Bernoulli **27**(1), 2021, 637–672
https://doi.org/10.3150/20-BEJ1253

Bayesian graph selection consistency under model misspecification

YABO NIU[*], DEBDEEP PATI[†] and BANI K. MALLICK[‡]

Department of Statistics, Texas A&M University, College Station, TX, USA.
E-mail: [*] *ybniu@stat.tamu.edu;* [†] *debdeep@stat.tamu.edu;* [‡] *bmallick@stat.tamu.edu*

Gaussian graphical models are a popular tool to learn the dependence structure in the form of a graph among variables of interest. Bayesian methods have gained in popularity in the last two decades due to their ability to simultaneously learn the covariance and the graph. There is a wide variety of model-based methods to learn the underlying graph assuming various forms of the graphical structure. Although for scalability of the Markov chain Monte Carlo algorithms, decomposability is commonly imposed on the graph space, its possible implication on the posterior distribution of the graph is not clear. *An open problem* in Bayesian decomposable structure learning is whether the posterior distribution is able to select a meaningful decomposable graph that is "close" to the true non-decomposable graph, when the dimension of the variables increases with the sample size. In this article, we explore specific conditions on the true precision matrix and the graph, which results in an affirmative answer to this question with a commonly used hyper-inverse Wishart prior on the covariance matrix and a suitable complexity prior on the graph space. In absence of structural sparsity assumptions, our strong selection consistency holds in a high-dimensional setting where $p = O(n^{\alpha})$ for $\alpha < 1/3$. We show when the true graph is non-decomposable, the posterior distribution concentrates on a set of graphs that are *minimal triangulations* of the true graph.

Keywords: decomposable graph; Gaussian graphical model; graph selection consistency; hyper-inverse Wishart distribution; minimal triangulation; model misspecification; partial correlation

1. Introduction

Graphical models provide a framework for describing statistical dependencies in (possibly large) collections of random variables [27]. In this article, we revisit the well-known problem of inference on the underlying graph using observed data from a Bayesian point of view. Research on Bayesian inference for natural exponential families and associated conjugate priors, called the Diaconis–Ylvisaker (DY priors), is pioneered by [13] and has a profound impact on the development of Bayesian Gaussian graphical models. Consider independent and identically distributed vectors $Y_1, Y_2, \ldots, Y_n$ drawn from a p-variate normal distribution with mean vector $\mathbf{0}$ and a sparse inverse covariance matrix Ω. The sparsity pattern in Ω can be encoded in terms of a graph G on the set of variables as follows. If the variables i and j do not share an edge in G, then $\Omega_{ij} = 0$. Hence, an undirected (or concentration) graphical model corresponding to G restricts the inverse covariance matrix Ω to a linear subspace of the cone of positive definite matrices.

A probabilistic framework for learning the dependence structure and the graph G requires specification of a prior distribution for (Ω, G). Conditional on G, a hyper-inverse Wishart (HIW) distribution [11] on $\Sigma = \Omega^{-1}$ and the corresponding induced class of distributions on Ω [36] are attractive choices of DY priors. A rich family of conjugate priors that subsumes the DY class is developed by [29]. Bayesian procedures corresponding to these Letac–Massam priors have been derived in a decision theoretic framework in the recent work of [33]. The key component of Bayesian structure learning is achieved through the specification of a prior distribution on the space of graphs. There is a need for a flexible but tractable family of such priors, capable of representing a variety of prior beliefs about

the conditional independence structure. During the last two decades, there has been a growing literature on the development of HIW priors focusing on decomposable graphs [9,10,19,20] and their non-decomposable counterparts [2,12,25,31,37,41]. Although deemed as a restrictive model choice in the space of graphs, in the interest of tractability and scalability, HIW priors on decomposable graphs continue to be widely used. In this paper, we focus on the HIW priors on decomposable graphs as this construction enjoys many advantages, such as computational efficiency due to its conjugate formulation and exact calculation of marginal likelihoods [38]. Stochastic search algorithms are empirically demonstrated to have good practical performances in these models. For detailed descriptions and comparisons of various Bayesian computational methods in these scenarios, see [15,24].

There has been a growing literature on model selection consistency in Gaussian graphical models from a frequentist point of view [16,30,34,42]. Beyond the literature on Gaussian graphical models, there has been an incredible amount of frequentist work in the context of estimating high-dimensional covariance matrix with rates of convergence of various regularized covariance estimators derived in [6, 7,17,26] among others. There is a relatively smaller literature on asymptotic properties of Bayesian procedures for covariance or precision matrices in graphical models; refer to [3,4]. Although the methodology in [4] can be used for graph selection, the theoretical results in [3,4] are focused on achieving an optimal rate of posterior convergence for the covariance and precision matrices. The literature on graph selection consistency in a Bayesian paradigm is surprisingly sparse [8,18,28]. In the context of decomposable graphs, the only article we were aware of is [18] which considered the behavior of Bayesian procedures that perform model selection for decomposable Gaussian graphical models. However, the analysis is restricted to the fixed dimensional regime and involves the behavior of the marginal likelihood ratios between graphs differing only by one edge. Although Bayes factors in the general case can be decomposed using Bayes factors between graphs that differ by one edge, it is not possible to derive the consistency results by simple aggregation of Bayes factors for single edge moves. Furthermore, in high dimension the number of single edge moves involved in the aggregation may increase with the graph size. In such cases, the consistency results necessitate more restrictive assumptions on the growth of the number of edges. For general graph selection consistency within a Bayesian framework, refer to the very recent articles [8,28] in the context of Gaussian directed acyclic graph (DAG) models.

In this article, focusing on the hyper-inverse Wishart g-prior [10] on the covariance matrix and a complexity prior on the graph, we derive sufficient conditions for strong selection consistency when $p = O(n^{\alpha})$ with $\alpha < 1/3$ considering both the cases when the true graph is decomposable and when it is not. The key conditions relate to the precise upper and lower bounds on the partial correlation coefficients and a suitable complexity prior on the space of graphs. We emphasize here that we do *not* need conditions to be verified on all subgraphs – all assumptions are easy to understand and relatively straightforward to verify. Regarding our findings, we discover that the HIW g-prior places a heavy penalty on missing true edges (false negatives), but a comparatively smaller penalty on adding false edges (false positives). Henceforth in the high-dimensional regime a carefully chosen complexity prior on the graph space is needed for penalizing false positives and achieving strong consistency.

In the well-specified case, the hierarchical model used here is a subset of [8] since hyper-inverse Wishart prior is a special case of DAG-Wishart prior proposed in [5] under *perfect* DAGs. However, the assumptions in this paper are distinctly different from those stated in [8]. In particular, our assumptions are on the magnitude of the elements of the partial correlation matrix rather than on the eigenvalues of the covariance matrix as in [8]. Also, the main focus of this article is to study the behavior of graph selection consistency under model misspecification, which cannot be addressed within a DAG framework. To the best of our knowledge, this is the first paper to show the strong selection consistency under HIW priors for high-dimensional graphs under model misspecification. In particular, we show that the posterior concentrates on decomposable graphs which are in some sense closest to the true non-decomposable graph. Interestingly, the pairwise Bayes factors between such graphs are stochastically

bounded. Our result under model-misspecification is inspired by [18], but extends to the case when p is growing with n and provides a rigorous proof of the convergence of the posterior distribution to the class of decomposable graphs which are closest to the true one. We also present a detailed simulation study both for the well-specified and misspecified case, which provides empirical justification for some of our technical results.

En-route, we develop precise bounds for the Bayes factor in favor of an alternative graph with respect to the true graph. The main proof technique is a combination of (a) *localization*: which involves breaking down the Bayes factor between any two graphs into local moves, that is, addition and deletion of one edge using decomposable graph chain rule [27] and (b) *correlation association*: which converts the Bayes factor between two graphs differing by an edge into a suitable function of sample partial correlations. By developing sharp concentration inequalities and tail bounds for sample partial correlations, we obtain bounds for ratios of local marginal likelihoods which are then combined to yield strong selection consistency results.

The remaining part of the paper is organized as follows. In Section 2, we introduce the necessary background and notations. Section 3 introduces the model with the HIW prior. Section 4 describes the main results of pairwise posterior ratio consistency and consistent graph selection when the true graph is decomposable. In Section 4, the results are presented progressively as follows. First, we provide a non-asymptotic sharp upper bound for pairwise Bayes factors. Next, we state the main theorem for posterior ratio consistency when p diverges with n where p is of the order n^{α} for $\alpha < 1/2$. Finally, we state the main theorem on strong graph selection consistency which further requires $\alpha < 1/3$. Section 5 states the main results on consistent graph selection under model misspecification and results on the equivalence of minimal triangulations. Numerical experiments are presented in Section 6 followed by a discussion in Section 7.

2. Preliminaries

In this section, we define a collection of notations required to describe the model and the prior. Section 2.1 introduces sample and population correlations and partial correlations, Section 2.2 sets up the notations for undirected graphs and briefly introduces the definitions and properties associated with decomposable graphs. Section 2.3 contains matrix abbreviations and notations used throughout the paper.

2.1. Correlation and partial correlation

Let $\boldsymbol{X}_p = (X_1, X_2, \ldots, X_p)^T$ denote a random vector which follows a p-dimensional Gaussian distribution and $\mathrm{x}^{(1)}, \mathrm{x}^{(2)}, \ldots, \mathrm{x}^{(n)}$ denote n independent and identically distributed (i.i.d.) sample observations from $\boldsymbol{X}_p$. Clearly, the $n \times p$ matrix formed by augmenting the n-dimensional column vectors x_i, denoted by $(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_p)$, is the same as $(\mathrm{x}^{(1)}, \mathrm{x}^{(2)}, \ldots, \mathrm{x}^{(n)})^T$ and $\bar{\mathrm{x}}_i = n^{-1}\mathbb{1}_n^T \mathrm{x}_i$, $i = 1, 2, \ldots, p$. Here $\mathbb{1}_n$ is an n-dimensional vector with all ones. Let I_n denote the $n \times n$ identity matrix and $\mathbb{E}$ denote the expectation with respect to the Gaussian random variable $\boldsymbol{X}_p$.

Definition 2.1 (Population correlation coefficient). The population correlation coefficient between X_i and X_j, $1 \le i, j \le p$, is defined as

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}},$$

where $\sigma_{ii} = \mathbb{E}(X_i - \mathbb{E}X_i)^2$ and $\sigma_{ij} = \mathbb{E}\{(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)\}$.

Definition 2.2 (Sample/Pearson correlation coefficient). The sample correlation coefficient between X_i and X_j, $1 \le i, j \le p$, is defined as

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}}\sqrt{\hat{\sigma}_{jj}}},$$

where $\hat{\sigma}_{ii} = (\mathrm{x}_i - \bar{\mathrm{x}}_i \mathbb{1}_n)^T (\mathrm{x}_i - \bar{\mathrm{x}}_i \mathbb{1}_n)/n$ and $\hat{\sigma}_{ij} = (\mathrm{x}_i - \bar{\mathrm{x}}_i \mathbb{1}_n)^T (\mathrm{x}_j - \bar{\mathrm{x}}_j \mathbb{1}_n)/n$.

Definition 2.3 (Population partial correlation coefficient). Let $S = \{i_1, i_2, \ldots, i_{|S|}\}$, where $1 \le i_1, i_2, \ldots, i_{|S|} \le p$ and $|S|$ is the cardinality of set S. Define $X_S = (X_{i_1}, X_{i_2}, \ldots, X_{i_{|S|}})^T$. The population partial correlation coefficient between X_i and X_j, where $i, j \notin S$ and $1 \le i, j \le p$, holding X_S fixed is defined as

$$\rho_{ij|S} = \frac{\sigma_{ij|S}}{\sqrt{\sigma_{ii|S}}\sqrt{\sigma_{jj|S}}},$$

where $\sigma_{ii|S} = \sigma_{ii} - \sigma_{Si}^T \sigma_{SS}^{-1} \sigma_{Si}$, $\sigma_{ij|S} = \sigma_{ij} - \sigma_{Si}^T \sigma_{SS}^{-1} \sigma_{Sj}$. And $\sigma_{Si} = \mathbb{E}\{(X_S - \mathbb{E}X_S)(X_i - \mathbb{E}X_i)\}$, $\sigma_{SS} = \mathbb{E}\{(X_S - \mathbb{E}X_S)^T (X_S - \mathbb{E}X_S)\}$.

Definition 2.4 (Sample partial correlation coefficient). Define $\mathrm{x}_S = (\mathrm{x}_{i_1}, \mathrm{x}_{i_2}, \ldots, \mathrm{x}_{i_{|S|}})$. The sample partial correlation coefficient between X_i and X_j, where $i, j \notin S$ and $1 \le i, j \le p$, holding X_S fixed is defined as

$$\hat{\rho}_{ij|S} = \frac{\hat{\sigma}_{ij|S}}{\sqrt{\hat{\sigma}_{ii|S}}\sqrt{\hat{\sigma}_{jj|S}}},$$

where $\hat{\sigma}_{ii|S} = \hat{\sigma}_{ii} - \hat{\sigma}_{Si}^T \hat{\sigma}_{SS}^{-1} \hat{\sigma}_{Si}$, $\hat{\sigma}_{ij|S} = \hat{\sigma}_{ij} - \hat{\sigma}_{Si}^T \hat{\sigma}_{SS}^{-1} \hat{\sigma}_{Sj}$. And $\hat{\sigma}_{Si} = (\mathrm{x}_S - \bar{\mathrm{x}}_S)^T (\mathrm{x}_i - \bar{\mathrm{x}}_i)/n$, $\hat{\sigma}_{SS} = (\mathrm{x}_S - \bar{\mathrm{x}}_S)^T (\mathrm{x}_S - \bar{\mathrm{x}}_S)/n$, $\bar{\mathrm{x}}_S = (\bar{\mathrm{x}}_{i_1} \mathbb{1}_n, \ldots, \bar{\mathrm{x}}_{i_{|S|}} \mathbb{1}_n)$.

2.2. Undirected decomposable graphs

Denote an undirected graph by $G = (V, E)$ with a vertex set $V = \{1, 2, \ldots, p\}$ and an edge set $E = \{(r, s) : e_{rs} = 1, 1 \le r < s \le p\}$ with $e_{rs} = 1$ if the edge (r, s) is present in G and 0 otherwise. For the purpose of a self-contained exposition, we first review some basic terminologies of graph theory. A *path* of length k in G from vertex u to v is a sequence of $k - 1$ distinct vertices of the form $u = v_0, v_1, \ldots, v_{k-1}, v_k = v$ such that $(v_{i-1}, v_i) \in E$ for all $i = 1, 2, \ldots, k$. The path is a *k-cycle* if the end points are the same, $u = v$. If there is a path from u to v, then we say u and v are *connected*. A subset $S \subseteq V$ is said to be an *uv-separator* if all paths from u to v intersect S. The subset S is said to *separate* A from B if it is an uv-separator for every $u \in A$, $v \in B$. A *chord* of a cycle is a pair of vertices that are not consecutive on the cycle, but are adjacent in G. A graph is *complete* if all vertices are joined by an edge. A *clique* is a complete subgraph that is maximal, i.e., maximally complete subgraph. See [27] for more graph related terminologies.

We shall focus on decomposable graphs in this paper. A graph is decomposable [27] if and only if its every cycle of length greater than or equal to four possesses a chord. A decomposable graph G can be represented by a perfect ordering of its cliques and separators. Refer to [27] for formal definitions of a clique and a separator, and other equivalent representations. An ordering of cliques $C_i \in \mathcal{C}$ and separators $S_i \in \mathcal{S}$, where $\mathcal{C} = \{C_i\}_{i=1}^k$ and $\mathcal{S} = \{S_i\}_{i=2}^k$, $(C_1, S_2, C_2, S_3, \ldots C_k)$, is said to be perfect if for every $i = 2, 3, \ldots, k$ the running intersection property ([27], page 15) is fulfilled, meaning that there exists a $j < i$ such that $S_i = C_i \cap H_{i-1} \subset C_j$ where $H_{i-1} = \bigcup_{j=1}^{i-1} C_j$. A junction tree for the

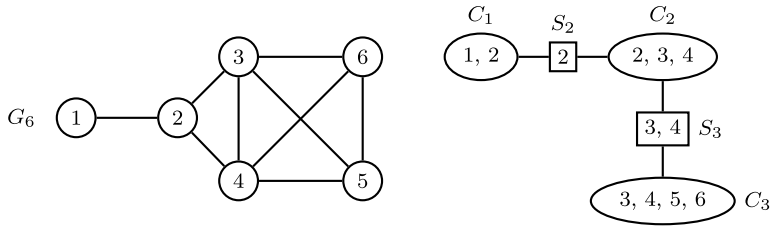


Figure 1. G_6 is a 6-node decomposable graph and its junction tree decomposition (right) has 3 *cliques* and 2 separators, that is, $C_1 = \{1, 2\}$, $S_2 = \{2\}$, $C_2 = \{2, 3, 4\}$, $S_3 = \{3, 4\}$, $C_3 = \{3, 4, 5, 6\}$.

decomposable graph G is a tree representation of the cliques. (For a non-decomposable graph, the junction tree consists of its prime components that are not necessarily cliques, i.e., not complete). A tree with a set of vertices equal to the set of cliques of G is said to be a junction tree if, for any two cliques C_i and C_j and any clique C on the unique path between C_i and C_j, we have $C_i \cap C_j \subset C$. A set of vertices shared by two adjacent nodes of the junction tree is complete and defines the separator of the two subgraphs induced by the two adjacent nodes. Figures 1 and 2 briefly illustrate a decomposable and a non-decomposable graph, both defined on 6 nodes.

2.3. Miscellaneous notations

For an $n \times p$ matrix Y, Y_C is defined as the submatrix of Y consisting of columns with indices in the clique C. Let $(\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_p) = (Y_1, Y_2, \ldots, Y_n)^T$, where y_i is the ith column of $\mathrm{Y}_{n\times p}$. If $C = \{i_1, i_2, \ldots, i_{|C|}\}$, where $1 \le i_1 < i_2 < \cdots < i_{|C|} \le p$, then $\mathrm{Y}_C = (\mathrm{y}_{i_1}, \mathrm{y}_{i_2}, \ldots, \mathrm{y}_{i_{|C|}})$. For any square matrix $A = (a_{ij})_{p\times p}$, define $A_C = (a_{ij})_{|C|\times|C|}$ where $i, j \in C$, and the order of entries carries into the new submatrix A_C. Therefore, $\mathrm{Y}_C^T\mathrm{Y}_C = (\mathrm{Y}^T\mathrm{Y})_C$. $\mathrm{MN}_{m\times n}(M, \Sigma_r, \Sigma_c)$ is an $m \times n$ matrix normal distribution with mean matrix M, Σ_r and Σ_c as covariance matrices between rows and columns, respectively.

