TOPICS ON BAYESIAN GAUSSIAN GRAPHICAL MODELS

A Dissertation

by

YABO NIU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Bani K. Mallick
Co-Chair of Committee,	Debdeep Pati
Committee Members,	Anirban Bhattacharya
	Huiyan Sang
	Yang Ni
	Yu Ding
Head of Department,	Jianhua Huang

August 2019

Major Subject: Statistics

ABSTRACT

Gaussian graphical models (GGMs) 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 and characterize uncertainty in the selection.

In this study, I first develop a Bayesian method to incorporate covariate information in the GGMs setup in a nonlinear seemingly unrelated regression framework. I propose a joint predictor and graph selection model and develop an efficient collapsed Gibbs sampler algorithm to search the joint model space. Furthermore, I investigate its theoretical variable selection properties. I demonstrate the proposed method on a variety of simulated data, concluding with a real data set from The Cancer Proteome Atlas (TCPA) project.

For scalability of the Markov chain Monte Carlo algorithms, decomposability is commonly imposed on the graph space. A wide variety of graphical conjugate priors are proposed jointly on the covariance matrix and the graph with improved algorithms to search along the space of decomposable graphs, rendering the methods extremely popular in the context of multivariate dependence modeling. An open problem in Bayesian decomposable structure learning is whether the posterior distribution is able to select a meaningful decomposable graph that it is “close” in an appropriate sense to the true non-decomposable graph, when the dimension of the variables increases with the sample size.

In the second part of this study, I explore specific conditions on the true precision matrix and the graph which results in an affirmative answer to this question using a commonly used hyper-inverse Wishart prior on the covariance matrix and a suitable complexity prior on the graph space, both in the well-specified and misspecified settings. In absence of structural sparsity assumptions, the strong selection consistency holds in a high dimensional setting where $p = O(n^{\alpha})$ for $\alpha < 1/3$. I show when the true graph is non-decomposable, the posterior distribution on the graph concentrates on a set of graphs that are minimal triangulations of the true graph.

DEDICATION

This dissertation is dedicated to

My mother, Hui Tang and my father, Jingui Niu who have raised me to be the person I am today;

My committee chairs, Dr. Bani K. Mallick and Dr. Debdeep Pati who have been guiding and supporting me through my entire Ph.D. study. I am so grateful to have them as my advisors.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my Ph.D. advisors Dr. Bani K. Mallick and Dr. Debdeep Pati for their support and guidance during my Ph.D. study. Their knowledge helped me go through countless obstacles in my research; their passion for Statistics inspired me to pursuit more challenging problems; their patience nourished me to grow stronger; their wisdom taught me to be a better researcher; their guidance leaded my way to the finish line like the lighthouse on the sea. The lessons and skills I acquired from them are beyond measure. I am so grateful to have them as my advisors, as my teachers and as my lifelong mentors. As an old Chinese idiom says, "Teacher for one day, father for ever ". They are like my family. I am forever in their debts and I owe my future career to them.

I would also like to thank my committee memebers Dr. Anirban Bhattachary, Dr. Huiyan Sang, Dr. Yang Ni, and Dr. Yu Ding for their support and guidance in my research. Thanks are also given to my collaborator Dr. Nilabja Guha for his help in proving the variable selection consistency. His training and patience helped me to learn and grow in my early research days. The amount of time he dedicated to guide me was tremendous. I will always cherish the time we worked together and it is an invaluable experience in my life.

My special thanks go to the Associate Department Head Dr. Michael Longnecker for his mentoring when I was an instructor of STAT 302. His advices and suggestions leaded me to be a better teacher. I would like to thank the Department of Statistics for providing me a healthy, friendly and supportive research environment to successfully complete my Ph.D. study. I would also like to thank Texas A&M University for creating such an amazing campus. I cannot remember how many times that ideas were floating in my mind while I was walking through such gorgeous campus during late night. I will forever be a grateful Aggie.

