure, in which the number of edges, i.e., the number of 1's among the off-diagonal entries of M_1, is set to be 100. The adjacency matrix of prior structure M_2 is generated by randomly changing 100α percent of the 1's entries of M_1 to 0 and the same number of 0's entries to 1. Given M_1, a symmetrical matrix with 1's on its diagonal is generated which has the same edge set as M_1. Each of nonzero entries in this matrix is generated from $N(0, 1)$ distribution. By tuning the diagonal element 1 to some value β, we can always get a positive definite matrix K, which will be used as the precision matrix in this study. Here we choose $\beta = 1.1 - \lambda_{\min}(A)$ in which $\lambda_{\min}(A)$ denotes the minimum eigenvalue of matrix A. For $p = 100, 200$, $\alpha = 0.7, 0.5, 0.3, 0.1$, and sample size $n = 100, 200$, Table 2 lists the results of TPR and FPR in each scenario for all the five model selection strategies. For Bayesian method, the inclusion probability for each edge is set to be $\theta = 200/p(p-1)$, which means that we do not assume any specific structure for the graph in its prior distribution other than the sparsity. For the prior G-Wishart distribution of precision matrix, $W_G(b, D)$, we set $b = 3$ and D the $p \times p$ identity matrix. The burn-in for sampling from the posterior distribution is set to be 5000. For a given vertex pair (X_i, X_j), if 50% precision matrix samples have nonzero (i, j)th entries, then we define an edge between X_i and X_j; Otherwise, no edge is defined between X_i and X_j. As for methods 2 to 5, the number of tuning parameters to build the model pool is set to be 100. The temperature parameter in SBIC is set based on the discrepancy rate between real and prior networks, i.e., $\alpha = 0.7, 0.5, 0.3, 0.1$. Note since method CV+Lasso does not involve any prior information, so it has the same TPR's and FPR's in all the four discrepancy situations in Table 2. It can be seen that the performance of Bayesian method is sensitive to the prior information. It should be noted that Bayesian method, which is implemented through the R package *BDgraph* (Mohammadi and Wit [52]), is much more time-consuming than the other four methods; CV+Lasso tends to select the graphs with too many false edges; EBIC+Lasso tends to omit too many true edges. Though BIC+rLasso has a better performance than CV+Lasso and EBIC+Lasso, in most cases, it still has a high probability to omit the true edges and select the false edges. As for our proposed strategy, it works well and reaches a good balance between TPR and FPR, and compared to other four

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

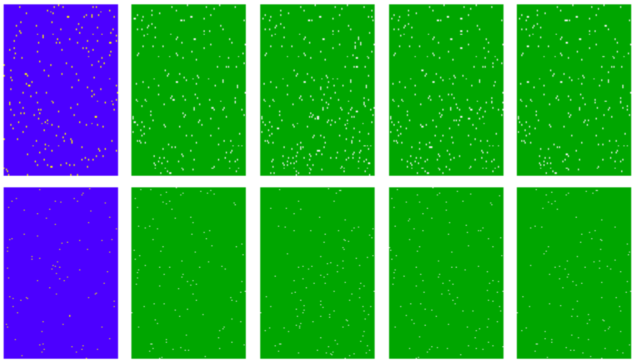

Figure 3: A realization of the real and prior networks in simulation studies in Section 5.2. Dots indicate the nonzero conditional associations between vertices. The five plots in the first row are for network with $p = 100$ vertices. The first one is the true network, and the discrepancy between the other four networks and true network are, from left to right, $\alpha = 70\%, 50\%, 30\%, 10\%$ respectively. These networks are used as prior networks respectively. Similar explanation applies to the networks in second row which have $p = 200$ vertices.

methods, yields higher TPR while lower FPR in most cases, especially when the discrepancy between prior structure and true structure becomes smaller.

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2: Performance comparison of different model selection strategies. The first number in the parentheses is true positive rate (TPR), and the second number is false positive rate (FPR). Number α stands for the proportion of the edges that have been falsely included in prior network.The number of replication for each result is $N = 100$.