Let $\mathbb{P}$ be the probability corresponding to the true data generating distribution. Denote $\mathcal{G}_k$ and $\mathcal{D}_k$ as the k-dimensional graph space and the k-dimensional decomposable graph space. Let $\mathcal{M}_t$ be the minimal triangulation space of G_t when G_t is non-decomposable. $a \asymp b$ denotes $C_1 b \le a \le C_2 b$ for positive constants C_1 and C_2. $a \precsim b$ denotes $a \le C_3 b$ for a positive constant C_3. $a \succsim b$ denotes $a \ge C_4 b$ for a positive constant C_4. For set relations, $A \subset B$ means A is a subset of B; $A \subsetneq B$ means $A \subset B$ and $A \ne B$; $A \not\subset B$ means A is not a subset of B. $|\cdot|$ determined by context can be the absolute value of a

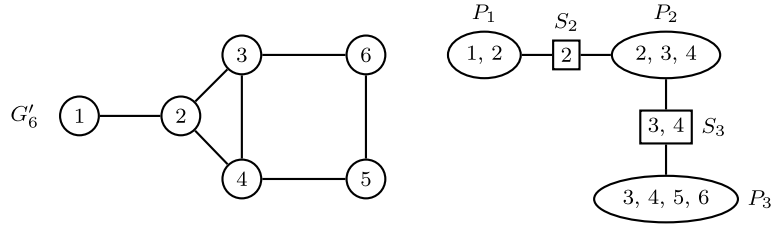


Figure 2. G_6' is a 6-node non-decomposable graph because it has a cycle of length four, $3 - 4 - 5 - 6 - 3$, that does not have a cord. Its junction tree decomposition (right) has 3 *prime* components and 2 separators, that is, $P_1 = \{1, 2\}$, $S_2 = \{2\}$, $P_2 = \{2, 3, 4\}$, $S_3 = \{3, 4\}$, $P_3 = \{3, 4, 5, 6\}$. Out of three prime components only P_1 and P_2 are cliques.

real number, the cardinality of a set or the determinant of a matrix. $\pi(\cdot)$ and $\pi(\cdot \mid \mathrm{Y})$ are the prior and the posterior distribution of graphs, respectively. Refer also to Table 2 for a detailed list of notations used in the theorem statements and the proofs.

3. Bayesian hierarchical model for graph selection

Suppose we observe independent and identically distributed p-dimensional Gaussian random variables $Y_i, i = 1, \ldots, n$. To describe the common distribution of Y_i, define a $p \times p$ covariance matrix Σ_G that depends on an undirected decomposable graph as defined in Section 2.2. Assume $Y_i \mid \Sigma_G, G \sim \mathrm{N}_p(0, \Sigma_G)$. In matrix notations,

$$\mathrm{Y}_{n\times p} \mid \Sigma_G, G \sim \mathrm{MN}_{n\times p}(\mathbf{0}_{n\times p}, I_n, \Sigma_G),$$

where $\mathrm{Y}_{n\times p} = (Y_1, Y_2, \ldots, Y_n)^T$ and $\mathbf{0}_{n\times p}$ is an $n \times p$ matrix with all zeros. The prior used here for covariance matrix Σ_G given a decomposable graph G is the hyper-inverse Wishart prior, described below. We emphasize here that although we restrict our model and prior to decomposable graphs only, our results in Section 5 allow for model misspecification.

3.1. The hyper-inverse Wishart distribution

Denoted by $\mathrm{HIW}_G(b, D)$ [10,11] the hyper-inverse Wishart (HIW) distribution is a distribution on the cone of $p \times p$ positive definite matrices with $b > 2$ degrees of freedom [24] and a fixed $p \times p$ positive definite matrix D such that the joint density factorizes on the junction tree of the given decomposable graph G as

$$p(\Sigma_G \mid b, D) = \frac{\prod_{C\in\mathcal{C}} p(\Sigma_C \mid b, D_C)}{\prod_{S\in\mathcal{S}} p(\Sigma_S \mid b, D_S)}, \tag{3.1}$$

where for each $C \in \mathcal{C}$, $\Sigma_C \sim \mathrm{IW}_{|C|}(b, D_C)$ with density

$$p(\Sigma_C \mid b, D_C) \propto |\Sigma_C|^{-(b+2|C|)/2} \operatorname{etr}\left\{-\frac{1}{2}\Sigma_C^{-1} D_C\right\},$$

where $|C|$ is the cardinality of the clique C and $\operatorname{etr}(\cdot) = \exp\{\operatorname{tr}(\cdot)\}$. $\mathrm{IW}_p(b, D)$ is an inverse Wishart distribution with b degrees of freedom and a fixed $p \times p$ positive definite matrix D. Its normalizing constant is

$$\left|\frac{1}{2}D\right|^{(b+p-1)/2} \Gamma_p^{-1}\left(\frac{b+p-1}{2}\right),$$

where $\Gamma_p(\cdot)$ is a multivariate gamma function. Refer to [10] for more details about this parametrization of the inverse Wishart distribution.

3.2. Bayesian inference on decomposable graphs

Since the joint density factorizes over cliques and separators in the same way as in (3.1), the likelihood function can be rewritten as

$$f(\mathrm{Y} \mid \Sigma_G) = (2\pi)^{-\frac{np}{2}} \frac{\prod_{C\in\mathcal{C}} |\Sigma_C|^{-\frac{n}{2}} \operatorname{etr}(-\frac{1}{2}\Sigma_C^{-1}\mathrm{Y}_C^T\mathrm{Y}_C)}{\prod_{S\in\mathcal{S}} |\Sigma_S|^{-\frac{n}{2}} \operatorname{etr}(-\frac{1}{2}\Sigma_S^{-1}\mathrm{Y}_S^T\mathrm{Y}_S)}. \tag{3.2}$$

From (3.1), the prior on Σ_G is

$$f(\Sigma_G \mid G) = \frac{\prod_{C\in\mathcal{C}} p(\Sigma_C \mid b, D_C)}{\prod_{S\in\mathcal{S}} p(\Sigma_S \mid b, D_S)} = \frac{\prod_{C\in\mathcal{C}} |\frac{1}{2}D_C|^{\frac{b+|C|-1}{2}} \Gamma_{|C|}^{-1}(\frac{b+|C|-1}{2}) |\Sigma_C|^{-\frac{b+2|C|}{2}} \operatorname{etr}(-\frac{1}{2}\Sigma_C^{-1}D_C)}{\prod_{S\in\mathcal{S}} |\frac{1}{2}D_S|^{\frac{b+|S|-1}{2}} \Gamma_{|S|}^{-1}(\frac{b+|S|-1}{2}) |\Sigma_S|^{-\frac{b+2|S|}{2}} \operatorname{etr}(-\frac{1}{2}\Sigma_S^{-1}D_S)}.$$

It is straightforward to obtain the marginal likelihood of the decomposable graph G,

$$f(\mathrm{Y} \mid G) = (2\pi)^{-\frac{np}{2}} \frac{h(G, b, D)}{h(G, b+n, D+\mathrm{Y}^T\mathrm{Y})} = (2\pi)^{-\frac{np}{2}} \frac{\prod_{C\in\mathcal{C}} w(C)}{\prod_{S\in\mathcal{S}} w(S)},$$

where

$$h(G, b, D) = \frac{\prod_{C\in\mathcal{C}} |\frac{1}{2}D_C|^{\frac{b+|C|-1}{2}} \Gamma_{|C|}^{-1}(\frac{b+|C|-1}{2})}{\prod_{S\in\mathcal{S}} |\frac{1}{2}D_S|^{\frac{b+|S|-1}{2}} \Gamma_{|S|}^{-1}(\frac{b+|S|-1}{2})},$$

$$w(C) = \frac{|D_C|^{\frac{b+|C|-1}{2}} |D_C + \mathrm{Y}_C^T\mathrm{Y}_C|^{-\frac{b+n+|C|-1}{2}}}{2^{-\frac{n|C|}{2}} \Gamma_{|C|}(\frac{b+|C|-1}{2}) \Gamma_{|C|}^{-1}(\frac{b+n+|C|-1}{2})}.$$

Throughout the remainder of the paper, we shall be working with the hyper-inverse Wishart g-prior [10], denoted as

$$\Sigma_G \mid G \sim \mathrm{HIW}_G\big(b, g\mathrm{Y}^T\mathrm{Y}\big), \tag{3.3}$$

where g is some suitably small fraction in $(0, 1)$ and $b > 2$ is a fixed constant. Following the recommendation in [10], we choose $g = 1/n$ for the rest of the paper. Intuitively, this choice of g avoids overwhelming the likelihood asymptotically as well as arbitrarily diffusing the prior. In that case,

$$w(C) = \frac{(n+1)^{-\frac{|C|(b+n+|C|-1)}{2}} |\mathrm{Y}_C^T\mathrm{Y}_C|^{-\frac{n}{2}}}{(2n)^{-\frac{n|C|}{2}} \Gamma_{|C|}(\frac{b+|C|-1}{2}) \Gamma_{|C|}^{-1}(\frac{b+n+|C|-1}{2})}.$$

The choice of focusing on the HIW g-prior in this paper is driven by the following two reasons. First, we can simplify the signal strength assumption in terms of the smallest nonzero entry in the partial correlation matrix, which serves as a natural interpretation of the edge strength compared to assumptions on the eigenvalues of the correlation matrix. Second, we conjecture that the results stated in Section 4 and Section 5 continue to hold for any choice of HIW priors. The proof techniques under HIW g-priors serve as representations to the principle ideas in the article and can be easily adapted to other variations of HIW priors.

To complete a fully Bayesian specification, we place a prior distribution $\pi(\cdot)$ on the decomposable graph G. Our theoretical results in Section 4 and Section 5 are independent of the prior choice on G if we consider a fixed p asymptotics. However, for p increasing with n we need a suitable penalty on the number of edges of the random graph to penalize the false positives. Here is a popular example [8,10,14,24,38] we use in the paper. Considering an undirected decomposable graph G, we assume the edges are independently drawn from a Bernoulli distribution with a common probability q,

$$\pi(G \mid q) \propto \left[\prod_{r<s} q^{e_{rs}} (1-q)^{1-e_{rs}}\right] \cdot \mathbb{1}_{\mathcal{D}_p}(G).$$

Here, q is the prior edge inclusion probability. We control the parameter q to induce sparsity on the number of edges. [24] recommends using $2/(p-1)$ as the hyper-parameter for the Bernoulli distribution in practice. For an undirected graph, it has its peak at p edges and the mode is smaller for decomposable graphs. We outline specific choices in Section 4 and Section 5 below.

To summarize, the Bayesian hierarchical model for Gaussian graphical models with decomposable graphs is as follows.

$$\begin{aligned} \mathrm{Y}_{n\times p} \mid \Sigma_G, G &\sim \mathrm{MN}_{n\times p}(\mathbf{0}_{n\times p}, I_n, \Sigma_G), \\ \Sigma_G \mid G &\sim \mathrm{HIW}_G\big(b, g\mathrm{Y}^T\mathrm{Y}\big), \\ \pi(G \mid q) &\propto \left[\prod_{r<s} q^{e_{rs}}(1-q)^{1-e_{rs}}\right] \cdot \mathbb{1}_{\mathcal{D}_p}(G). \end{aligned}$$

This model and its numerous variations have been studied extensively during the past few decades and many efficient algorithms have been proposed for posterior sampling [1,20,24]. The most commonly used one is the add-delete Metropolis–Hastings sampler which is based on the reversible jump Markov chain Monte Carlo (MCMC) algorithm proposed in [20,21]. The algorithm samples decomposable graphs through local updates to traverse through $\mathcal{D}_p$. To highlight the key steps of the MCMC procedure, let G denote the current decomposable graph at a certain step of the MCMC. Then

- Propose a new decomposable graph G' by adding or deleting an edge with equal probability from the current decomposable graph G while maintaining decomposability. This can be achieved by applying the procedure in [20].
- Calculate the acceptance probability

$$\alpha\big(G', G\big) = \min\left\{1, \frac{f(\mathrm{Y} \mid G')\pi(G')}{f(\mathrm{Y} \mid G)\pi(G)}\right\}.$$

- Let G_{new} be the updated decomposable graph. Sample $\Sigma_{G_{\text{new}}}$ from its posterior distribution

$$\Sigma_{G_{\text{new}}} \mid \mathrm{Y}, \quad G_{\text{new}} \sim \mathrm{HIW}_{G_{\text{new}}}\big\{b+n, (g+1)\mathrm{Y}^T\mathrm{Y}\big\}.$$

  This step can be performed by using the sampling procedure in [9].

4. Theoretical results in the well-specified case

In this section, we present our main consistency results when the true graph is assumed to be decomposable. The assumptions and theorems introduced here serve as the foundation for the consistency results when the model is misspecified. Understanding the mechanism of how the results are derived for decomposable graphs is crucial to develop the theoretical tools when the true graph is nondecomposable. This section provides an alternative perspective of selection consistency problems for perfect DAGs [8] based on a set of new assumptions. A detailed comparison of assumptions with those in [8] is given in Section 4.4. The proofs are deferred to the Appendix and Supplementary Materials [32].

Before introducing the assumptions, we adapt previous notations to the high-dimensional graph selection problem. Let $\mathrm{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ and $\Omega_0 = \Sigma_0^{-1}$ be the corresponding precision matrix. Without loss of generality, we assume all column means of Y are zero. Let $\rho_{ij|V\setminus\{i,j\}}$ denote the true partial correlation between nodes i and j given the rest of the nodes $V\setminus\{i, j\}$. Let ρ_L and ρ_U be the

smallest and largest (in absolute value) *non-zero* population partial correlations, that is,

$$\rho_L = \min_{\substack{1\le i<j\le p \\ (i,j)\in E_t}} |\rho_{ij|V\setminus\{i,j\}}|, \qquad \rho_U = \max_{\substack{1\le i<j\le p \\ (i,j)\in E_t}} |\rho_{ij|V\setminus\{i,j\}}|.$$

Furthermore, assume all partial correlations (with the form $\rho_{ij|V\setminus\{i,j\}}$) are uniformly bounded above by 1 to avoid degeneracy in the Gaussian distribution. More precisely, we assume $0 < \rho_L \le \rho_U < 1$.

Let $G_t = (V, E_t)$ denote the true decomposable graph induced by Ω_0. Let $G_a = (V, E_a)$ be any alternative decomposable graph other than the true graph G_t. $E_a^1 = E_t \cap E_a$ denotes the set of true edges in G_a. Notice, when $E_t \subsetneq E_a$, we have $E_a^1 = E_t$. Letting $|\cdot|$ denote the cardinality of a set, $|E_t|$ is the number of edges in G_t (number of true edges) and $|E_a^1|$ is the number of true edges in G_a. Define $G_c = (V, E_c)$, where $E_c = \{(i, j) : e_{ij} = 1, 1 \le i < j \le p\}$ and $|E_c| = p(p-1)/2$, to be the complete graph. By definition G_c is a decomposable graph. We use $G_a \ne G_t$ to denote $E_a \ne E_t$, $G_a \not\subset G_t$ for $E_a \not\subset E_t$, and $G_a \subsetneq G_t$ for $E_a \subsetneq E_t$. In the following, we introduce a set of general assumptions essential to establish the theoretical results of the graph selection consistency. The main results will have additional restrictions on the parameters $(\alpha, \lambda, \sigma, \gamma)$ introduced below in the general assumptions.

Assumption 4.1 (Graph size).

$$p \precsim n^{\alpha} \quad \text{where } 0 < \alpha < 1.$$

This assumption states that the dimension p of graphs has to grow slower than n, unlike [8] (hereafter CKG19) where p is allowed to grow up to a sub-exponential rate in n. This is simply because the sparsity assumptions in our paper are completely different from those in CKG19. As clarified below, we allow graphs to be "dense" while graphs considered by CKG19 are sparse in a systematic way. Similar to CKG19, [28] (hereafter LLL19) allows p to be sub-exponential in n but with some compromise in allowing dense graphs.

Assumption 4.2 (Signal strength and identifiability).

$$\rho_L \succsim n^{-\lambda} \quad \text{where } 0 \le \lambda < \frac{1}{2}.$$

This assumption indicates that the smallest nonzero partial correlation (in absolute value) cannot converge to zero faster than $1/\sqrt{n}$. Interpreting ρ_L as the "signal size", the assumption restricts the smallest signal size such that the hierarchical model in Section 3 is identifiable. We compare this assumption with analogous assumptions from CKG19 and LLL19 in Section 4.4.

Assumption 4.3 (Maximum number of edges in G_t).

$$|E_t| \precsim n^{\sigma} \quad \text{where } 0 < \sigma \le 2\alpha.$$

In the well-specified case (Theorems 4.2 and 4.3), the maximum number of edges in G_t is not restricted, that is, G_t is allowed to be complete, meaning that we do not impose any sparsity conditions on the true graph. For model misspecification, some restrictions on $|E_t|$ are needed to ensure the selection consistency. CKG19 and LLL19 mainly focus on ultra high-dimensional cases ($p > n$ and $p \gg n$) with sparsity assumptions on G_t which apply to the $p < n$ case as well.

Assumption 4.4 (Prior edge inclusion probability).

$$q \asymp e^{-C_q n^{\gamma}} \quad \text{where } 0 < \gamma < 1, 0 < C_q < \infty.$$

The assumption states that $-\log(q)$ has to be a fractional power in n and cannot increase faster than n. A similar assumption on prior edge inclusion probability can be found in CKG19. A detailed comparison of assumptions on q is provided in Section 4.4. On the other hand, LLL19 uses a different prior specification, called the empirical sparse Cholesky (ESC) prior which has close connections with the prior setting in CKG19. However, the ESC prior does not impose elementwise sparsity with independent Bernoulli distributions, thus it loses the interpretation of the prior edge inclusion probability.

Assumption 4.5 (Imperfect linear relationship).

$$1 - \rho_U \gtrsim n^{-k} \quad \text{where } k \geq 0.$$

In general, ρ_U is strictly less than 1 to avoid degeneracy in the Gaussian distribution. However, it is allowed to grow with n to 1 at any polynomial rate (or slower). This limitation on the growth is due to a technical condition on the tail behavior of the distribution of non-zero sample partial correlations. See Section S.1 in the Supplementary Materials [32] for more details.