Last but not least, none of these would be possible without the support of my beloved mother, Hui Tang and my beloved father, Jingui Niu. Their honesty, integrity and selflessness taught me to be the person who I am today. Their encouragement, persistence and thoughtfulness keep me marching forward in my research and in my life. I cannot imagine the sacrifice and compromise they made to their lives in order to create a better life for me. I am indebted to them forever and I am honored to be their son. I hope one day I make them proud.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Bani K. Mallick and Dr. Debdeep Pati and Dr. Anirban Bhattacharya, Dr. Huiyan Sang, Dr. Yang Ni of the Department of Statistics and Dr. Yu Ding of the Department of Industrial and Systems Engineering.

The data analyzed for Chapter 2 was provided by Dr. Veerabhadran Baladandayuthapani of the Department of Biostatistics at University of Michigan. The proof of Bayes factor consistency in Chapter 2 was a collaboration with Dr. Nilabja Guha at the Department of Mathematical Science at University of Massachusetts, Lowell.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by the Department of Statistics at Texas A&M University.

NOMENCLATURE

GGM	Gaussian Graphical model
BF, PR	Bayes factor, posterior ratio
HIW	hyper-inverse Wishart distribution/prior
MCMC	Markov chain Monte Carlo
SUR	seemingly unrelated regression
DAG	directed acyclic graph
i.i.d.	independent and identically distributed
$\mathbb{P}$	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
$a \asymp b$	$C_1 a \leq b \leq C_2 a$ for constants C_1, C_2
$a \precsim b$	$a \leq C_3 b$ for a constant C_3
$A \subset B$, $A \not\subset B$	A is a subset of B, A is not a subset of B
$A \subsetneq B$	$A \subset B$ and $A \neq B$
$\lvert\cdot\rvert$	absolute value, cardinality of sets or determinant of matrices by context
$\pi(\cdot)$, $\pi(\cdot \mid \mathrm{Y})$	prior distribution and posterior distribution of graphs
Y, Y_i^T, y_i	$n \times p$ data matrix, row of Y, column of Y
ρ_{ij}, $\rho_{ij\mid S}$	correlation and partial correlation between X_i and X_j given X_S

$\hat{\rho}_{ij}$, $\hat{\rho}_{ij\mid S}$	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 $\rho_{ij\mid V\setminus\{i,j\}}$, where $(i,j) \in E_t$
C_i, $\mathcal{C}(\mathscr{C})$, S_i, $\mathcal{S}(\mathscr{S})$	clique, set of cliques, separator, set of separators
G_t, G_a, G_c	the true graph, any decomposable graph, the complete graph
G_m, G_0	the minimal triangulation when G_t is non-decomposable, empty graph
$\hat{G}$	posterior mode in the decomposable graph space
E_t, E_a, E_c, E_a^1	edge set of G_t, G_a, G_c and $E_a^1 = E_a \cap E_t$
p, V	graph dimension, 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	cardinality of separator S, prior edge inclusion probability
Δ'_ϵ, $\Delta'_\epsilon(n)$, $\Delta''_\epsilon(n)$	probability regions of sample partial correlations
Π_{xy}	the set of all sets that separates node 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

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION [*]

Probabilistic graphical models provide a helpful tool to describe and visualize the dependence structures among random variables. Graphical models which describe conditional dependencies can provide insights into properties and relationships between random variables. A graph comprises of vertices (nodes) connected by edges (links or arcs). In a probabilistic graphical model, the random variables (single or vector) are represented by the vertices and probabilistic relationships between these variables are expressed by the edges. An edge may or may not carry directional information. In this dissertation we concentrate on undirected Gaussian graphical models (GGMs) where the edges do not carry any directional information. Furthermore in this model, the variables follow a multivariate normal distribution with a particular structure on the inverse of the covariance matrix, called the precision or the concentration matrix.