			$\alpha = 70\%$	$\alpha = 50\%$	$\alpha = 30\%$	$\alpha = 10\%$
p=100	n=100	Bayesian	(0.220, 0.0074)	(0.400, 0.0054)	(0.601, 0.0038)	(0.829, 0.0016)
		CV+gLasso	(0.531, 0.0479)	(0.531, 0.0479)	(0.531, 0.0479)	(0.531, 0.0479)
		EBIC+gLasso	(0.259, 0.0005)	(0.321, 0.0009)	(0.322, 0.0006)	(0.354, 0.0011)
		BIC+rLasso	(0.358, 0.0080)	(0.463, 0.0044)	(0.552, 0.0050)	(0.598, 0.0069)
		SBIC+TS	(0.395, 0.0084)	(0.614, 0.0085)	(0.780, 0.0066)	(0.915, 0.0021)
	n=200	Bayesian	(0.238, 0.0047)	(0.404, 0.0044)	(0.608, 0.0034)	(0.826, 0.0015)
		CV+glasso	(0.641, 0.0518)	(0.641, 0.0518)	(0.641, 0.0518)	(0.641, 0.0518)
		EBIC+glasso	(0.427, 0.0004)	(0.469, 0.0008)	(0.521, 0.0010)	(0.488, 0.0014)
		BIC+rlasso	(0.500, 0.0056)	(0.558, 0.0053)	(0.702, 0.0053)	(0.898, 0.0029)
		SBIC+TS	(0.565, 0.0057)	(0.703, 0.0105)	(0.829, 0.0067)	(0.943, 0.0024)
p=200	n=100	Bayesian	(0.227, 0.0014)	(0.391, 0.0012)	(0.594, 0.0009)	(0.817, 0.0003)
		CV+glasso	(0.456, 0.0158)	(0.456, 0.0158)	(0.456, 0.0158)	(0.456, 0.0158)
		EBIC+glasso	(0.277, 0.0000)	(0.292, 0.0001)	(0.324, 0.0001)	(0.317, 0.0002)
		BIC+rlasso	(0.389, 0.0028)	(0.487, 0.0038)	(0.726, 0.0028)	(0.810, 0.0024)
		SBIC+TS	(0.434, 0.0028)	(0.690, 0.0033)	(0.774, 0.0017)	(0.928, 0.0005)
	n=200	Bayesian	(0.229, 0.0013)	(0.418, 0.0010)	(0.606, 0.0008)	(0.825, 0.0003)
		CV+glasso	(0.571, 0.0151)	(0.571, 0.0151)	(0.571, 0.0151)	(0.571, 0.0151)
		EBIC+glasso	(0.406, 0.0001)	(0.422, 0.0001)	(0.442, 0.0002)	(0.479, 0.0002)
		BIC+rlasso	(0.572, 0.0037)	(0.634, 0.0033)	(0.774, 0.0036)	(0.948, 0.0013)
		SBIC+TS	(0.622, 0.0041)	(0.746, 0.0035)	(0.805, 0.0015)	(0.940, 0.0005)

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. Metabolite Network in Human Gut

Metabolites in human body are intrinsically related with different diseases. Understanding the relationship between metabolites are helpful to design appropriate treatment. To this end, multiple methods have been proposed in literature to identify the structure of metabolite networks . For example, Gao et al [27], Karnovsky et al [40] used the biochemical domain knowledge to construct the metabolite network. Barupal et al [3], Grapov et al [29] constructed the network based on structural similarity and mass spectral similarity of metabolites. The metabolite prior networks in this paper are constructed using the similar method to that in Gao et al [27], Karnovsky et al [40].

The dataset involved comes from the New Hampshire Birth Cohort Study, an ongoing prospective cohort study of women and their young children Madan et al [46]. 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 Banerjee et al [2], Brim et al [8], Pathmasiri et al [57], Sumner et al [73, 74]. Chenomx NMR Suite 8.4 Professional software (Edmonton, Alberta, Canada) was used to determine relative concentration (Weljie et al [82]) of selected metabolites from a curation of list of metabolites that are associated with host-microbiome metabolism, see Li et al [44], Paul et al [58]. 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 [78]. In the following, we consider how to learn the structure of the network among these metabolites using the algorithm proposed in Section 3.

Specifically, we use pathway analysis to construct the prior structure. These pathway information 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. Note that 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 be $Z_i = \{Z_{i1}, \cdots, Z_{im_i}\}$ and $Z_j = \{Z_{j1}, \cdots, Z_{jm_j}\}$ respectively. Denote the common pathways of X_i and X_j by $Z_{ij} = Z_i \cap Z_j$ and let

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

If $s_{ij} \geq 0.8$, then we define an edge between X_i and X_j in the prior graph. With threshold equal to 0.8, there are 27 edges in the prior graph. With threshold equal to 0.6, there are 117 edges in the prior graph. We use the difference of