4.1. Pairwise Bayes factor consistency for fixed p

In this section, we investigate the behavior of the pairwise Bayes factor

$$\text{BF}(G_a; G_t) = \frac{f(\text{Y} \mid G_a)}{f(\text{Y} \mid G_t)}, \tag{4.1}$$

for fixed p, ρ_L and ρ_U. We aim to derive sufficient conditions on the likelihood (3.2) and the prior on Σ_G given by (3.3) such that the Bayes factor (4.1) converges to 0 as $n \to \infty$ for any graph $G_a \neq G_t$.

Theorem 4.1 (Upper bounds for pairwise Bayes factors). *Let p be fixed. Given any decomposable graph $G_a \neq G_t$, there exists a set Δ_a, such that on the set Δ_a, if $n > \max\{p + b, 4p\}$, we have*

1. *when $G_t \not\subset G_a$,*

$$\text{BF}(G_a; G_t) < \exp\left\{-\frac{n\rho_L^2}{2} + \delta(n)\right\},$$

2. *when $G_t \subsetneq G_a$,*

$$\text{BF}(G_a; G_t) < \left(e^{p^2}\right) \cdot n^{-\frac{1}{2}(|E_a| - |E_t|)(1 - 2/\tau^*)},$$

and

$$\mathbb{P}(\Delta_a) \geq 1 - \frac{42p^2}{(1-\rho_U)^2}(n-p)^{-\frac{1}{4\tau^*}} \left\{\frac{1}{\tau^*}\log(n-p)\right\}^{-\frac{1}{2}},$$

where $\tau^* > 2$ and $\delta(n) = p^2 \log n + \sqrt{n \log n} + 3p^2 \log p$ satisfying $\delta(n)/n \to 0$, as $n \to \infty$.

When $G_t \not\subset G_a$, G_a lacks at least one true edge in G_t. The first part of Theorem 4.1 says that the Bayes factor in favor of true-edge deletions is exponentially small with the actual rate depending on

ρ_L. When $G_t \subsetneq G_a$, G_a contains all true edges in G_t and at least one false edge. The Bayes factor in favor of false-edge additions decreases to zero only at a polynomial rate depending on the number of false edges in G_a.

The behavior of Bayes factors observed from Theorem 4.1 depends solely on the property of the HIW prior as no prior on the graph space is involved and no sparsity condition on G_t is imposed. The HIW prior enforces a stronger penalty on true-edge deletions than false-edge additions. In high-dimensional cases, as the complexity of the graph space grows, the polynomial penalty on false-edge additions is not enough to overcome the complexity of the graph space. In such situations, a prior on the graph space is needed to achieve the selection consistency (refer to Theorems 4.2 and 4.3). In high-dimensional cases, the HIW prior alone does not favor parsimonious models, that is, it does not guarantee the selection of sparse graphs when G_t is sparse. The next two corollaries are direct consequences of Theorem 4.1.

Although Theorem 4.1 holds for finite dimensional graphs, its proof techniques serve as the foundation to the subsequent theorems on decomposable graphs presented in this article. Therefore, it is important to discuss the key steps of the proof. First, to find an upper bound of Bayes factors between any two decomposable graphs, the Bayes factors are decomposed in terms of "local moves" consisting of adding or deleting edges between two decomposable graphs. Hence, the Bayes factor between any two decomposable graphs which differ by one edge can be seen as the basic unit in the proof. Lemmas A.1–A.4 in Appendix A provide the theoretical support for this procedure. Second, we enumerate all such Bayes factors and express them as functions of sample partial correlations specific to each Bayes factor. Finally, in Section S.1 we develop results on the tail behavior of sample partial correlations, which enable us to replace the sample partial correlations with the upper or lower bound defined in Assumptions 4.2 and 4.5.

Corollary 4.1 (Finite graph pairwise Bayes factor consistency). *Let G_a be any decomposable graph and $G_a \neq G_t$. If the graph dimension p is a fixed constant,* $\mathrm{BF}(G_a; G_t) \xrightarrow{\mathbb{P}} 0$, *as* $n \to \infty$.

Corollary 4.2 (Finite graph strong selection consistency). *Let p be fixed. Define*

$$\widetilde{\pi}(G_t \mid \mathrm{Y}) = \frac{1}{1 + \sum_{G_a \neq G_t} \mathrm{BF}(G_a; G_t)}. \tag{4.2}$$

We have $\widetilde{\pi}(G_t \mid \mathrm{Y}) \xrightarrow{\mathbb{P}} 0$, *as* $n \to \infty$.

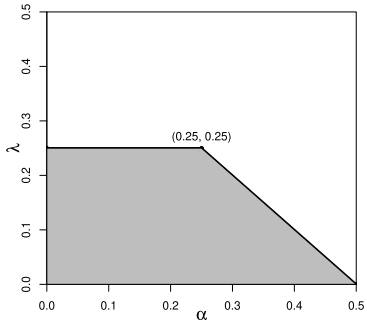
Notice, the number of Bayes factor terms in the denominator on the right-hand side of (4.2) is finite. Thus, the strong selection consistency is trivial given Theorem 4.1.

4.2. Posterior ratio consistency for growing p

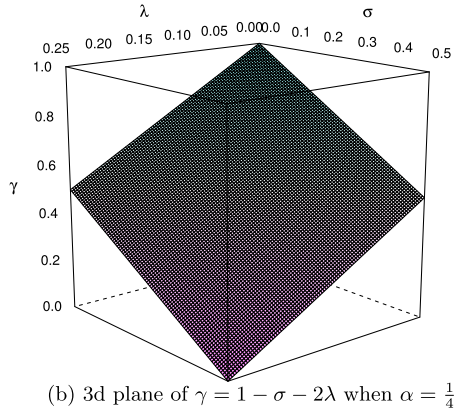
Next, we examine the convergence of the following posterior ratio,

$$\mathrm{PR}(G_a; G_t) = \frac{f(\mathrm{Y} \mid G_a)\pi(G_a)}{f(\mathrm{Y} \mid G_t)\pi(G_t)},$$

when the dimension of graphs p grows with the sample size n.



(a) the region of (α, λ) described by (4.3)



(b) 3d plane of $\gamma = 1 - \sigma - 2\lambda$ when $\alpha = \frac{1}{4}$

Figure 3. Possible range of $(\alpha, \lambda, \sigma, \gamma)$ required for the high-dimensional graph posterior ratio consistency. (a) The gray area shows all possible values of (α, λ) in the assumptions of Theorem 4.2. The range of λ is $(0, 1/4)$ for all $\alpha \in (0, 1/4)$. As α increases passing $1/4$, the range of λ becomes narrower. (b) For $\alpha = 1/4$, then $\lambda \in [0, 1/4)$ and $\sigma \in (0, 1/2]$. The plotted plane is the upper bound of γ in Theorem 4.2, that is, $\gamma = 1 - \sigma - 2\lambda$. The lower bound of γ is the bottom plane $\gamma = 0$ in this plot. The area between the upper and bottom planes contains all possible values of $(\lambda, \sigma, \gamma)$ when $\alpha = 1/4$.

Theorem 4.2 (High-dimensional graph posterior ratio consistency). *Let G_a be any decomposable graph and $G_a \neq G_t$. If Assumptions 4.1–4.5 are satisfied with*

$$0 < \alpha < \frac{1}{2}, \qquad 0 \leq \lambda < \min\left\{\frac{1}{4}, \frac{1-2\alpha}{2}\right\}, \tag{4.3}$$

by choosing γ in the interval $(0, 1 - \sigma - 2\lambda)$, we have $\mathrm{PR}(G_a; G_t) \xrightarrow{\mathbb{P}} 0$, as $n \to \infty$.

We provide some intuitions of the constraints imposed on $(\alpha, \lambda, \sigma, \gamma)$ in Theorem 4.2. For posterior ratio consistency, the graph size p cannot grow at a rate equal to or faster than $\sqrt{n}$. In Figure 3(a), if $\alpha > 1/4$ and increases to $1/2$, the range of λ becomes narrower. This means the rate at which ρ_L converges to zero needs to be slower to achieve the selection consistency. Therefore, for larger dimensional graphs a stronger identifiability is needed for consistent recovery of the true graph.

From Theorem 4.1, recall that HIW priors do not enforce a strong penalty on false-edge additions. When the graph size grows with n, a prior on the number of edges is needed. The penalty on the number of edges is controlled by γ, that is, larger γ means smaller edge inclusion probability and larger penalty on dense graphs, and vice versa. Only the upper bound of γ is restricted by σ and λ, see Figure 3(b). The upper bound of γ decreases when σ increases (the total number of edges $|E_t|$ increases). This ensures the consistency when G_t is dense.

The upper bound of γ also decreases when λ increases (ρ_L converges to zero faster). As ρ_L goes to zero faster, the identifiability issue becomes more severe. Based on Theorem 4.1, the risk of false-edge additions (polynomial) is lower than the risk of true-edge deletions (exponential). In order to overcome the identifiability issue, increasing the edge inclusion probability would result in a lower risk. This explains the decrease in the upper bound of γ when λ increases. Furthermore, there is no restriction on σ in Assumption 4.3 suggesting that no sparsity assumption on the true graph is needed for the consistency.

4.3. Strong graph selection consistency for growing p

In this section, we examine the behavior of

$$\pi(G \mid \mathrm{Y}) = \frac{f(\mathrm{Y} \mid G)\pi(G)}{\sum_{G' \in \mathcal{D}_p} f(\mathrm{Y} \mid G')\pi(G')},$$

as $n, p \to \infty$.

Theorem 4.3 (Strong graph selection consistency). *Let G_a be any decomposable graph and $G_a \neq G_t$. If Assumptions 4.1–4.5 are satisfied with*

$$0 < \alpha < \frac{1}{3}, \qquad 0 \leq \lambda < \min\left\{\frac{1-\alpha}{4}, \frac{1-3\alpha}{2}\right\}, \tag{4.4}$$

by choosing γ in the interval $(\alpha, 1-\sigma-2\lambda)$, we have

$$\pi(G_t \mid \mathrm{Y}) \xrightarrow{\mathbb{P}} 1 \quad \text{as } n \to \infty.$$

Strong selection consistency demands all posterior ratios to be converging simultaneously at a sufficiently fast rate so that the sum is convergent. Since the number of alternative graphs is of the order 2^{p^2}, to make the summation convergent, we require further assumptions on the model complexity and an accompanying stronger penalty on the number of edges. We achieve this by shrinking the dimension of the graph space ($\alpha < 1/3$) and inducing a slightly stronger sparsity (by selecting larger γ) on the prior over the graph space. As α increases, the upper bound of λ (the fastest rate allowed for ρ_L to converge to zero) becomes more restrictive and so does its feasible range. This alleviates the identifiability problem caused by the growth of p.

The upper bound of γ is the same as in Theorem 4.2, but λ has a different range in Theorem 4.3, see Figure 4(b). While there is no specific lower bound of γ in Theorem 4.2, there is a lower bound on γ increasing with α in Theorem 4.3. This ensures when p grows with n, the upper bound of the edge inclusion probability q becomes smaller imposing a stronger penalty coming from the prior on the graph space. Finally, as in Theorem 4.2 we do not need any further restriction on σ in Assumption 4.3 meaning that the true graph is allowed to be dense or even complete in the strong selection consistency.

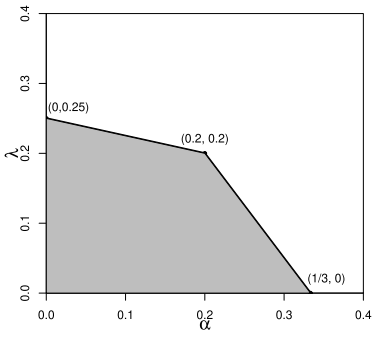
In practice, one might be interested in a consistent point estimate rather than the entire posterior distribution. In Bayesian inference for discrete configurations, the posterior mode provides a natural point estimate. In the following, we investigate the consistency of the posterior mode obtained from our Bayesian hierarchical model as a simple by-product of Theorems 4.2 and 4.3. Define $\hat{G}$ to be the posterior mode in the decomposable graph space, that is,

$$\hat{G} = \underset{G \in \mathcal{D}_p}{\operatorname{argmax}}\, \pi(G \mid \mathrm{Y}).$$

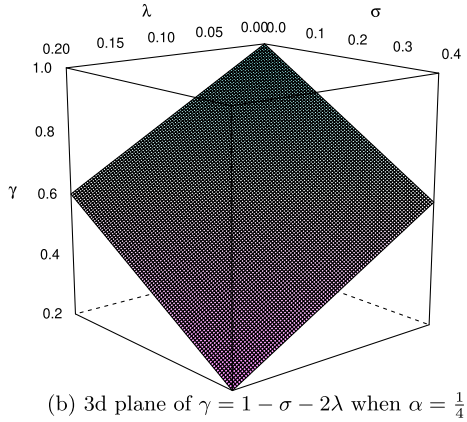
Then the following is true.

Corollary 4.3 (Posterior mode consistency when G_t is decomposable). *Under the assumptions of Theorem 4.3, the probability which the posterior mode $\hat{G}$ is equal to the true graph G_t goes to one, that is,*

$$\mathbb{P}(\hat{G} = G_t) \to 1 \qquad \text{as } n \to \infty.$$



(a) the region of (α, λ) described by (4.4)



(b) 3d plane of $\gamma = 1 - \sigma - 2\lambda$ when $\alpha = \frac{1}{4}$

Figure 4. Possible range of $(\alpha, \lambda, \sigma, \gamma)$ required for the high-dimensional strong graph selection consistency. (a) The gray area shows all possible values of (α, λ) in Theorem 4.3. For $\alpha \in (0, 1/5]$, the range of λ narrows slowly with a rate of $1/4$; for $\alpha \in (1/5, 1/3)$, the upper bound of λ decreases faster at a rate of $3/2$. (b) For $\alpha = 1/5$, then $\lambda \in [0, 1/5)$ and $\sigma \in (0, 2/5]$. The plotted plane is the upper bound of γ in Theorem 4.3, that is, $\gamma = 1 - \sigma - 2\lambda$. The lower bound of γ is the bottom plane $\gamma = 0.2$ in this plot. The area between the upper and bottom planes contains all possible values of $(\lambda, \sigma, \gamma)$ when $\alpha = 1/5$.

4.4. Comparing assumptions with CKG19 and LLL19

The results in Section 4.2 and Section 4.3 are established with certain restrictions on the parameters $(\alpha, \lambda, \sigma, \gamma)$ in Assumptions 4.1–4.5. Recently, CKG19 and LLL19 developed the DAG selection consistency with the DAG-Wishart prior [5] and the ESC prior [28], respectively. For any undirected decomposable graph there exists a directed acyclic graph whose skeleton is the same as the decomposable graph [8]. Also, the DAG-Wishart prior is equivalent to the hyper-inverse Wishart prior if one considers decomposable graphs. The ESC prior is related to the DAG-Wishart prior from an autoregressive perspective of Gaussian DAG models [28]. Thus, most assumptions in LLL19 can be related to CKG19. Due to the similarities between the hierarchical models in CKG19 and this paper, we mainly focus on comparing the differences of assumptions between these two papers. For a fair comparison of the assumptions, we consider p to be increasing only at a fractional exponent in n.

4.4.1. *Assumptions on eigenvalues of* Ω_0

In CKG19, the lower bound of the smallest eigenvalue of the true precision matrix Ω_0 has to decrease at a rate slower than $n^{-1/8}$ and the upper bound of the largest eigenvalue cannot grow faster than $n^{1/8}$. In LLL19, only a fixed constant and its reciprocal are used to restrict the smallest and largest eigenvalues of the true precision matrix, but it can be relaxed to a similar assumption as in CKG19 with some scarification of other conditions. We do not need any assumptions on the eigenvalues of the true precision matrix. Instead, we only restrict the smallest nonzero partial correlation ρ_L in absolute value. This assumption will be further examined below.

In general, there is no association between eigenvalues and partial correlations of a positive definite matrix. Here is a simple example to demonstrate this. Consider a $(p+2)$-dimensional positive definite precision matrix

$$\begin{bmatrix} A & \mathbf{0}_{2\times p} \\ \mathbf{0}_{p\times 2} & I_p \end{bmatrix} \quad \text{where } A = \begin{bmatrix} a & b \\ b & a \end{bmatrix} \text{ and } a^2 - b^2 > 0, b \neq 0.$$

The two nontrivial eigenvalues are $a \pm b$ and the only non-zero partial correlation is b/a. For example, let $a = n^s$ and $b = n^t$, where $s > t > 0$. Setting $s - t < 1/2$, the partial correlation does not go to zero faster than $1/\sqrt{n}$. But in this case, both eigenvalues $a \pm b = O(n^s)$ can diverge to infinity at any arbitrary polynomial rate. On the other hand, restricting the eigenvalues of Ω_0 does not guarantee ρ_L can be controlled. Let $a = n^{-s}$ and $b = n^{-t}$, where $t > s > 0$. If we restrict the eigenvalues to decrease slower than $n^{-1/8}$ with $s < 1/8$. In this case, $t - s > t - 1/8$ which means the partial correlation can converge to zero at any arbitrary polynomial rate. We state that there exist examples which eigenvalues and partial correlations are related as well. We omit the details for simplicity.

4.4.2. *Assumptions on the sparsity of G_t*

Considering decomposable graphs, CKG19 restricts the number of edges directed from any node of the true graph to be less than $n^{1/4}$. Assumption 4 in CKG19 involving the signal size and eigenvalue restrictions limits this number even further. But Theorems 4.2 and 4.3 do not enforce any sparsity assumptions on G_t. For $p = o(n^{1/2})$, CKG19 requires the number of children for each node in G_t to grow slower than $n^{3/4}$, while Theorem 4.2 does not have any constraints on $|E_t|$ (it can be as large as $p(p-1)/2$). On the other hand, LLL19 sets the same upper bound for the cardinality of the parent set and children set of each node which is allowed to increase with the sample size. When $p < n$, LLL19 does not have any constraints on the total number of edges.

In the strong selection consistency results, CKG19 adds another restriction on the total number of edges. Assuming $p = o(n^{1/3})$ in CKG19, this implies that the total number of edges has to increase slower than $n^{7/12}$, while Theorem 4.3 only requires it to be slower than $n^{2/3}$ where G_t is allowed to be G_c. Overall, for $p < n$, CKG19 enforces stronger sparsity conditions on G_t at least in the well-specified case.

4.4.3. *Assumptions on the prior edge inclusion probability*

In CKG19, the prior edge inclusion probability always penalizes dense graphs through the maximum number of children for each node, while in this paper the prior edge inclusion probability is chosen based on ρ_L, $|E_t|$ and p (only in Theorem 4.3) under the well-specified case. There is no direct comparison between these two assumptions due to very different sparsity assumptions and perspectives under which the consistency is studied in both papers. As discussed before, the ESC prior in LLL19 does not have an interpretation of the prior edge inclusion probability.

4.4.4. *Assumptions on the signal size*

In CKG19, the signal size is defined as the smallest (in absolute value) non-zero entry in the lower triangular matrix of the modified Cholesky decomposition of Ω_0, denoted by s_n. The beta-min condition in LLL19 restricts the square of the signal size defined in CKG19. In this paper, we consider ρ_L, the smallest non-zero partial correlation (in absolute value), as the analogue of the signal size. Although the definitions are different, s_n and ρ_L both relate to some quantities about the number of edges. In CKG19 the signal size goes to zero slower when the maximum number of children for each node increases; in this paper the signal size converges to zero slower when $|E_t|$ increases; in LLL19 the signal size cannot converge to zero faster than $1/\sqrt{n}$ which is similar to our general assumption on ρ_L. Both signal sizes serve as a measure of identifiability in the selection consistency.

In the following, we look at an example to show that neither of the assumptions in CKG19 and this paper implies the other. Let Ω denote the precision matrix of a complete graph with the maximum number of children d in the modified Cholesky decomposition $\Omega = L(aI_p)^{-1}L^T$, where $a > 0$. The following calculation does not depend on the choice of a as long as it is greater than zero. For simplicity,

all non-zero lower triangular entries are assumed to be the same, say s. Thus, s is also the signal size defined in CKG19. The smallest partial correlation in terms of d and s is given by

$$\rho_L = \frac{1}{\sqrt{1/s^2 + d}}.$$

First, when $s \to \infty$, $\rho_L \to 1/\sqrt{d}$. In CKG19, d cannot grow faster than $n^{\frac{1}{4}}$, hence the assumption about ρ_L in this paper is always satisfied. When $s \to 0$, based on Assumption 4 in CKG19, the signal size must be converging to zero at a rate slower than $n^{-1/4}$ while d grows slower than $1/s^2$. In this case, ρ_L must be slower than $n^{-1/4}$ as well. Thus, the assumptions of the strong selection consistency (Theorem 4.3) in this paper are not met since it requires ρ_L to be slower than $n^{-1/5}$. But Theorem 4.2 still holds.

4.4.5. *Assumptions on HIW g-priors and DAG-Wishart priors*

For the HIW g-prior used in this paper, the assumption that its degrees of freedom $b > 2$ is analogous to Assumption 5(i) in CKG19. Also, the fact that $(1/n)\mathrm{Y}^T\mathrm{Y}$ has a constant order of eigenvalues is essentially the same as Assumption 5(ii) in CKG19. Although the ESC prior in LLL19 is closely related to the DAG-Wishart prior under certain conditions, there is no direct connection with the HIW g-prior.

Based on the comparisons presented in this section, the assumptions in CKG19 (and LLL19) do not imply the ones in this paper and vice versa even when $p < n$. The assumptions in all three papers are only the sufficient conditions and violation of any assumption does not necessarily compromise the consistency.

5. Theoretical results under model misspecification

In this section, we investigate the effect of model misspecification when the underlying true graph G_t is non-decomposable. We develop theoretical results for non-decomposable graphs based on the theorems and techniques in the well-specified case. The general assumptions remain the same as in Assumptions 4.1–4.5 but different versions of restrictions on $(\alpha, \lambda, \sigma, \gamma)$ are introduced to handle the misspecification. We begin with some definitions on triangulations and minimal triangulations of a graph.

Definition 5.1 (Triangulation). Let $G = (V, E)$ be a non-decomposable graph. A graph $G^\Delta = (V, E \cup F)$ where $E \cap F = \varnothing$ is called a triangulation of G if G^Δ is decomposable. The edges in F are called *fill-in* edges.

Definition 5.2 (Minimal triangulation [22,35]). Let $G^\Delta = (V, E \cup F)$ be a triangulation of the non-decomposable graph $G = (V, E)$. G^Δ is a minimal triangulation of G if $(V, E \cup F')$ is non-decomposable for every $F' \subsetneq F$. In other words, a triangulation is minimal if and only if the removal of any single fill-in edge from it results in a non-decomposable graph.

Definition 5.2 captures the important aspect of minimal triangulations, that is, no fill-in edge in a minimal triangulation is redundant. To better understand this concept, see the following example in Figure 5. Graph G_{o6} is not decomposable since it contains a cycle with length 6, i.e., $1-2-3-4-5-6-1$. Graph $G_{m_1}, G_{m_2}, G_{m_3}, G_{\text{tri}}$ are all triangulations of G_{o6} since they are all decomposable and the edge set of G_{o6} is a subset of theirs. Furthermore, $G_{m_1}, G_{m_2}, G_{m_3}$ are minimal triangulations of

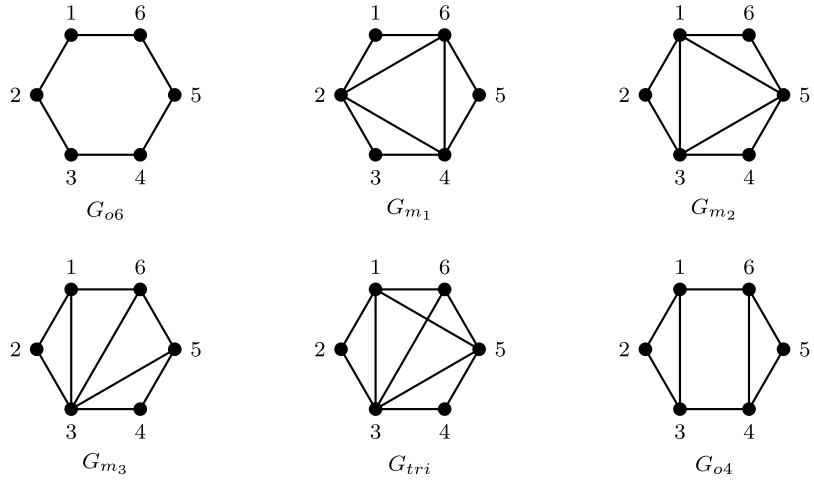


Figure 5. G_{o6} is not decomposable; G_{m_1},G_{m_2},G_{m_3} are all minimal triangulations of G_{o6}; G_{tri} is a triangulation of G_{o6} but not minimal; G_{o4} is not a triangulation of G_{o6}.

G_{o6} since removing any fill-in edge in them results in non-decomposable graphs. G_{tri} is not a minimal triangulation of G_{o6} since the removal of edge $3-6$ or $1-5$ results in decomposable graphs G_{m_2} or G_{m_3}, respectively. It has redundant fill-in edges, thus not minimal. G_{o4} is not a triangulation of G_{o6} since it is not decomposable due to the cycle with length 4, i.e., $1-3-4-6-1$. This example shows that minimal triangulations are not unique. For a summary of minimal triangulations, see [22] for more details. Next, we state theorems of graph selection consistency when the true graph G_t is non-decomposable.

Theorem 5.1 (Convergence of minimal triangulations for finite graphs). *Assume the true graph G_t is non-decomposable. When the graph dimension p is a fixed constant (ρ_L and ρ_U are fixed constants), we have the following results.*

1. *Let G_m be any minimal triangulation of G_t and G_a be any decomposable graph that is not a minimal triangulation of G_t. If $\rho_U \neq 1$, then $\text{BF}(G_a; G_m) \xrightarrow{\mathbb{P}} 0$, as $n \to \infty$.*
2. *If $\rho_U \neq 1$, we have $\sum_{G_m \in \mathcal{M}_t} \pi(G_m \mid \text{Y}) \xrightarrow{\mathbb{P}} 1$, as $n \to \infty$, where $\mathcal{M}_t$ is the minimal triangulation space of G_t.*

Theorem 5.1 states that Bayes factors favor minimal triangulations over any other decomposable graphs. This leads to the pairwise Bayes factor consistency for minimal triangulations. The second part of the theorem ensures that the posterior of a candidate graph under model misspecification eventually concentrates within the minimal triangulation space, that is, the sum of all posterior probabilities of minimal triangulations converges to 1. The minimal triangulations serve as a proxy of G_t. They contain all edges in G_t with the minimal number of fill-in edges. Thus, the minimal number of false positive edges and no false negative edges are both guaranteed during model fitting.

Theorem 5.2 (Equivalence of minimal triangulations for finite graphs). *Assume the true graph G_t is non-decomposable. Let G_{m_1} and G_{m_2} be any two different minimal triangulations of G_t (with the same number of fill-in edges). When the graph dimension p is a fixed constant, the Bayes factor*

between them are stochastically bounded, that is, for any $0 < \epsilon < 1$, *there exist two positive finite constants* $A_1(\epsilon) < 1$ *and* $A_2(\epsilon) > 1$, *such that*

$$\mathbb{P}\big\{A_1 < \mathrm{BF}(G_{m_1}; G_{m_2}) < A_2\big\} > 1 - \epsilon \quad \textit{for } n > p + \max\big\{3, b, 6\log\big(10p^2/\epsilon\big)\big\}.$$

Although minimal triangulation graphs are not unique, Theorem 5.2 says that the Bayes factor between any two minimal triangulations is stochastically bounded. Asymptotically the posterior probability may not converge to 1 for a single minimal triangulation of G_t. In fact, the posterior assigns non-vanishing probabilities to different minimal triangulations. The highest posterior probability model depends on the given data set. Refer to the second simulation in Section 6 for an example. In [18], an asymptotic version of Theorem 5.3 is presented. In contrast, our result is non-asymptotic and provides an exact form of the constants, see Section B.2 in the Supplementary Materials [32] for details.

The Bayes factors considered in Theorems 5.1 and 5.2 are in fact between two decomposable graphs, except one of them is a minimal triangulation (which is decomposable) of the true graph. Therefore, the proofs follow similar steps as in the proof of Theorem 4.1.

Theorem 5.3 (Convergence of minimal triangulations for high-dimensional graphs). *Assume the true graph* G_t *is not decomposable. When the graph dimension* p *grows with* n, *we have the following results.*

1. *Let* G_m *be any minimal triangulation of* G_t *and* G_a *be any decomposable graph that is not a minimal triangulation of* G_t. *Assume*

$$0 < \alpha < \frac{1}{2}, \qquad 0 \le \lambda < \min\left\{\frac{1}{4}, \frac{1-2\alpha}{2}\right\}, \qquad 0 < \sigma < 1 - 2\alpha - 2\lambda.$$

   Choose γ *in the interval* $(2\alpha, 1 - \sigma - 2\lambda)$. *Then under Assumptions 4.1–4.5, we have* $\mathrm{PR}(G_a; G_m) \xrightarrow{\mathbb{P}} 0$, *as* $n \to \infty$.
2. *If*

$$0 < \alpha < \frac{1}{3}, \qquad 0 \le \lambda < \min\left\{\frac{1-\alpha}{4}, \frac{1-3\alpha}{2}\right\}, \qquad 0 < \sigma < 1 - 3\alpha - 2\lambda.$$

   And we choose γ *in the interval* $(3\alpha, 1 - \sigma - 2\lambda)$, *then under Assumptions 4.1–4.5, we have* $\sum_{G_m \in \mathcal{M}_t} \pi(G_m \mid \mathrm{Y}) \xrightarrow{\mathbb{P}} 1$, *as* $n \to \infty$, *where* $\mathcal{M}_t$ *is the minimal triangulation space of* G_t.

Theorem 5.3 is the high-dimensional version of convergence of minimal triangulations under model misspecification. The consistency results are the same as in Theorem 5.1 with some additional assumptions on the parameters $(\alpha, \lambda, \sigma, \gamma)$. Comparing with Theorems 4.2 and 4.3, assumptions on α and ρ_L remain the same. The upper bound of γ is also the same as in the well-specified case. Under model misspecification, a sparsity assumption on the number of edges through σ and a larger lower bound on γ are needed to ensure the consistency of minimal triangulations.

In the well-specified case, we only consider two cases in the proofs, i.e., $G_t \subsetneq G_a$ and $G_t \not\subset G_a$. In the case of model misspecification, there are three scenarios which are $G_m \subsetneq G_a$, $G_m \not\subset G_a$ with $|E_a^1| < |E_t|$, and $G_m \not\subset G_a$ with $G_t \subsetneq G_m, G_a$. The first two scenarios mirror the two cases in the well-specified case. The third scenario is added because of properties of minimal triangulations. Due to the proof techniques adopted here, in order to show the consistency for the third scenario, additional constraints on γ have to be placed. It results in lowering the edge inclusion probability which in return imposes a sparsity condition on the total number of edges. This occurs in both parts of Theorem 5.3. We

state that the assumptions in Theorem 5.3 are only sufficient conditions needed to achieve consistency. The current proof techniques are not suitable to remove the extra assumption on the number of edges.

Theorem 5.4 (Equivalence of minimal triangulations for high-dimensional graphs). *Assume the true graph G_t is not decomposable and the graph dimension p grows with n. Let G_{m_1} and G_{m_2} be any two different minimal triangulations of G_t. If the number of fill-in edges is finite, then the Bayes factor between them are stochastically bounded.*

This is an extension of Theorem 5.2 when p grows with n. Based on Theorem 5.4, the equivalence among minimal triangulations is true when the number of fill-in edges is finite. Adding infinitely many fill-in edges prompts the minimal triangulations to drift further away from each other; thus, the equivalence may not be valid in such case. It is worth noting that any decomposable subgraph of the true graph is not a good posterior estimate. This is simply due to the fact that false-edge deletions result in an exponential decay of Bayes factors in favor of decomposable subgraphs. Analogous to Corollary 4.3, we show in Corollary 5.1 that when the true graph G_t is not decomposable, the posterior mode is one of the minimal triangulations and is consistent.

Corollary 5.1 (Consistency of the posterior mode when G_t is non-decomposable). *Under the assumptions of the second part of Theorem 5.3, the posterior mode $\hat{G}$ is in the minimal triangulation space $\mathcal{M}_t$ of the true graph G_t with probability converging to one, that is,*

$$\mathbb{P}(\hat{G} \in \mathcal{M}_t) \to 1 \quad \text{as } n \to \infty.$$

6. Simulations

We conduct two sets of simulations for demonstrating the convergence of Bayes factors in the well-specified case (Theorem 4.1) and in the misspecified case (Theorem 5.1) for fixed p.

6.1. Simulation 1: Demonstration of pairwise Bayes factor convergence rate

In this section, we conduct a simulation study in $\mathcal{D}_3$ to demonstrate the convergence rate of pairwise Bayes factors. Since there is no non-decomposable graph with 3 nodes, $\mathcal{D}_3$ is the same as $\mathcal{G}_3$. All 8 graphs in $\mathcal{D}_3$ are enumerated in Figure 6.

The underlying covariance matrix Σ_3 and its precision matrix Ω_3 are shown below along with the correlation matrix R_3 and the partial correlation matrix $\overline{R}_3$. Samples are drawn independently from $N_3(\mathbf{0}, \Sigma_3)$. The range of the sample size simulated is from 100 to 10,000 with an increment of 100. The Bayes factor for each sample size is averaged over 1000 simulation replicates. The degrees of freedom b in the HIW g-prior is chosen to be 3. The first six pairwise Bayes factors in logarithmic scale are shown in Figure 7(a) and the logarithm of $\mathrm{BF}(G_c; G_t)$ is shown separately in Figure 7(b) due to its slower convergence rate. To better understand the simulation results, asymptotic leading terms of pairwise Bayes factors in logarithmic scale and the empirically estimated slopes for n or $\log n$ are listed in the second and third columns of Table 1. To calculate the leading terms, sample partial correlations or sample correlations are replaced with their population counterparts that do not depend on n. The slopes of logarithms of the first six Bayes factors in Figure 7(a) are calculated in Table 1 by performing linear regressions on n. The last slope in Table 1 is calculated with a linear regression on $\log n$; refer to

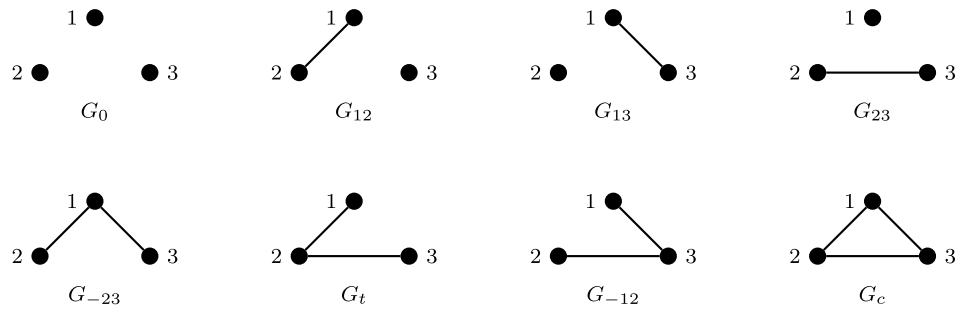


Figure 6. Enumerating all 3-node decomposable graphs in $\mathcal{D}_3$ with G_t as the true graph, G_0 as the null graph and G_c as the complete graph.

Figure 7(b). Table 1 shows that the theoretical asymptotic leading terms match well with their empirical values.

$$\Sigma_3 = \begin{bmatrix} 0.7119 & -0.4237 & 0.1695 \\ -0.4237 & 0.8475 & -0.3390 \\ 0.1695 & -0.3390 & 0.6356 \end{bmatrix}, \qquad \Omega_3 = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0.8 \\ 0 & 0.8 & 2 \end{bmatrix},$$

$$R_3 = \begin{bmatrix} 1.0000 & -0.5456 & 0.2520 \\ -0.5456 & 1.0000 & -0.4619 \\ 0.2520 & -0.4619 & 1.0000 \end{bmatrix}, \qquad \overline{R}_3 = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.4 \\ 0 & 0.4 & 1 \end{bmatrix}.$$

From the simulation results, we can see that missing at least one true edge of G_t in G_a results in the Bayes factor converging to zero exponentially. This is perfectly illustrated by all six Bayes factors in Figure 7(a). On the other hand, adding false edges in G_a results in the Bayes factor going to zero at a polynomial rate which is much slower than the case of missing a true edge, see Figure 7(b). These discoveries are consistent with Table 1 and our proofs.