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

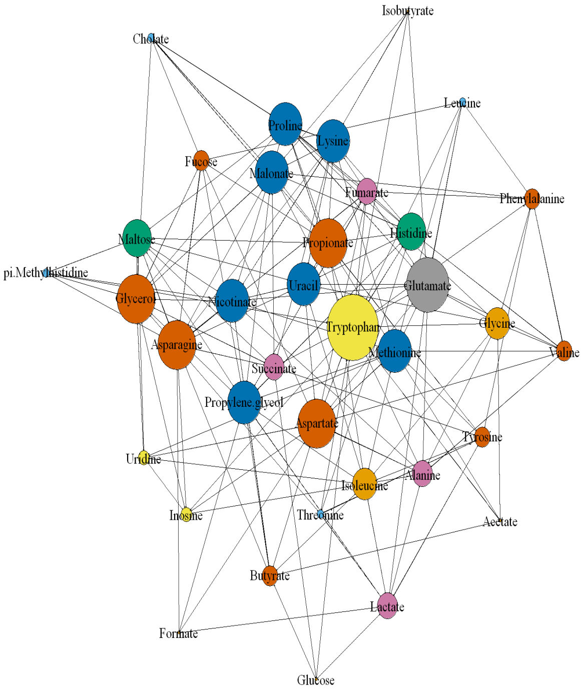

Figure 4: The edges that appear in the final network while have not been included in prior network.

these two number as the expected number of edges in difference graph between the prior 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 (13) and (14), where $\lambda_{\max}$ represents the minimal λ at which the graph has no edge. Then based on SBIC (11) and two-step algorithm, we obtain the final network. Comparison of the prior and the final network reveals that there are 153 edges added to and 3 edges removed from the prior network. Figure 4 shows the added edges. The three removed edges are (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 a pathway does not contain any information about association between metabolites, then such prior net-

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

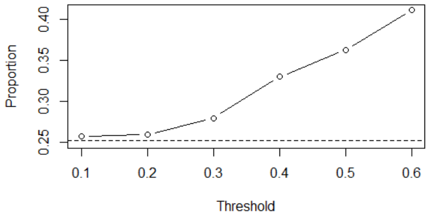

Figure 5: The proportion of the added edges in prior network E_s as a function of threshold s. The bottom dashed line corresponds to the line under the null hypothesis.

work can be regarded as built just randomly. Then the probability p_1 that an edge is removed 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 pathways have statistically significant effect on the association between metabolites.

One potential concern about the previous analysis is that 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 $s_{ij} \geq s$ otherwise not. Note for $s = 0.8$ there are 153 added edges among total 603 edges, apart from the 27 prior edges, that have been selected by the proposed method. 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 on these results, we can say that pathway does contain the association information between metabolites. The more pathways two metabolites share, the more likely their concentrations are related. Figure 5 shows the empirical probability of nonzero 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

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3: Edges in Figure 4 that can not be covered by conventional 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

subjective belief.

We have confirmed that part of the association among metabolites can be attributed to pathway. The next question we aim to address is that whether all the association among metabolites can be explained solely by pathway. To try to answer this question, first we define a densest prior structure among metabolites based on pathway. Specifically, whenever two metabolites have any pathway in common, we define 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 cannot explain all the association between 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 the rest are related 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 (Morshed et al [54, 55]). In light of these existing studies, it is not surprising to find out their wide connections with other metabolites even they do not share any pathway.

In summary, metabolic pathway can explain part of the association between 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 algorithm can discover these omitted associations efficiently.

7. Conclusion

We develop 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

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

part, we propose a novel model selection criterion called structural BIC which is a generalization of extended BIC. In second part, we propose a two-step algorithm to construct the candidate model pool which incorporates the prior structure into the candidate model through edge enrichment and pruning. It is proved that under some mild conditions the structural BIC is a consistent criterion for graphical model selection. Simulation results validate the efficacy and robustness of the algorithm.

We apply the proposed algorithm to the concentration data of metabolite in human gut 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 between 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 find new relationships between metabolites that can be omitted by conventional pathway analysis.

It is possible to use other types of prior network for metabolites, e.g, the structural similarity based prior network. Other biological network such as gene regulation network or microbial interaction network can also be analyzed based on our method if the prior structure can be properly defined. The algorithm can be 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 (R01LM012012; R01LM012723; P20ES018175; P01ES022832; UG3OD023275), US Environmental Protection Agency grant (RD83459901).