Next, we compare the rates in the convergence of the first six Bayes factors. The convergence rate associated with missing two true edges of G_t is faster than missing only one true edge, that is, $\mathrm{BF}(G_0; G_t)$ vs. $\mathrm{BF}(G_{23}; G_t)$ and $\mathrm{BF}(G_0; G_t)$ vs. $\mathrm{BF}(G_{12}; G_t)$. The convergence rate is faster when the missing true edge of G_t corresponds to a larger partial correlation (or correlation) in absolute value, that is, $\mathrm{BF}(G_{-12}; G_t)$ vs. $\mathrm{BF}(G_{-23}; G_t)$ and $\mathrm{BF}(G_{23}; G_t)$ vs. $\mathrm{BF}(G_{12}; G_t)$. One interesting fact is although G_0 and G_{13} are both missing two true edges of G_t, with G_{13} having an additional false edge of G_t

Table 1. Asymptotic leading terms and simulation slopes of Bayes factors in logarithmic scale

Bayes factor	Asymptotic leading term	Simulation slope
$\mathrm{BF}(G_0; G_t)$	$\{\log(1-\rho_{12}^2)+\log(1-\rho_{23}^2)\}\cdot n/2 = -0.2967\cdot n$	-0.2963
$\mathrm{BF}(G_{13}; G_t)$	$\{\log(1-\rho_{12}^2)+\log(1-\rho_{23\|1}^2)\}\cdot n/2 = -0.2639\cdot n$	-0.2637
$\mathrm{BF}(G_{23}; G_t)$	$\log(1-\rho_{12}^2)\cdot n/2 = -0.1767\cdot n$	-0.1765
$\mathrm{BF}(G_{-12}; G_t)$	$\log(1-\rho_{12\|3}^2)\cdot n/2 = -0.1438\cdot n$	-0.1439
$\mathrm{BF}(G_{12}; G_t)$	$\log(1-\rho_{23}^2)\cdot n/2 = -0.1120\cdot n$	-0.1198
$\mathrm{BF}(G_{-23}; G_t)$	$\log(1-\rho_{23\|1}^2)\cdot n/2 = -0.0872\cdot n$	-0.0873
$\mathrm{BF}(G_c; G_t)$	$-0.5\cdot \log n$	-0.5106

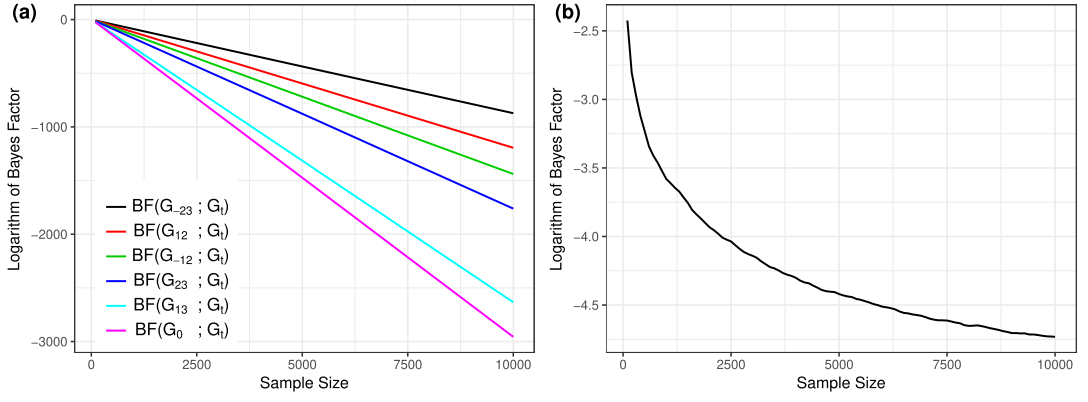


Figure 7. Simulation results of pairwise Bayes factors of $\mathcal{D}_3$ in logarithmic scale. (a) Six Bayes factors where $G_t \not\subset G_a$ (at least missing one true edge of G_t). (b) Bayes factor $\mathrm{BF}(G_c; G_t)$ when $G_t \subsetneq G_a = G_c$ (only false-edge additions).

compared to G_0, the convergence rate of the Bayes factor for G_{13} is slower than that for G_0. The reason is clear from Table 1. As the absolute value of the correlation between nodes 2 and 3 ($|\rho_{23}| = 0.4619$) is larger than the absolute value of the partial correlation between them given node 1 ($|\rho_{23|1}| = 0.4$); thus, the leading term of $\mathrm{BF}(G_0; G_t)$ is smaller than that of $\mathrm{BF}(G_{13}; G_t)$. The effect due to the false edge $1 - 3$ (polynomial rate) is overwhelmed by the leading term (exponential rate). It is evident that HIW priors place a larger penalty on false negative edges compared to false positive edges. This confirms Theorem 4.1. Similar conclusions can be made by comparing $\mathrm{BF}(G_{23}; G_t)$ and $\mathrm{BF}(G_{-12}; G_t)$, also by comparing $\mathrm{BF}(G_{12}; G_t)$ and $\mathrm{BF}(G_{-23}; G_t)$.

6.2. Simulation 2: Examination of model misspecification

In this section, we illustrate the stochastic equivalence between minimal triangulations when the true graph is non-decomposable. The smallest non-decomposable graph is a cycle of length 4 without a chord. So we focus our simulation in $\mathcal{D}_4$. Since the number of decomposable graphs increases exponentially with the dimension of graphs, we only select 5 alternative graphs in $\mathcal{D}_4$ besides the minimal triangulation, see Figure 8. The true covariance matrix Σ_4 and its precision matrix Ω_4 are listed below along with the correlation matrix R_4 and the partial correlation matrix $\overline{R}_4$. All simulation settings are the same as the simulation in $\mathcal{D}_3$.

$$\Sigma_4 = \begin{bmatrix} 1.8364 & -1.0909 & 0.8909 & -1.3636 \\ -1.0909 & 1.0606 & -0.7273 & 0.9091 \\ 0.8909 & -0.7273 & 0.9273 & -0.9091 \\ -1.3636 & 0.9091 & -0.9091 & 1.6364 \end{bmatrix}, \qquad \Omega_4 = \begin{bmatrix} 2 & 1.2 & 0 & 1 \\ 1.2 & 3 & 1.2 & 0 \\ 0 & 1.2 & 3 & 1 \\ 1 & 0 & 1 & 2 \end{bmatrix},$$

$$R_4 = \begin{bmatrix} 1.0000 & -0.7817 & 0.6827 & -0.7866 \\ -0.7817 & 1.0000 & -0.7334 & 0.6901 \\ 0.6827 & -0.7334 & 1.0000 & -0.7380 \\ -0.7866 & 0.6901 & -0.7380 & 1.0000 \end{bmatrix}, \qquad \overline{R}_4 = \begin{bmatrix} 1 & 0.49 & 0 & 0.50 \\ 0.49 & 1 & 0.40 & 0 \\ 0 & 0.40 & 1 & 0.41 \\ 0.50 & 0 & 0.41 & 1 \end{bmatrix}.$$

When the true graph G_t is non-decomposable, the two minimal triangulations of G_t act like two proxy graphs of G_t. We plot the first four pairwise Bayes factors where $G_{m_i} \not\subset G_a$ $(i = 1, 2)$ for G_{m_1}

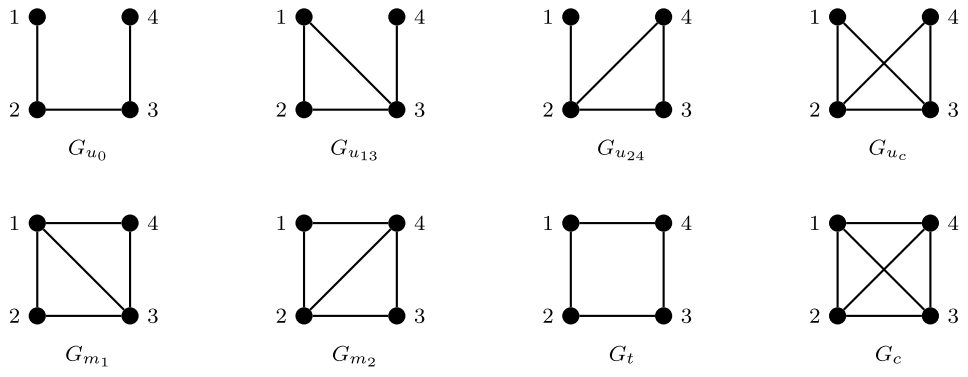


Figure 8. Some selected graphs in $\mathcal{G}_4$, including G_t as the true graph which is non-decomposable. G_{m_1} and G_{m_2} are two minimal triangulations of G_t.

and G_{m_2} in logarithmic scale together in Figure 9(a) and (b), respectively. The logarithm of the Bayes factor between the two minimal triangulations is in Figure 9(c). Finally, we plot the Bayes factors between G_c (one triangulation of G_t, but not minimal) and both minimal triangulations in logarithmic scale in Figure 9(d).

Figure 9(a) and (b) are Bayes factors with true-edge deletions from G_{m_1} and G_{m_2}, respectively. They behave the same as in the well-specified case, meaning that true-edge deletions cause exponential decay of Bayes factors under model misspecification as well. Figure 9(d) shows that Bayes factors with false-edge additions decay at a polynomial rate under model misspecification. This is also the same as in the well-specified case. Based on the simulation result in Figure 9(c), we can see the Bayes factor between the two minimal triangulations G_{m_1} and G_{m_2} appears to be stochastically bounded.

7. Discussion

In this paper, we provide a complete theoretical foundation for high-dimensional decomposable graph selection under model misspecification. When the graph dimension is finite, Fitch, Jones and Massam [18] present pairwise Bayes factor consistency results and stochastic equivalence among minimal triangulations. We provide more general results of both pairwise consistency and strong selection consistency in high-dimensional scenarios. To the best of our knowledge, these are the first complete results on this topic so far. Our current choice of edge inclusion probability requires the knowledge of total number of edges in the true graph. We anticipate this can be relaxed by placing an appropriate prior on γ.

In our results, the graph size cannot be equal to or exceed $n^{1/2}$ and $n^{1/3}$ for posterior ratio consistency and strong selection consistency, respectively. The limitation of the growth rate of the graph dimension is caused by the convergence rate of sample partial correlations and sample correlations. With the current techniques, without further investigating the relationship among sample partial correlations, these results cannot be improved. Observe that in the i.i.d. case without any sparsity assumptions, such an assumption on the growth of the graph dimension is common [23,39]. We conjecture that it may not be possible to relax the growth rate of p for achieving strong selection consistency using the current formulation of HIW priors. This is simply because HIW priors do not penalize false edges significantly enough so that in high dimension a prior on the graph space is needed to achieve both pairwise and strong selection consistency. Also any other sparsity restriction on the elements of the

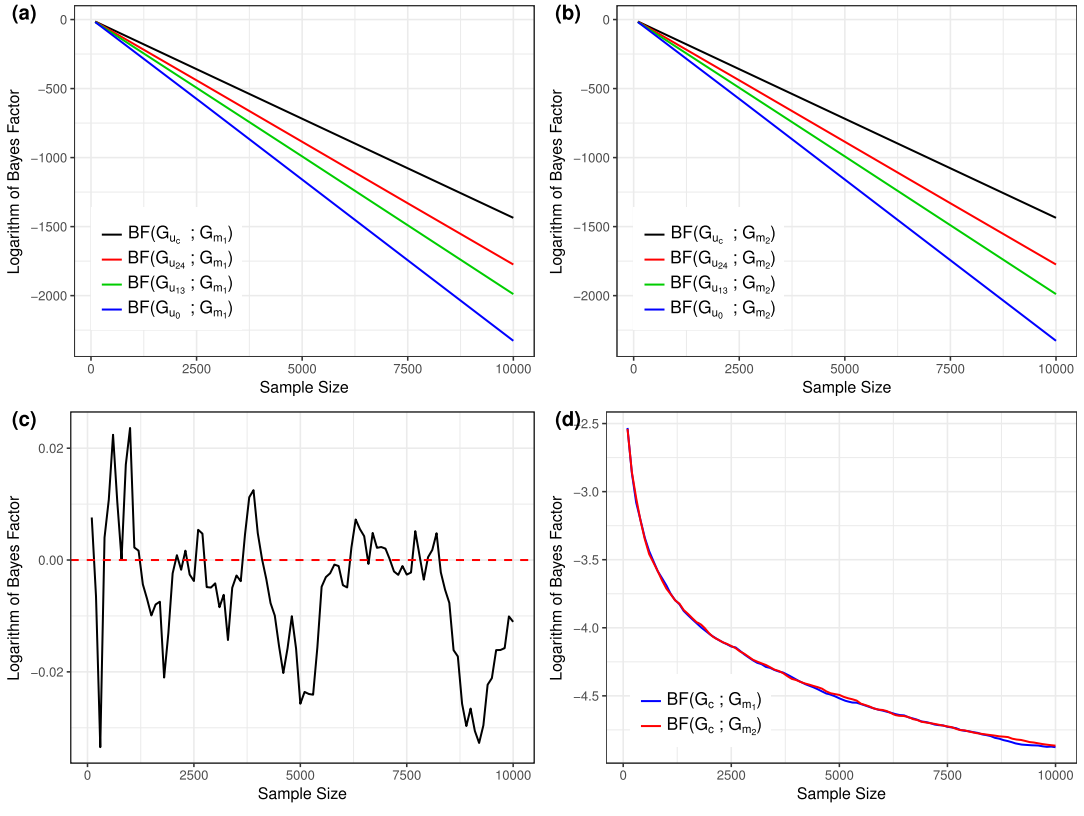


Figure 9. Simulation results of pairwise Bayes factors of $\mathcal{D}_4$ in logarithmic scale. (a) Bayes factor $\mathrm{BF}(G_a; G_{m_1})$ when $G_{m_1} \not\subset G_a$ (missing true edges). (b) Bayes factor $\mathrm{BF}(G_a; G_{m_2})$ when $G_{m_2} \not\subset G_a$ (missing true edges). (c) Bayes factor between the two minimal triangulations of G_t, $\mathrm{BF}(G_{m_2}; G_{m_1})$. (d) Bayes factor $\mathrm{BF}(G_a; G_{m_i})$ when $G_{m_i} \subsetneq G_a = G_c$, $i = 1, 2$ (only false-edge additions).

precision matrix is not supported by HIW priors due to its inability to enforce sufficient shrinkage conditional on the graph. This limits extending the technical results to ultra-high-dimensional cases by enforcing additional sparsity assumptions on the elements of the precision matrix. This apparent "flaw" lies in the construction of HIW priors and cannot be improved by adding any reasonable penalty on the graph space.

For technical simplicity, our results are based on the HIW g-prior only. We conjecture that the consistency results continue to hold for general HIW priors. Moreover, extensions to non-decomposable graphical models can be done by using G-Wishart priors, but major bottlenecks are expected stemming from the lack of a closed form for the normalizing constant. A recent work [40] on the development of exact formulas for the normalizing constant can be used to show consistency results of Bayes factors for general graphs. A future area of research will also involve studying repercussion of model misspecification on other related approaches of Bayesian graph learning [18,24].

Overview of results in the Appendix. Note that the proofs in the Appendix require a set of auxiliary results. The readers are deferred to the Supplementary Materials [32] for these results. The Appendix begins with a set of graph theoretic results required to prove Theorem 4.1. Then we provide the proof

Table 2. Summary of notations

Symbol	Definition
$\mathbb{P}$	The probability corresponding to the true data generating distribution
$\mathcal{G}_k, \mathcal{D}_k$	k-dimensional graph space, k-dimensional decomposable graph space
$\mathcal{M}_t$	The minimal triangulation space of G_t when G_t is non-decomposable
$a \asymp b$	$C_1 b \le a \le C_2 b$ for positive constants C_1, C_2
$a \precsim b, a \succsim b$	$a \le C_3 b$ for a positive constant C_3, $a \ge C_4 b$ for a positive constant C_4
$A \subset B$, $A \not\subset B$	A is a subset of B, A is not a subset of B
$A \subsetneq B$	$A \subset B$ and $A \ne B$
$\lvert\cdot\rvert$	Absolute value, cardinality of a set or determinant of a matrix by context
$\pi(\cdot), \pi(\cdot \mid \mathrm{Y})$	The prior distribution and posterior distribution of graphs
$\mathrm{Y}, Y_i^T, \mathrm{y}_i$	The $n \times p$ data matrix, the ith row of Y, the ith column of Y
$\rho_{ij}, \rho_{ij\lvert S}$	The correlation and partial correlation between X_i and X_j given X_S
$\hat{\rho}_{ij}, \hat{\rho}_{ij\lvert S}$	The sample correlation and partial correlation between X_i and X_j given X_S
ρ_L, ρ_U	The lower and upper bound for all $\lvert\rho_{ij\lvert V\setminus\{i,j\}}\rvert$, where $(i, j) \in E_t$
$C_i, \mathcal{C}, S_i, \mathcal{S}$	A clique, the set of cliques, a separator, the set of separators
G_t, G_a, G_c	The true graph, any decomposable graph, the complete graph
G_m, G_0	A minimal triangulation of G_t, the empty graph
$\hat{G}$	The posterior mode in the decomposable graph space
E_t, E_a, E_c, E_a^1	Edge sets of G_t, G_a, G_c and $E_a^1 = E_a \cap E_t$
p, V	The graph dimension, the vertex set, where $V = \{1, 2, \ldots, p\}$
$x, \overline{x}, \widetilde{x}$	Nodes in the graph
i, j	Determined by context, nodes in the graph or indices of nodes
$S, \overline{S}, \widetilde{S}$	Separators in the graph
d_S, q	The cardinality of separator S, the prior edge inclusion probability
$\Delta'_\epsilon, \Delta'_\epsilon(n), \Delta''_\epsilon(n)$	Probability regions of sample partial correlations
Π_{xy}	The set of all sets that separates nodes x and y, where $(x, y) \notin E_t$
$G_{\pm(x,y)\in E_t}$	A graph with/without true edge (x, y)
$G_{\pm(x,y)\notin E_t}$	A graph with/without false edge (x, y)
$\overline{G}_i^{c\to a}, \widetilde{G}_i^{t\to c}$	The ith graph in the sequence from G_c to G_a and G_t to G_c

of Theorem 4.1 followed by the proof of Theorem 4.2. We also provide the proofs of theorems related to minimal triangulations, that is, Theorems 5.1 and 5.2.