Supplementary Materials

Software in the form of R code is available at *https://github.com/hoenlab/SBIC*. The Supplementary Materials for this article are available in Appendix A-C, which include the marginal distributions of homogeneous Boltzmann distribution, the proofs of Theorem 1 and 2 and the prior graph for metabolomic data in section 5.

References

[1] Akaike, H. Statistical predictor identification. Ann. Inst. Statist. Math. 22, 203-217.

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[2] 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.

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

[4] Bogdan M., Ghosh J K., Doerge R W. (2004) Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics, 167(2): 989–99.

[5] Bogdan M., Ghosh J K., Doerge R W. (2008) Selecting ExplanatoryVariables with theModifiedVersion of the Bayesian Information Criterion, Quality and Reliability Engineering International, 24(6): 627-641.

[6] Bogdan M, Frommlet F, Biecek P, Cheng R, Ghosh JK, Doerge RW. (2008). Extending the modified bayesian information criterion (mBIC) to dense markers and multiple interval mapping. Biometrics, 64(4):1162-1169 DOI: 10.1111/j.1541-0420.2008.00989.x PMID: 18266892.

[7] 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.

[8] 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).

[9] B*ü*hlmann, P. and van de Geer, S. (2011). Statistics for high-dimensional data. Springer.

[10] Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 76, 503-514.

[11] Carvalho, C. M., Massam, H., and West, M. (2007). Simulation of Hyperinverse Wishart Distributions in Graphical Models. Biometrika, 94(3): 647–659. MR2410014. doi: http://dx.doi.org/10.1093/biomet/asm056.

[12] 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.

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

[16] Chickering DM. (2002) Optimal structure identification with greedy search. Journal of Machine Learning Research, 3: 507-554.

[17] Chow C K., Liu C N. (1968) Approximating discrete probability distribution with dependence tress. IEEE Trans Inf Theory, 14, 462-467.

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

[19] Dobra, A., Hans, C., Jones, B., Nevins, J. R., and West, M. (2004). Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis, 90: 196– 212. MR2064941. doi: http://dx.doi.org/10.1016/j.jmva.2004.02.009.

[20] Dobra, A. and Lenkoski, A. (2011). Copula Gaussian Graphical Models and Their Application to Modeling Functional Disability Data. The Annals of Applied Statistics, 5(2A): 969–993. MR2840183. doi: http://dx.doi.org/10.1214/10-AOAS397.

[21] Edwards D., de Abreu GCG., Labouriau R. (2010) Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. BMC Bioinform, 11: 18.

[22] Friedman J., Hastie T., Tibshirani R. (2008) R package *glasso*, URL: *http://www-stat.stanford.edu/tibs/glasso*, Version: 1.11.

[23] Friedman J., Hastie T., Tibshirani R. (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432-41.

[24] Friedman J., Hastie T., Tibshirani R., Narasimhan B., Simon N., Qian J (2019), R package: *glmnet*, URL: *https://glmnet.stanford.edu*, Version 3.0-2.

[25] D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. The Annals of Statistics, 22:1947–1975, 1994.

[26] Foygel Rina., Drton Mathias. (2010). Extended Bayesian information criteria for Gaussian graphical models, NIPS.

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[27] 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.

[28] Geisser, S. (1975). The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 70, 320-328.

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

[30] Hojsgaad S., Edwards D., Lauritzen S. (2012) Graphical Models with R. New York: Springer.

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

[32] 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.

[33] Janková, Jana.; van de Geer, Sara. (2017) Honest confidence regions and optimality in high-dimensional precision matrix estimation. TEST 26, 143–162. https://doi.org/10.1007/s11749-016-0503-5.

[34] Janková, Jana; van de Geer, Sara. (2015). Confidence intervals for high-dimensional inverse covariance estimation. Electron. J. Statist. 9, no. 1, 1205–1229. doi:10.1214/15-EJS1031. https://projecteuclid.org/euclid.ejs/1433195859

[35] Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2004). Experiments in Stochastic Computation for High-Dimensional Graphical Models. Statistical Science, 20: 388–400.

[36] Jalali A., Johnson C., Ravikumar P. (2011). On Learning Discrete Graphical Models using Greedy Methods. Advances in Neural Information Processing Systems 24 (NIPS 2011), 1935–43.

[37] Kim, Y., Kwon, S., Choi, H. (2012) Consistent Model Selection Criteria on High Dimensions, Journal of Machine Learning Research 13 (2012) 1037-1057.

[38] Kirshner S., Smyth P., Robertson AW. (2004) Conditional Chow-Liu tree structures for modeling discrete-valued vector time series. UAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence July 2004: 317–324.