Appendix A: Pairwise Bayes factor consistency and posterior ratio consistency – any graph G_a versus the true graph G_t

Lemma A.1 (Decomposable graph chain rule [27]). *Let $G = (V, E)$ be a decomposable graph and let $G' = (V, E')$ be a subgraph of G that also is decomposable with $\lvert E \setminus E'\rvert = k$. Then there is an increasing sequence $G' = G_0 \subset G_1 \cdots \subset G_{k-1} \subset G_k = G$ of decomposable graphs that differ by exactly one edge.*

Assume $G_t \not\subset G_a$, then $\lvert E_t\rvert > \lvert E_a^1\rvert$. By Lemma A.1, there exists a decreasing sequence of decomposable graphs from G_c to G_a that differ by exactly one edge, say $\{\overline{G}_i^{c\to a}\}_{i=0}^{\lvert E_c\rvert-\lvert E_a\rvert}$, where $G_c = \overline{G}_0^{c\to a} \supsetneq \overline{G}_1^{c\to a} \supsetneq \cdots \supsetneq \overline{G}_{\lvert E_c\rvert-\lvert E_a\rvert-1}^{c\to a} \supsetneq \overline{G}_{\lvert E_c\rvert-\lvert E_a\rvert}^{c\to a} = G_a$. There are $\lvert E_c\rvert - \lvert E_a\rvert$ steps for moving from G_c to G_a. Let $\{\rho_{\overline{x}_i\overline{y}_i\lvert\overline{S}_i}\}_{i=1}^{\lvert E_c\rvert-\lvert E_a\rvert}$ be the corresponding population partial correlation (or correlation,

when $\overline{S}_i = \varnothing$) sequence and $\{\mathrm{BF}(\overline{G}_i^{c\to a}; \overline{G}_{i-1}^{c\to a})\}_{i=1}^{|E_c|-|E_a|}$ be the corresponding Bayes factor sequence for each step. By that, we mean in the ith step, edge $(\overline{x}_i, \overline{y}_i)$ is removed; $\rho_{\overline{x}_i\overline{y}_i|\overline{S}_i}$ and $\mathrm{BF}(\overline{G}_i^{c\to a}; \overline{G}_{i-1}^{c\to a})$ are the population partial correlation and the Bayes factor accordingly, $i = 1, 2, \ldots, |E_c| - |E_a|$. $\overline{S}_i$ is the specific separator corresponding to the ith step. Among them $|E_t| - |E_a^1|$ steps are removals of true edges that are true-edge deletion cases; $|E_c| - |E_a| - |E_t| + |E_a^1|$ steps are removals of false edges that can be seen as false-edge deletion cases.

Lemma A.2 (Origin of the exponential rate in true-edge deletion cases). *Assume $G_t \not\subset G_a$. There exists at least one sequence of partial correlations $\{\rho_{\overline{x}_i\overline{y}_i|\overline{S}_i}\}_{i=1}^{|E_c|-|E_a|}$ for moving from G_c to G_a such that among all population partial correlations in the sequence that are corresponding to the removal of true edges, at least one is non-zero and it is not a population correlation ($\overline{S}_i \neq \varnothing$).*

Proof. There are many sequences of $\{(\overline{x}_i, \overline{y}_i)\}_{i=1}^{|E_c|-|E_a|}$ (in different orders) that can achieve moving from G_c to G_a and still maintain decomposability along the way. Let $(\overline{x}_*, \overline{y}_*) \in E_t \backslash E_a^1$. Thus, $(\overline{x}_*, \overline{y}_*) \in \{(\overline{x}_i, \overline{y}_i)\}_{i=1}^{|E_c|-|E_a|}$. Choose $(\overline{x}_1, \overline{y}_1) = (\overline{x}_*, \overline{y}_*)$. This means the first step is the removal of a true edge in $E_t \backslash E_a^1$ from G_c. (The rest of steps can be arbitrary as long as they maintain decomposability.) Let $\overline{S}_*$ be the corresponding separator for the first step. Thus, we know $\overline{S}_* = V\backslash\{\overline{x}_*, \overline{y}_*\} \neq \varnothing$, since $(\overline{x}_*, \overline{y}_*)$ is removed from G_c. In fact, the removal of any edge from a complete graph still maintains decomposability, that is, $\overline{G}_1^{c\to a}$ is a decomposable graph. Since $(\overline{x}_*, \overline{y}_*) \in E_t$, by the pairwise Markov property, $\rho_{\overline{x}_*,\overline{y}_*|V\backslash\{\overline{x}_*,\overline{y}_*\}} \neq 0$. And $\rho_L \leq |\rho_{\overline{x}_*,\overline{y}_*|V\backslash\{\overline{x}_*,\overline{y}_*\}}| \leq \rho_U$. Therefore, the partial correlation sequence starting with $\rho_{\overline{x}_*,\overline{y}_*|V\backslash\{\overline{x}_*,\overline{y}_*\}}$ is the one which satisfies all conditions according to the lemma since $\rho_{\overline{x}_*,\overline{y}_*|V\backslash\{\overline{x}_*,\overline{y}_*\}} \neq 0$ and it is corresponding to the removal of a true edge $(\overline{x}_*, \overline{y}_*)$. Thus, we have proved the existence and found the sequence as well. □

Lemma A.3 (The inheritance of separators). *Let $G = (V, E)$ and $G' = (V, E')$ be two undirected graphs (not necessary to be decomposable). Assume $E \subseteq E'$. If $S \subsetneq V$ separates node $x \in V$ from node $y \in V$ in G', where $(x, y) \notin E'$, then S also separates them in G.*

Proof. First, if x and y are not connected in G (x and y are not adjacent as well), any node set can separate them by definition, so does S. The lemma holds trivially in this case. When x and y are connected in G, assume S does not separate x from y in G. By the definition of separators (with the fact that x and y are connected), there exists a path from x to y in G, say $x = v_0, v_1, \ldots, v_{l-1}, v_l = y$ and $v_i \notin S$, for all $i = 0, 1, \ldots, l$. Since $E \subseteq E'$, the path from x to y, $\{v_i\}_{i=1}^{l-1}$, is still a path from x to y in G'. By the definition of separators again, we know that S does not separate x from y in G'. But this contradicts with the assumption in the lemma. Therefore, S separates x from y in G. □

Assume $G_t \subsetneq G_a$, thus $|E_t| = |E_a^1|$. By Lemma A.1, there exists an increasing sequence of decomposable graphs from G_t to G_a that differ by exactly one edge, say $\{\widetilde{G}_i^{t\to a}\}_{i=0}^{|E_a|-|E_t|}$, where $G_t = \widetilde{G}_0^{t\to a} \subsetneq \widetilde{G}_1^{t\to a} \subsetneq \ldots \subsetneq \widetilde{G}_{|E_a|-|E_t|-1}^{t\to a} \subsetneq \widetilde{G}_{|E_a|-|E_t|}^{t\to a} = G_a$. There are $|E_a| - |E_t|$ steps for moving from G_t to G_a. All of them are additions of false edges that are false-edge addition cases. Let $\{\rho_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}\}_{i=1}^{|E_a|-|E_t|}$ be the corresponding population partial correlation (or correlation, when $\widetilde{S}_i = \varnothing$) sequence and $\{\mathrm{BF}(\widetilde{G}_i^{t\to a}; \widetilde{G}_{i-1}^{t\to a})\}_{i=1}^{|E_a|-|E_t|}$ be the corresponding Bayes factor sequence for each step. By that, we mean in the ith step, edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ is added; $\rho_{\widetilde{x}_i,\widetilde{y}_i|\widetilde{S}_i}$ and $\mathrm{BF}(\widetilde{G}_i^{t\to a}; \widetilde{G}_{i-1}^{t\to a})$ are the population partial correlation and the Bayes factor accordingly, $i = 1, 2, \ldots, |E_a| - |E_t|$. $\widetilde{S}_i$ is the specific separator corresponding to the ith step.

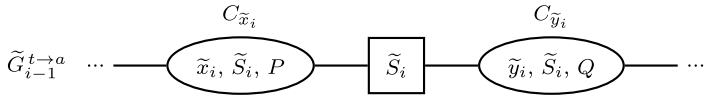


Figure 10. $\widetilde{G}_{i-1}^{t\to a}$ before adding edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ where $\widetilde{S}_i \neq \varnothing$.

Lemma A.4 (Origin of the polynomial rate in false-edge addition cases). *Assume $G_t \subsetneq G_a$. For any edge sequence $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{|E_a|-|E_t|}$ from G_t to G_a described above, all population partial correlations in $\{\rho_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}\}_{i=1}^{|E_a|-|E_t|}$ are zero (or correlation, when $\widetilde{S}_i = \varnothing$).*

Proof. Assume in the ith step, we add edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ to graph $\widetilde{G}_{i-1}^{t\to a}$ and $\widetilde{S}_i$ is the corresponding separator, where $1 \leq i \leq |E_a| - |E_t|$.

First, when $\widetilde{S}_i \neq \varnothing$, this case is showed in Figure 10. Since edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ is added in the ith step, by Lemma S.3.1, $C_{\widetilde{x}_i}$ and $C_{\widetilde{y}_i}$ are adjacent in some junction tree of $\widetilde{G}_{i-1}^{t\to a}$ where $C_{\widetilde{x}_i}$ and $C_{\widetilde{y}_i}$ are the cliques that contain $\widetilde{x}_i$ and $\widetilde{y}_i$, respectively. And $\widetilde{S}_i$ is the separator between them, i.e., $\widetilde{S}_i = C_{\widetilde{x}_i} \cap C_{\widetilde{y}_i}$. By the property of junction trees, we know $\widetilde{S}_i$ separates $\widetilde{x}_i$ from $\widetilde{y}_i$ in $\widetilde{G}_{i-1}^{t\to a}$. Since $\{\widetilde{G}_i^{t\to a}\}_{i=0}^{|E_a|-|E_t|}$ is an increasing sequence by edge, by Lemma A.3, we know $\widetilde{S}_i$ also separates $\widetilde{x}_i$ from $\widetilde{y}_i$ in $\widetilde{G}_0^{t\to a} = G_t$. By the global Markov property, $\rho_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i} = 0$.

Next, when $\widetilde{S}_i = \varnothing$, see Figure 11, we show $\rho_{\widetilde{x}_i\widetilde{y}_i} = 0$. By the property of junction trees, we know node $\widetilde{x}_i$ and $\widetilde{y}_i$ are disconnected. Furthermore, in the current graph $\widetilde{G}_{i-1}^{t\to a}$, nodes before clique $C_{\widetilde{x}_i}$ (including nodes in $C_{\widetilde{x}_i}$) and nodes after clique $C_{\widetilde{y}_i}$ (including nodes in $C_{\widetilde{y}_i}$) are disconnected. Since $G_t \subsetneq \widetilde{G}_{i-1}^{t\to a}$, then this is also true in G_t. Thus, nodes before clique $C_{\widetilde{x}_i}$ (including nodes in $C_{\widetilde{x}_i}$) and nodes after clique $C_{\widetilde{y}_i}$ (including nodes in $C_{\widetilde{y}_i}$) are disconnected in G_t. We can rearrange the precision matrix of G_t into a block matrix such that the block which $\widetilde{x}_i$ is in and the block which $\widetilde{y}_i$ is in are independent. Therefore, node $\widetilde{x}_i$ and $\widetilde{y}_i$ are marginally independent in G_t, $\rho_{\widetilde{x}_i\widetilde{y}_i} = 0$. Notice when $G_a = G_c$ this lemma still holds. □

For the rest of proofs, when $G_t \not\subset G_a$, moving from G_c to G_a is restricted to the sequence mentioned in Lemma A.2 (deleting a true edge at the beginning); when $G_t \subsetneq G_a$, moving from G_t to G_a (or G_c) can be sequences with any order of adding false edges (as long as decomposability is satisfied) according to Lemma A.4. Notice, there can be many other sequences of decomposable graphs for moving from G_c to G_a and G_t to G_a. We chose the sequences according to Lemma A.2 and A.4 mainly to simplify the proofs. The specific choices of sequences of decomposable graphs do not affect the final consistency results, but they do determine the sufficient conditions of each theorem. Other sequences of decomposable graphs can also be used to prove the consistency results, but they may result in different assumptions. Following the notations in Lemma A.2 and A.4, we have the decomposition of the Bayes

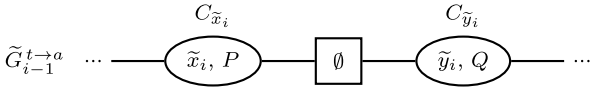


Figure 11. $\widetilde{G}_{i-1}^{t\to a}$ before adding edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ where $\widetilde{S}_i = \varnothing$.

factor in favor of G_a as follows. (i) When $G_t \not\subset G_a$,

$$\begin{aligned}
\mathrm{BF}(G_a; G_t) &= \frac{f(\mathrm{Y} \mid G_a)}{f(\mathrm{Y} \mid G_t)} = \frac{f(\mathrm{Y} \mid G_a)}{f(\mathrm{Y} \mid G_c)} \cdot \frac{f(\mathrm{Y} \mid G_c)}{f(\mathrm{Y} \mid G_t)} \\
&= \frac{p(\mathrm{Y} \mid G_a)}{p(\mathrm{Y} \mid \overline{G}^{c\to a}_{|E_c|-|E_a|-1})} \frac{p(\mathrm{Y} \mid \overline{G}^{c\to a}_{|E_c|-|E_a|-1})}{p(\mathrm{Y} \mid \overline{G}^{c\to a}_{|E_c|-|E_a|-2})} \cdots \frac{p(\mathrm{Y} \mid \overline{G}^{c\to a}_{2})}{p(\mathrm{Y} \mid \overline{G}^{c\to a}_{1})} \frac{p(\mathrm{Y} \mid \overline{G}^{c\to a}_{1})}{p(\mathrm{Y} \mid G_c)} \\
&\quad \times \frac{p(\mathrm{Y} \mid G_c)}{p(\mathrm{Y} \mid \widetilde{G}^{t\to c}_{|E_c|-|E_t|-1})} \frac{p(\mathrm{Y} \mid \widetilde{G}^{t\to c}_{|E_c|-|E_t|-1})}{p(\mathrm{Y} \mid \widetilde{G}^{t\to c}_{|E_c|-|E_t|-2})} \cdots \frac{p(\mathrm{Y} \mid \widetilde{G}^{t\to c}_{2})}{p(\mathrm{Y} \mid \widetilde{G}^{t\to c}_{1})} \frac{p(\mathrm{Y} \mid \widetilde{G}^{t\to c}_{1})}{p(\mathrm{Y} \mid G_t)} \\
&= \prod_{i=1}^{|E_c|-|E_a|} \mathrm{BF}\big(\overline{G}^{c\to a}_{i}; \overline{G}^{c\to a}_{i-1}\big) \cdot \prod_{i=1}^{|E_c|-|E_t|} \mathrm{BF}\big(\widetilde{G}^{t\to c}_{i}; \widetilde{G}^{t\to c}_{i-1}\big) \\
&= \mathrm{BF}_{c\to a} \cdot \mathrm{BF}_{t\to c}, \\
\mathrm{PR}(G_a; G_t) &= \frac{p(G_a \mid \mathrm{Y})}{p(G_t \mid \mathrm{Y})} = \frac{f(\mathrm{Y} \mid G_a)\pi(G_a)}{f(\mathrm{Y} \mid G_t)\pi(G_t)} = \mathrm{BF}(G_a; G_t)\frac{\pi(G_a)}{\pi(G_t)} \\
&= \mathrm{BF}_{c\to a} \cdot \mathrm{BF}_{t\to c} \cdot \left(\frac{q}{1-q}\right)^{|E_a|-|E_t|}.
\end{aligned}$$

$\mathrm{BF}_{c\to a}$ contains $|E_c| - |E_a|$ terms, in which $|E_t| - |E_a^1|$ terms are true-edge deletion cases and $|E_c| - |E_a| - |E_t| + |E_a^1|$ terms are false-edge deletion cases. $\mathrm{BF}_{t\to c}$ has $|E_c| - |E_t|$ terms that are all false-edge addition cases. (ii) When $G_t \subsetneq G_a$,

$$\begin{aligned}
\mathrm{BF}(G_a; G_t) &= \prod_{i=1}^{|E_a|-|E_t|} \mathrm{BF}\big(\widetilde{G}^{t\to a}_{i}; \widetilde{G}^{t\to a}_{i-1}\big) = \mathrm{BF}_{t\to a}, \\
\mathrm{PR}(G_a; G_t) &= \mathrm{BF}_{t\to a} \cdot \left(\frac{q}{1-q}\right)^{|E_a|-|E_t|}.
\end{aligned}$$

Here, we introduce the notation like $\mathrm{BF}_{c\to a}$. The use of such notations is to directly point out the moving direction in the sequence of decomposable graphs, since the traditional Bayes factor notation, in this case $\mathrm{BF}(G_a; G_c)$, may be confused for denoting the movement from G_a to G_c but in reality we look at it as the movement from G_c to G_a. These two directions are very different, one corresponds to adding edges and the other corresponds to deleting edges. The additional notation is introduced to make these two cases distinct.

A.1. Proof of Theorem 4.1

First, for any $\tau^* > 2$, let $\epsilon_{1,n} = \sqrt{\frac{\log(n-p)}{\tau^*(n-p)}}$. Then define

$$R'_{ij|S} = \big\{|\hat{\rho}_{ij|S} - \rho_{ij|S}| < \epsilon_{1,n}\big\}.$$

Given any decomposable graph $G_a \neq G_t$, when $G_t \not\subset G_a$, by Lemma A.2, we have the edge sequence $\{(\overline{x}_i, \overline{y}_i)\}_{i=1}^{|E_c|-|E_a|}$ for moving from G_c to G_a and let $(\overline{x}_1, \overline{y}_1) = (\overline{x}_*, \overline{y}_*)$ be the first in the sequence

where a true edge is deleted from G_c. Let $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{|E_c|-|E_t|}$ and $\{\widetilde{S}_i\}_{i=1}^{|E_c|-|E_t|}$ be the edge sequence and the corresponding separator sequence for moving from G_t to G_c according to Lemma A.4. Let

$$\Delta_{t\not\subset a,\epsilon_1} = \big(R'_{\overline{x}_*\overline{y}_*|V\setminus\{\overline{x}_*,\overline{y}_*\}}\big) \cap \left(\bigcap_{i=1}^{|E_c|-|E_t|} R'_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}\right).$$

Since $\rho_U \neq 1$, by the proof of Lemma S.1.1, we have

$$\mathbb{P}(\Delta_{t\not\subset a,\epsilon_1}) \geq \mathbb{P}\big(\Delta'_{\epsilon_1}\big) \geq 1 - \frac{42p^2}{(1-\rho_U)^2}(n-p)^{-\frac{1}{4\tau^*}}\left\{\frac{1}{\tau^*}\log(n-p)\right\}^{-\frac{1}{2}}.$$

When $G_t \subsetneq G_a$, let $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{|E_a|-|E_t|}$ and $\{\widetilde{S}_i\}_{i=1}^{|E_a|-|E_t|}$ be the edge sequence and the corresponding separator sequence for moving from G_t to G_a according to Lemma A.4. (Notice here we use the same edge and separator notations as in G_t to G_c for consistency reason and G_t to G_a can be seen as a part of G_t to G_c.) Let

$$\Delta_{t\subsetneq a,\epsilon_1} = \bigcap_{i=1}^{|E_a|-|E_t|} R'_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}.$$

Since $\rho_U \neq 1$, by the proof of Lemma S.1.1, we also have

$$\mathbb{P}(\Delta_{t\subsetneq a,\epsilon_1}) \geq \mathbb{P}\big(\Delta'_{\epsilon_1}\big) \geq 1 - \frac{42p^2}{(1-\rho_U)^2}(n-p)^{-\frac{1}{4\tau^*}}\left\{\frac{1}{\tau^*}\log(n-p)\right\}^{-\frac{1}{2}}.$$

Thus, $\Delta_{a,\epsilon_1} = \Delta_{t\not\subset a,\epsilon_1}$ when $G_t \not\subset G_a$ and $\Delta_{a,\epsilon_1} = \Delta_{t\subsetneq a,\epsilon_1}$ when $G_t \subsetneq G_a$. For the following proof, we restrict it to the event Δ_{a,ϵ_1}. Next, we consider two scenarios for Bayes factor consistency, that is, $G_t \not\subset G_a$ and $G_t \subsetneq G_a$.

First, when $G_t \not\subset G_a$ and $G_t \neq G_c$, we have $|E_t| > |E_a^1|$ and $|E_c| > |E_t|$. We begin by simplifying the upper bound of $\mathrm{BF}_{t\to c}$. (For $G_t = G_c$, $\mathrm{BF}_{t\to c} = 1$.) By Lemmas S.3.1 and A.4,

$$\begin{aligned}
\mathrm{BF}_{t\to c} &= \prod_{i=1}^{|E_c|-|E_t|} \mathrm{BF}\big(\widetilde{G}_i^{t\to c}; \widetilde{G}_{i-1}^{t\to c}\big) < \prod_{i=1}^{|E_c|-|E_t|}\left(\frac{g}{g+1}\right)\sqrt{\frac{b+n+d_{\widetilde{S}_i}}{b+d_{\widetilde{S}_i}-\frac{1}{2}}}\big(1-\hat{\rho}^2_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}\big)^{-\frac{n}{2}} \\
&< \left(\frac{2}{n}\right)^{\frac{|E_c|-|E_t|}{2}}\left\{1-\frac{\log(n-p)}{\tau^*(n-p)}\right\}^{-(|E_c|-|E_t|)\frac{n}{2}} \quad \text{when } n > b+p \\
&< \left(\frac{2}{n}\right)^{\frac{|E_c|-|E_t|}{2}}\exp\left(\frac{n}{n-p-1/\tau^*\log n}\cdot\frac{|E_c|-|E_t|}{2\tau^*}\cdot\log n\right) \\
&< \left(\frac{2}{n}\right)^{\frac{|E_c|-|E_t|}{2}}\exp\left(\frac{|E_c|-|E_t|}{\tau^*}\cdot\log n\right) \quad \text{when } n > 4p \\
&< \exp\left\{p^2 - \left(\frac{1}{2}-\frac{1}{\tau^*}\right)\big(|E_c|-|E_t|\big)\log n\right\}.
\end{aligned}$$

Next, we examine $\mathrm{BF}_{c\to a}$. Based on Lemma A.2 and its proof, we divide it into two parts, that is, true-edge deletion cases and false-edge deletion cases. For true-edge deletion cases, we

use $\{(\overline{x}_i^d, \overline{y}_i^d)\}_{i=1}^{|E_t|-|E_a^1|}$ to denote the sequence of true edges and $\{\overline{S}_i^d\}_{i=1}^{|E_t|-|E_a^1|}$ are the corresponding separator sequence. For false-edge deletion cases, we use $\{(\overline{x}_i^a, \overline{y}_i^a)\}_{i=1}^{|E_c|-|E_a|-|E_t|+|E_a^1|}$ and $\{\overline{S}_i^a\}_{i=1}^{|E_c|-|E_a|-|E_t|+|E_a^1|}$. Since p is finite, by the definition of ρ_L, then ρ_L is a positive finite constant.

$$
\begin{aligned}
\mathrm{BF}_{c\to a} &= \prod_{i=1}^{|E_c|-|E_a|} \mathrm{BF}\big(\overline{G}_i^{c\to a}; \overline{G}_{i-1}^{c\to a}\big) \\
&< \prod_{i=1}^{|E_t|-|E_a^1|} \left(1+\frac{1}{g}\right) \sqrt{\frac{b+d_{\overline{S}_i^d}}{b+n+d_{\overline{S}_i^d}-\frac{1}{2}}} \big(1-\hat{\rho}^2_{\overline{x}_i^d \overline{y}_i^d | \overline{S}_i^d}\big)^{\frac{n}{2}} \\
&\quad \times \prod_{i=1}^{|E_c|-|E_a|-|E_t|+|E_a^1|} \left(1+\frac{1}{g}\right) \sqrt{\frac{b+d_{\overline{S}_i^a}}{b+n+d_{\overline{S}_i^a}-\frac{1}{2}}} \big(1-\hat{\rho}^2_{\overline{x}_i^a \overline{y}_i^a | \overline{S}_i^a}\big)^{\frac{n}{2}} \\
&< \big\{2p(n+1)\big\}^{\frac{|E_c|-|E_a|}{2}} \big(1-\hat{\rho}^2_{\overline{x}_* \overline{y}_* | V\setminus\{\overline{x}_*, \overline{y}_*\}}\big)^{\frac{n}{2}} \quad \text{wlog assume } p > b \\
&< \big\{2p(n+1)\big\}^{\frac{|E_c|-|E_a|}{2}} \big\{1-\big(\epsilon_1 - |\rho_{\overline{x}_* \overline{y}_* | V\setminus\{\overline{x}_*, \overline{y}_*\}}|\big)^2\big\}^{\frac{n}{2}} \\
&< \big\{2p(n+1)\big\}^{\frac{|E_c|-|E_a|}{2}} \exp\left(-\frac{n\rho_L^2}{2} + n\epsilon_1 - \frac{n\epsilon_1^2}{2}\right) \\
&< \big\{2p(n+1)\big\}^{\frac{|E_c|-|E_a|}{2}} \exp\left\{-\frac{n\rho_L^2}{2} + \sqrt{n\log n} - \frac{1}{2\tau^*}\log(n-p)\right\} \quad \text{when } n > 2p \\
&< \exp\left\{-\frac{n\rho_L^2}{2} + p^2\log n + \sqrt{n\log n} - \frac{1}{2\tau^*}\log(n-p) + 2p^2\log p\right\} \quad \text{when } n > 1.
\end{aligned}
$$

Let $\delta(n) = p^2\log n + \sqrt{n\log n} + 3p^2\log p$ and $\delta(n)/n \to 0$ as $n \to \infty$. Hence,

$$
\mathrm{BF}(G_a; G_t \mid G_t \not\subset G_a) = \mathrm{BF}_{c\to a} \cdot \mathrm{BF}_{t\to c} < \exp\left\{-\frac{n\rho_L^2}{2} + \delta(n)\right\}.
$$

When $G_t \subsetneq G_a$, by Lemmas S.3.1 and A.4 we have

$$
\mathrm{BF}(G_a; G_t \mid G_t \subsetneq G_a) = \prod_{i=1}^{|E_a|-|E_t|} \mathrm{BF}\big(\widetilde{G}_i^{t\to a}; \widetilde{G}_{i-1}^{t\to a}\big) < \exp\left\{p^2 - \left(\frac{1}{2} - \frac{1}{\tau^*}\right)\big(|E_a| - |E_t|\big)\log n\right\}.
$$

A.2. Proof of Theorem 4.2

From $\lambda < \frac{1}{2} - \alpha$, we have $\alpha + \lambda < \frac{1}{2}$; from $\lambda < \frac{1}{4}$, we have $2\lambda < \frac{1}{2}$. For any β^* that satisfies

$$
\max\left\{2\lambda, \alpha+\lambda, \frac{1-\gamma}{2}\right\} < \beta^* < \frac{1}{2},
$$

let $\epsilon_{2,n} = (n-p)^{-\beta^*}$. Then define

$$
R''_{ij|S} = \big\{|\hat{\rho}_{ij|S} - \rho_{ij|S}| < \epsilon_{2,n}\big\}.
$$

Given any decomposable graph $G_a \neq G_t$, when $G_t \not\subset G_a$, by Lemma A.2, we have the edge sequence $\{(\overline{x}_i, \overline{y}_i)\}_{i=1}^{|E_c|-|E_a|}$ for moving from G_c to G_a and let $(\overline{x}_1, \overline{y}_1) = (\overline{x}_*, \overline{y}_*)$ be the first in the sequence where a true edge is deleted from G_c. Let $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{|E_c|-|E_t|}$ and $\{\widetilde{S}_i\}_{i=1}^{|E_c|-|E_t|}$ be the edge sequence and the corresponding separator sequence for moving from G_t to G_c according to Lemma A.4. Let

$$\Delta_{t\not\subset a,\epsilon_2}(n) = \big(R''_{\overline{x}_*\overline{y}_*|V\setminus\{\overline{x}_*,\overline{y}_*\}}\big) \cap \left(\bigcap_{i=1}^{|E_c|-|E_t|} R''_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}\right).$$

Since $0 < \beta^* < \frac{1}{2}$ and Assumption 4.5, by Lemma S.1.2, when $n \to \infty$,

$$\mathbb{P}\big\{\Delta_{t\not\subset a,\epsilon_2}(n)\big\} \geq \mathbb{P}\big\{\Delta'_{\epsilon_2}(n)\big\} \geq 1 - \frac{42p^2}{(1-\rho_U)^2}(n-p)^{\beta^*-\frac{1}{2}} \exp\left\{-\frac{1}{4}(n-p)^{1-2\beta}\right\} \to 1.$$

When $G_t \subsetneq G_a$, let $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{|E_a|-|E_t|}$ and $\{\widetilde{S}_i\}_{i=1}^{|E_a|-|E_t|}$ be the edge sequence and the corresponding separator sequence for moving from G_t to G_a according to Lemma A.4. (Notice here we use the same edge and separator notations as in G_t to G_c for consistency reason and G_t to G_a can be seen as a part of G_t to G_c.) Let

$$\Delta_{t\subsetneq a,\epsilon_2}(n) = \bigcap_{i=1}^{|E_a|-|E_t|} R''_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}.$$

Since $0 < \beta^* < \frac{1}{2}$ and Assumption 4.5, by Lemma S.1.2, when $n \to \infty$,

$$\mathbb{P}\big\{\Delta_{t\subsetneq a,\epsilon_2}(n)\big\} \geq \mathbb{P}\big\{\Delta'_{\epsilon_2}(n)\big\} \geq 1 - \frac{42p^2}{(1-\rho_U)^2}(n-p)^{\beta^*-\frac{1}{2}} \exp\left\{-\frac{1}{4}(n-p)^{1-2\beta}\right\} \to 1.$$

Thus, $\Delta_{a,\epsilon_2}(n) = \Delta_{t\not\subset a,\epsilon_2}(n)$ when $G_t \not\subset G_a$ and $\Delta_{a,\epsilon_2}(n) = \Delta_{t\subseteq a,\epsilon_2}(n)$ when $G_t \subsetneq G_a$. For the following proof, we restrict it to the event $\Delta_{a,\epsilon_2}(n)$. Similar to the proof of Theorem 4.1, we consider two scenarios here for posterior ratio consistency, i.e., $G_t \not\subset G_a$ and $G_t \subsetneq G_a$.

First, when $G_t \not\subset G_a$ and $G_t \neq G_c$, we have $|E_t| > |E_a^1|$ and $|E_c| > |E_t|$. (For $G_t = G_c$, $\mathrm{BF}_{t\to c} = 1$.) By Lemmas S.3.1 and A.4,

$$\begin{aligned}
\mathrm{BF}_{t\to c} &= \prod_{i=1}^{|E_c|-|E_t|} \mathrm{BF}\big(\widetilde{G}_i^{t\to c}; \widetilde{G}_{i-1}^{t\to c}\big) < \left(\frac{2}{n}\right)^{\frac{|E_c|-|E_t|}{2}} \big\{1-(n-p)^{-2\beta^*}\big\}^{-(|E_c|-|E_t|)\frac{n}{2}} \quad \text{when } n > b+p \\
&< \left(\frac{2}{n}\right)^{\frac{|E_c|-|E_t|}{2}} \left\{1+\frac{2}{(n-p)^{2\beta^*}}\right\}^{(|E_c|-|E_t|)\frac{n}{2}} \quad \text{when } n > \max\big\{2p, 2^{1/(2\beta^*)+1}\big\} \\
&< \exp\left\{\frac{np^2}{(n-p)^{2\beta^*}} - \frac{|E_c|-|E_t|}{4}\log n\right\} \quad \text{when } n > 4.
\end{aligned}$$

Similar to the proof of Theorem 4.1, we have

$$\begin{aligned}
\mathrm{BF}_{c\to a} &= \prod_{i=1}^{|E_c|-|E_a|} \mathrm{BF}\big(\overline{G}_i^{c\to a}; \overline{G}_{i-1}^{c\to a}\big) < \big\{2p(n+1)\big\}^{\frac{|E_c|-|E_a|}{2}} \big(1-\hat{\rho}^2_{\overline{x}_*\overline{y}_*|V\setminus\{\overline{x}_*,\overline{y}_*\}}\big)^{\frac{n}{2}} \quad \text{when } p > b \\
&< \big\{2p(n+1)\big\}^{\frac{|E_c|-|E_a|}{2}} \exp\left(-\frac{n\rho_L^2}{2} + n\epsilon_2 - \frac{n\epsilon_2^2}{2}\right)
\end{aligned}$$

$$< \{2p(n+1)\}^{\frac{|E_c|-|E_a|}{2}} \exp\left\{-\frac{n\rho_L^2}{2} + \frac{n}{(n-p)^{\beta^*}} - \frac{1}{2}n^{1-2\beta^*}\right\}$$

$$< \exp\left\{-\frac{n\rho_L^2}{2} + \frac{n}{(n-p)^{\beta^*}} - \frac{1}{2}n^{1-2\beta^*} + 3p^2\log n\right\}.$$

When $n > 3\exp\{(1-2\beta^*)^{-2}\}$, we have $n(n-p)^{-2\beta^*} > 3\log n$. Hence,

$$\mathrm{BF}(G_a; G_t \mid G_t \not\subset G_a) < \exp\left\{-\frac{n\rho_L^2}{2} + \frac{n}{(n-p)^{\beta^*}} - \frac{1}{2}n^{1-2\beta^*} + \frac{2np^2}{(n-p)^{2\beta^*}}\right\}.$$

Therefore, when $G_t \not\subset G_a$, for $n > (\log 2/C_q)^{1/\gamma}$,

$$\mathrm{PR}(G_a; G_t \mid G_t \not\subset G_a) < \exp\left\{-\frac{n\rho_L^2}{2} + \frac{n}{(n-p)^{\beta^*}} - \frac{1}{2}n^{1-2\beta^*} + \frac{2np^2}{(n-p)^{2\beta^*}} + \big(|E_a| - |E_t|\big)\log(2q)\right\}.$$

By the construction of β^*, we have

$$1 - 2\lambda > \max\big\{2\alpha, 1-2\beta^*, 1-\beta^*, 1+2\alpha-2\beta^*\big\},$$

and $1 - 2\lambda > \sigma + \gamma$. Therefore, $-n\rho_L^2/2$ is the leading term in the upper bound of $\mathrm{PR}(G_a; G_t \mid G_t \not\subset G_a)$. Thus, $\mathrm{PR}(G_a; G_t) \to 0$, as $n \to \infty$ when $G_t \not\subset G_a$.

When $G_t \subsetneq G_a$, by Lemmas S.3.1 and A.4, we have

$$\mathrm{BF}(G_a; G_t \mid G_t \subsetneq G_a) < \exp\left\{\frac{(|E_a| - |E_t|)n}{(n-p)^{2\beta^*}}\right\},$$

$$\mathrm{PR}(G_a; G_t \mid G_t \subsetneq G_a) < \exp\left\{\frac{(|E_a| - |E_t|)n}{(n-p)^{2\beta^*}} + \big(|E_a| - |E_t|\big)\log(2q)\right\}.$$

Since $\beta^* > \frac{1-\gamma}{2}$, then $(|E_a| - |E_t|)\log(2q)$ is the leading term above and $|E_a| - |E_t| > 0$. Therefore, $\mathrm{PR}(G_a; G_t | G_t \subsetneq G_a) \to 0$, as $n \to \infty$.

Appendix B: Equivalence of minimal triangulations when G_t is not decomposable

Let $G_m = (V, E_m)$ be any minimal triangulation of G_t, where $E_m = E_t \cup F$, $F \neq \varnothing$. In here G_a denotes any decomposable graph other than minimal triangulations of G_t. Since G_m is a minimal triangulation, then $E_a \neq E_t \cup F'$, where $F' \subseteq F$. Different from when G_t is decomposable, there are three cases here: (1) $|E_a^1| < |E_m^1| = |E_t|$, thus $G_m \not\subset G_a$; (2) $|E_a^1| = |E_m^1| = |E_t|$ and $G_m \subsetneq G_a$; (3) $|E_a^1| = |E_m^1| = |E_t|$ and $G_m \not\subset G_a$. But in case (3) there exists at least one minimal triangulation of G_t which is a subset of G_a. And in both (2) and (3), we have $|E_m| < |E_a|$.