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[39] 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.

[40] 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.

[41] Lauritzen S L. (1996). Graphical Models. Oxford University Press.

[42] Lauritzen, S. L. and Sheehan, N. A. (2003). Graphical Models for Genetic Analyses. Statistical Science, 18(4): 489–514. MR2059327. doi: http://dx.doi.org/10.1214/ss/1081443232.

[43] Leday G. G. and Richardson S. (2019). Fast Bayesian inference in large Gaussian graphical models. Biometrics, 75(4): 1288-1298.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] Ma J, Shojaie Ali, Michailidis George. (2016) Network-based pathway enrichment analysis with incomplete network information. Bioinformatics, 32(20): 3165-3174.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[51] Mitsakakis, N., Massam, H., and Escobar, M. D. (2011). A Metropolis-Hastings Based Method for Sampling from the G-Wishart Distribution in Gaussian Graphical Models. Electronic Journal of Statistics, 5: 18–30.

[52] Mohammadi R, Wit EC (2019). BDgraph: Bayesian Structure Learning in Graphical Models using Birth-Death MCMC. R package version 2.59, URL http://CRAN.R-project.org/ package=BDgraph.

[53] Mohammadi R (2019). ssgraph: Bayesian Graphical Estimation using Spike-and-Slab Priors. R package version 1.8.

[54] 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.

[55] 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.

[56] Mukherjee, Sach., Speed, Terence. (2008). Network Inference using Informative Priors. Proceedings of the National Academy of Sciences of the United States of America. 105. 14313-8. 10.1073/pnas.0802272105.

[57] 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.

[58] 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.

[59] Ramsey, J., Glymour, M., Sanchez-Romero, R. et al. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. Int J Data Sci Anal 3, 121–129 (2017). https://doi.org/10.1007/s41060-016-0032-z.

[60] Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. Biometrika 76, 369-374

[61] 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.

[62] Ray A., Sanghavi S., Shakkottai S. Improved Greedy Algorithms for Learning Graphical Models (2015). IEEE Transactions on Information Theory, 61(6): 3457-3468.

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[63] BY Ren Z, Sun T, Zhang CH, Zhou H. (2015) Asymptotic Normality and Optimalities in Estimation of Large Gaussian Graphical Models. The Annals of Statistics, 43(3): 991–1026, DOI: 10.1214/14-AOS1286.

[64] 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.

[65] 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.

[66] Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. Journal of Statistical Planning and Inference, 136(7): 2144–2162. MR2235051. doi: http://dx.doi.org/10.1016/j.jspi.2005.08.031.

[67] 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 .

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

[69] Schwarz, G. (1978). Estimating the dimensions of a model. Ann. Statist. 6, 461-464.

[70] Shao, J. (1993). Linear model selection by cross-validation. J. Amer. Statist. Assoc. 88, 486-494.

[71] J. Shao. An asymptotic theory for linear model selection. Statistica Sinica, 7:221–264, 1997.

[72] Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. The Annals of Applied Statistics, 4(4): 2024–2048.

[73] 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.

[74] 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.

bioRxiv preprint doi: https://doi.org/10.1101/815423; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[75] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. Ser. B 36, 111-147.

[76] Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (methodological), Wiley, 58 (1): 267–288.

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

[78] 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.

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

[80] Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. Bayesian Analysis, 7(4): 867–886. MR3000017. doi: http://dx.doi.org/10.1214/ 12-BA729.

[81] Wang, H. and Carvalho, C. M. (2010). Simulation of hyper-inverse Wishart distributions for non-decomposable graphs. Electronic Journal of Statistics, 4: 1470–1475. MR2741209. doi: http://dx.doi.org/10.1214/10-EJS591.

[82] 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.

[83] Zhang, P. (1993). Model selection via multifold cross validation. Ann. Statist. 21, 299-313.

[84] Zhang R, Ren Z, Chen W (2018) SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. PLOS Computational Biology 14(8): e1006369. https://doi.org/10.1371/journal.pcbi.1006369

[85] Zhang Y., Shen X. Model selection procedure for high-dimensional data. Statistical Analysis and Data Mining, 3:350–358, 2010.

[86] Zhang, Y., Lv, J., Liu, H., Zhu, J., Su, J., Wu, Q., Qi, Y., Wang, F., and Li, X. (2010). HHMD: the human histone modification database. Nucleic Acids Research, 38: 149–154.

[87] Zhao P., Yu B. (2006). On Model Selection Consistency of Lasso. J. Mach. Learn. Res. 7, 2541–2563.