For case (1), when $|E_a^1| < |E_m^1| = |E_t|$, i.e., one of the two cases where $G_m \not\subset G_a$, we inherit all notations from Lemma A.2, $\{\overline{x}_i, \overline{y}_i\}_{i=1}^{|E_c|-|E_a|}$ is the edge sequence from G_c to G_a and $\{\rho_{\overline{x}_i\overline{y}_i|\overline{S}_i}\}_{i=1}^{|E_c|-|E_a|}$ is the corresponding population partial correlation sequence. And Lemma A.2 still holds here, that is, we can find a sequence of partial correlations $\{\rho_{\overline{x}_i\overline{y}_i|\overline{S}_i}\}_{i=1}^{|E_c|-|E_a|}$ which has at least one population partial correlation corresponding to the removal of a true edge is non-zero and not a correlation. The proof carries out the same as in Lemma A.2, just let the first step of moving from G_c to G_a be the

deletion of a true edge which is missing in G_a. For case (3), where $|E_a^1| = |E_m^1| = |E_t|$ but $G_m \not\subset G_a$, when moving from G_c to G_a, all steps are false-edge deletion cases. There is no true-edge deletion case here since G_a has all the true edges of G_t.

For case (2), when $G_m \subsetneq G_a$ and $|E_a^1| = |E_m^1| = |E_t|$, we still use $\{(\widetilde{x}_i, \widetilde{y}_y)\}_{i=1}^{|E_a|-|E_m|}$ to denote the sequence of edges which are added in each steps from G_m to G_a and $\{\rho_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}\}_{i=1}^{|E_a|-|E_m|}$ is the corresponding population partial correlation sequence. A similar version of Lemma A.4 still holds here.

Lemma B.1. *For any edge sequence* $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{|E_a|-|E_m|}$ *from* G_m *to* G_a *described above in case* (2)*, all population partial correlations in* $\{\rho_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}\}_{i=1}^{|E_a|-|E_m|}$ *are zero* (*or correlation, when* $\widetilde{S}_i = \varnothing$).

Proof. This proof follows similarly to the proof of Lemma A.4. Assume in the ith step we add edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ to graph $\widetilde{G}_{i-1}^{m\to a}$ and $\widetilde{S}_i$ is the corresponding separator.

When $\widetilde{S}_i \neq \varnothing$. Since adding edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ to graph $\widetilde{G}_{i-1}^{m\to a}$ maintains decomposability of graph $\widetilde{G}_i^{m\to a}$. By Lemma S.3.1, $\widetilde{x}_i$ and $\widetilde{y}_i$ are in two cliques which are adjacent in the current junction tree of $\widetilde{G}_{i-1}^{m\to a}$. Thus, by the property of junction trees, we know $\widetilde{S}_i$ separates $\widetilde{x}_i$ from $\widetilde{y}_i$ in $\widetilde{G}_{i-1}^{m\to a}$. Since this is an increasing sequence in terms of edges from G_m to G_a, thus $G_m \subsetneq \widetilde{G}_{i-1}^{m\to a}$. And due to the minimal triangulation, $G_t \subsetneq G_m \subsetneq \widetilde{G}_{i-1}^{m\to a}$. By Lemma A.3, $\widetilde{S}_i$ separates node $\widetilde{x}_i$ from $\widetilde{y}_i$ in G_t, $\rho_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i} = 0$.

When $\widetilde{S}_i = \varnothing$, $\widetilde{x}_i$ and $\widetilde{y}_i$ are disconnected in the current graph $\widetilde{G}_{i-1}^{m\to a}$. Then they are also disconnected in G_t. Thus, they are marginally independent in G_t, $\rho_{\widetilde{x}_i\widetilde{y}_i} = 0$. □

Remark B.1. For $|E_a^1| = |E_t|$ and $|E_a| - |E_a^1| = 0, \ldots, |F| - 1$, no decomposable G_a exists; for $|E_a^1| = |E_t|$ and $|E_a| - |E_a^1| > |F|$, at least one decomposable G_a exists; but for $|E_a^1| < |E_t|$ and $|E_a| - |E_a^1| \geq 0$, a decomposable G_a may not exist. The Bayes factor $\mathrm{BF}(G_a; G_m)$ under $|E_a^1| < |E_t|$ and $|E_a| - |E_a^1| \geq 0$ is only valid when a decomposable G_a exists, otherwise it is defined to be zero.

B.1. Proof of Theorem 5.1

Part 1. For any given decomposable graph G_a that is not a minimal triangulation of G_t, let

$$\tau^* > \max\left\{2, \frac{2(|E_c| - |E_m|)}{|E_a| - |E_m|}\right\}.$$

The construction of Δ_{a,ϵ_1} is the same as in the proof of Theorem 4.1. After that, we restrict the following proof to the set Δ_{a,ϵ_1}. For case (1), when $|E_a^1| < |E_m^1| = |E_t|$, we have

$$\mathrm{BF}_{m\to c} < \exp\left\{p^2 - \left(\frac{1}{2} - \frac{1}{\tau^*}\right)\big(|E_c| - |E_m|\big)\log n\right\} \to 0,$$

$$\mathrm{BF}_{c\to a} < \exp\left\{-\frac{n\rho_L^2}{2} + p^2\log n + \sqrt{n\log n} - \frac{1}{2\tau^*}\log(n-p) + 2p^2\log p\right\} \to 0.$$

Hence,

$$\mathrm{BF}\big(G_a; G_m \mid G_m \not\subset G_a, \big|E_a^1\big| < \big|E_m^1\big|\big) = \mathrm{BF}_{c\to a}\cdot\mathrm{BF}_{m\to c} \to 0.$$

For case (2), when $G_m \subsetneq G_a$, that is, $|E_a^1| = |E_m^1| = |E_t|$ and $|E_a| > |E_m|$, we have

$$\mathrm{BF}(G_a; G_m \mid G_m \subsetneq G_a) < \exp\left\{p^2 - \left(\frac{1}{2} - \frac{1}{\tau^*}\right)\big(|E_a| - |E_m|\big)\log n\right\} \to 0.$$

For case (3), when $|E_a^1| = |E_m^1| = |E_t|$ and $G_m \not\subset G_a$, also $|E_a| > |E_m|$, we have

$$\mathrm{BF}_{m\to c} < 2^{p^2} n^{-\frac{|E_c|-|E_a|}{2}} \exp\left[-\left\{\frac{|E_a|-|E_m|}{2(|E_c|-|E_m|)} - \frac{1}{\tau^*}\right\}\big(|E_c|-|E_m|\big)\log n\right],$$

$$\mathrm{BF}_{c\to a} < (4p)^{p^2} n^{\frac{|E_c|-|E_a|}{2}} \quad \text{when } n > 1.$$

Hence,

$$\begin{aligned}&\mathrm{BF}\big(G_a; G_m \mid G_m \not\subset G_a, \big|E_a^1\big| = \big|E_m^1\big|\big)\\ &\quad < (8p)^{p^2} \exp\left[-\left\{\frac{|E_a|-|E_m|}{2(|E_c|-|E_m|)} - \frac{1}{\tau^*}\right\}\big(|E_c|-|E_m|\big)\log n\right] \to 0.\end{aligned}$$

Therefore, $\mathrm{BF}(G_a; G_m) \to 0$, as $n \to \infty$.

Part 2. Let $G_{m_1}, G_{m_2}, \ldots, G_{m_l}$ be all the minimal triangulations of G_t, where l is a positive finite integer, since the graph dimension is finite. By Part 1, on the set Δ_{a,ϵ_1}, $\mathrm{BF}(G_{m_i}; G_a) \to \infty$, $i = 1, 2, \ldots, l$, where $G_a \notin \mathcal{M}_t$. Therefore,

$$\begin{aligned}\sum_{G_m\in\mathcal{M}_t} \pi(G_m \mid \mathrm{Y}) &= \frac{\sum_{i=1}^{l} p(\mathrm{Y} \mid G_{m_i})}{\sum_{i=1}^{l} p(\mathrm{Y} \mid G_{m_i}) + \sum_{G_a\notin\mathcal{M}_t} p(\mathrm{Y} \mid G_a)}\\ &= \frac{1}{1 + \sum_{G_a\notin\mathcal{M}_t} \frac{1}{\sum_{i=1}^{l} \mathrm{BF}(G_{m_i}; G_a)}} \to 1 \quad \text{as } n \to \infty.\end{aligned}$$

B.2. Proof of Theorem 5.2

Let $\{\hat{\rho}_{m_1,i}\}_{i=1}^{|E_c|-|E_{m_1}|}$ and $\{\rho_{m_1,i}\}_{i=1}^{|E_c|-|E_{m_1}|}$ be the sample and population partial correlation sequence corresponding to each step from G_{m_1} to G_c. By Lemma B.1, $\rho_{m_1,i} = 0$, $i = 1, 2, \ldots, |E_c| - |E_{m_1}|$. By Lemma S.1.4, for any $0 < \epsilon < 1$, there exist $0 < M_1(\epsilon) < 1/4$ and $M_2(\epsilon) > 3$ (the choice of M_1 and M_2 is the same as in the proof of Lemma S.1.4), we have $\mathbb{P}(\Delta_\epsilon^0) > 1 - \epsilon/2$, for $n > p + 3$. Let

$$R_{m_1,i} = \left\{\frac{M_1}{n} < \hat{\rho}_{m_1,i}^2 < \frac{M_2}{n-p}\right\}, \qquad \Delta_{m_1} := \bigcap_{i=1}^{|E_c|-|E_{m_1}|} R_{m_1,i}.$$

Then

$$\mathbb{P}(\Delta_{m_1}) \geq \mathbb{P}\big(\Delta_\epsilon^0\big) \geq 1 - \epsilon/2.$$

By Lemma S.3.1, when $n > b + p$, we have

$$\begin{aligned}&\left(\frac{1}{2n}\right)^{\frac{|E_c|-|E_{m_1}|}{2}} \prod_{i=1}^{|E_c|-|E_{m_1}|} \big(1-\hat{\rho}_{m_1,i}^2\big)^{-\frac{n(|E_c|-|E_{m_1}|)}{2}}\\ &\quad < \mathrm{BF}(G_c; G_{m_1}) < \left(\frac{2}{n}\right)^{\frac{|E_c|-|E_{m_1}|}{2}} \prod_{i=1}^{|E_c|-|E_{m_1}|} \big(1-\hat{\rho}_{m_1,i}^2\big)^{-\frac{n(|E_c|-|E_{m_1}|)}{2}}.\end{aligned}$$

Under the event Δ_{m_1}, when $n > p + M_2$,

$$\left(\frac{e^{M_1}}{2n}\right)^{\frac{|E_c|-|E_{m_1}|}{2}} < \mathrm{BF}(G_c; G_{m_1}) < \left(\frac{2e^{2M_2}}{n}\right)^{\frac{|E_c|-|E_{m_1}|}{2}}.$$

Thus we have

$$\mathbb{P}\left\{\left(\frac{e^{M_1}}{2n}\right)^{\frac{|E_c|-|E_{m_1}|}{2}} < \mathrm{BF}(G_c; G_{m_1}) < \left(\frac{2e^{2M_2}}{n}\right)^{\frac{|E_c|-|E_{m_1}|}{2}}\right\} > 1 - \frac{\epsilon}{2}.$$

Similarly,

$$\mathbb{P}\left\{\left(\frac{2e^{2M_2}}{n}\right)^{-\frac{|E_c|-|E_{m_1}|}{2}} < \mathrm{BF}(G_{m_2}; G_c) < \left(\frac{e^{M_1}}{2n}\right)^{-\frac{|E_c|-|E_{m_1}|}{2}}\right\} > 1 - \frac{\epsilon}{2}.$$

Therefore, let $A_1 = \frac{1}{4}e^{-M_2}$ and $A_2 = 4e^{2M_2 p^2}$, we have $\mathbb{P}\{A_1 < \mathrm{BF}(G_{m_1}; G_{m_2}) < A_2\} > 1 - \epsilon$.

Acknowledgements

Debdeep Pati acknowledges support from NSF DMS (1854731, 1916371) and NSF CCF 1934904 (HDR-TRIPODS). Bani K. Mallick acknowledges support from NIH R01CA194391 (NCI) and NSF CCF 1934904 (HDR-TRIPODS).

Supplementary Material

Supplement to "Bayesian graph selection consistency under model misspecification" (DOI: 10.3150/20-BEJ1253SUPP; .pdf). There are five sections in the Supplementary Materials [32]. In Section S.1, we provide a set of auxiliary results related to the concentration and tail behavior of partial correlations. The bounds for Bayes factors of local moves are developed in Section S.2 and Section S.3. For the proofs of Theorems 4.3, 5.3, 5.4 and Corollaries 4.3, 5.1, see Section S.4 and Section S.5.

References

[1] Armstrong, H., Carter, C.K., Wong, K.F.K. and Kohn, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Stat. Comput.* **19** 303–316. MR2516221 https://doi.org/10.1007/s11222-008-9093-8

[2] Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* **92** 317–335. MR2201362 https://doi.org/10.1093/biomet/92.2.317

[3] Banerjee, S. and Ghosal, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electron. J. Stat.* **8** 2111–2137. MR3273620 https://doi.org/10.1214/14-EJS945

[4] Banerjee, S. and Ghosal, S. (2015). Bayesian structure learning in graphical models. *J. Multivariate Anal.* **136** 147–162. MR3321485 https://doi.org/10.1016/j.jmva.2015.01.015

[5] Ben-David, E., Li, T., Massam, H. and Rajaratnam, B. (2011). High dimensional Bayesian inference for Gaussian directed acyclic graph models. Preprint. Available at arXiv:1109.4371.

[6] Bickel, P.J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 https://doi.org/10.1214/009053607000000758

[7] Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949 https://doi.org/10.1198/jasa.2011.tm10560
[8] Cao, X., Khare, K. and Ghosh, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Ann. Statist.* **47** 319–348. MR3909935 https://doi.org/10.1214/18-AOS1689
[9] Carvalho, C.M., Massam, H. and West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94** 647–659. MR2410014 https://doi.org/10.1093/biomet/asm056
[10] Carvalho, C.M. and Scott, J.G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96** 497–512. MR2538753 https://doi.org/10.1093/biomet/asp017
[11] Dawid, A.P. and Lauritzen, S.L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267 https://doi.org/10.1214/aos/1176349260
[12] Dellaportas, P., Giudici, P. and Roberts, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhyā* **65** 43–55. MR2016776
[13] Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238
[14] Dobra, A., Hans, C., Jones, B., Nevins, J.R., Yao, G. and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212. MR2064941 https://doi.org/10.1016/j.jmva.2004.02.009
[15] Donnet, S. and Marin, J.-M. (2012). An empirical Bayes procedure for the selection of Gaussian graphical models. *Stat. Comput.* **22** 1113–1123. MR2950089 https://doi.org/10.1007/s11222-011-9285-5
[16] Drton, M. and Perlman, M.D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statist. Sci.* **22** 430–449. MR2416818 https://doi.org/10.1214/088342307000000113
[17] El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011 https://doi.org/10.1214/07-AOS559
[18] Fitch, A.M., Jones, M.B. and Massam, H. (2014). The performance of covariance selection methods that consider decomposable models only. *Bayesian Anal.* **9** 659–684. MR3256059 https://doi.org/10.1214/14-BA874
[19] Giudici, P. (1996). Learning in graphical Gaussian models. In *Bayesian Statistics*, 5 (*Alicante*, 1994). *Oxford Sci. Publ.* 621–628. New York: Oxford Univ. Press. MR1425431
[20] Giudici, P. and Green, P.J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86** 785–801. MR1741977 https://doi.org/10.1093/biomet/86.4.785
[21] Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810 https://doi.org/10.1093/biomet/82.4.711
[22] Heggernes, P. (2006). Minimal triangulations of graphs: A survey. *Discrete Math.* **306** 297–317. MR2204109 https://doi.org/10.1016/j.disc.2005.12.003
[23] Johnstone, I.M. (2010). High dimensional Bernstein–von Mises: Simple examples. In *Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat. (IMS) Collect.* **6** 87–98. Beachwood, OH: IMS. MR2798513
[24] Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.* **20** 388–400. MR2210226 https://doi.org/10.1214/088342305000000304
[25] Khare, K., Rajaratnam, B. and Saha, A. (2018). Bayesian inference for Gaussian graphical models beyond decomposable graphs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 727–747. MR3849341 https://doi.org/10.1111/rssb.12276
[26] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459 https://doi.org/10.1214/09-AOS720
[27] Lauritzen, S.L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. New York: The Clarendon Press. MR1419991
[28] Lee, K., Lee, J. and Lin, L. (2019). Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse Cholesky factors. *Ann. Statist.* **47** 3413–3437. MR4025747 https://doi.org/10.1214/18-AOS1783
[29] Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35** 1278–1323. MR2341706 https://doi.org/10.1214/009053606000001235

[30] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[31] Moghaddam, B., Khan, E., Murphy, K.P. and Marlin, B.M. (2009). Accelerating Bayesian structural inference for non-decomposable Gaussian graphical models. In *Advances in Neural Information Processing Systems* 1285–1293.

[32] Niu, Y., Pati, D. and Mallick, B.K. (2020). Supplement to "Bayesian graph selection consistency under model misspecification." https://doi.org/10.3150/20-BEJ1253SUPP

[33] Rajaratnam, B., Massam, H. and Carvalho, C.M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.* **36** 2818–2849. MR2485014 https://doi.org/10.1214/08-AOS619

[34] Raskutti, G., Yu, B., Wainwright, M.J. and Ravikumar, P.K. (2009). Model selection in Gaussian graphical models: High-dimensional consistency of lregularized MLE. In *Advances in Neural Information Processing Systems* 1329–1336.

[35] Rose, D.J., Tarjan, R.E. and Lueker, G.S. (1976). Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.* **5** 266–283. MR0408312 https://doi.org/10.1137/0205021

[36] Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika* **87** 99–112. MR1766831 https://doi.org/10.1093/biomet/87.1.99

[37] Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Stat.* **29** 391–411. MR1925566 https://doi.org/10.1111/1467-9469.00297

[38] Scott, J.G. and Carvalho, C.M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comput. Graph. Statist.* **17** 790–808. MR2649067 https://doi.org/10.1198/106186008X382683

[39] Spokoiny, V. (2013). Bernstein–von Mises theorem for growing parameter dimension. Preprint. Available at arXiv:1302.3430.

[40] Uhler, C., Lenkoski, A. and Richards, D. (2018). Exact formulas for the normalizing constants of Wishart distributions for graphical models. *Ann. Statist.* **46** 90–118. MR3766947 https://doi.org/10.1214/17-AOS1543

[41] Wang, H. and Carvalho, C.M. (2010). Simulation of hyper-inverse Wishart distributions for non-decomposable graphs. *Electron. J. Stat.* **4** 1470–1475. MR2741209 https://doi.org/10.1214/10-EJS591

[42] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824 https://doi.org/10.1093/biomet/asm018

Received March 2020 and revised July 2020