1.1 Undirected Decomposable Graphs

Denote an undirected graph by $G = (V, E)$ with a vertex set $V = \{1, 2, \ldots, q\}$ and an edge set $E = \{(r, s) : e_{rs} = 1, 1 \leq r < s \leq q\}$ with $e_{rs} = 1$ if the edge (r, s) is present in G and 0 otherwise. 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, maximally complete subgraph. See [1] for more graph related terminologies.

[*] Reprinted with permission from arXiv, “Bayesian Graph Selection Consistency Under Model Misspecification”, arXiv preprint arXiv:1901.04134, 2019, by Niu, Yabo and Pati, Debdeep and Mallick, Bani K. In accordance arXiv copyright no modifications have been made except formatting.

We shall focus on decomposable graphs in this dissertation. A graph is decomposable [1] 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 [1] 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 $\mathcal{S} = \{S_i\}_{i=2}^{k}$, $(C_1, S_2, C_2, S_3, \dots C_k)$, is said to be perfect if for every $i = 2, 3, \dots, k$ the running intersection property [1] (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 $H_{i-1} = \cup_{j=1}^{i-1} C_j$. A junction tree for the 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. 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 nodes. Denote by $\mathcal{D}_k$ the space of all decomposable graphs on k notes. Figures 1.1 and 1.2 briefly illustrate a decomposable and a non-decomposable graph, both defined on 6 nodes.

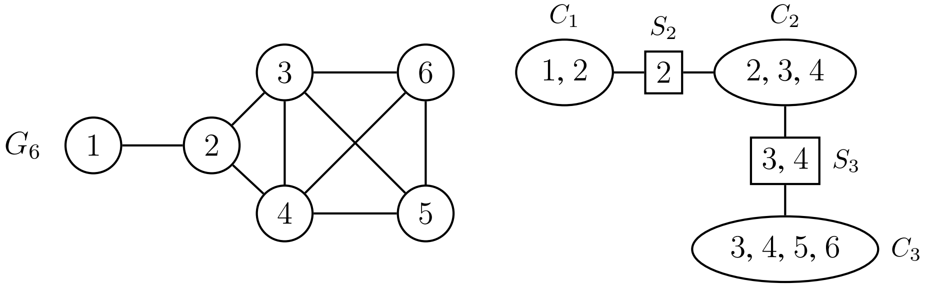


Figure 1.1: G_6 is a 6-node decomposable graph and its junction tree decomposition (right) has 3 **cliques** and 2 separators, i.e. $C_1 = \{1, 2\}$, $S_2 = \{2\}$, $C_2 = \{2, 3, 4\}$, $S_3 = \{3, 4\}$, $C_3 = \{3, 4, 5, 6\}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

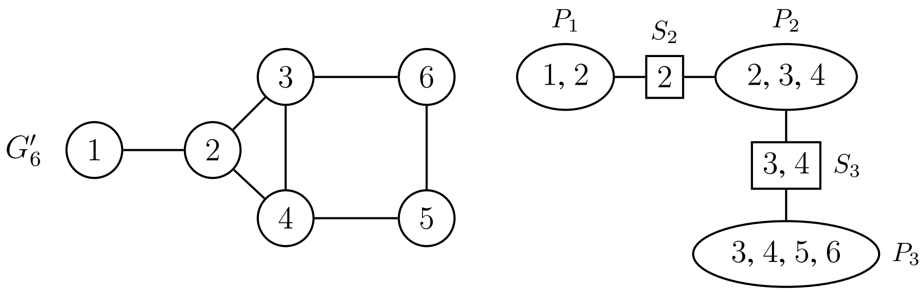


Figure 1.2: G_6' is a 6-node non-decomposable graph because its cycle of four, $3-4-5-6$, does not have a cord. Its junction tree decomposition (right) has 3 **prime** components and 2 separators, i.e. $P_1=\{1,2\}$, $S_2=\{2\}$, $P_2=\{2,3,4\}$, $S_3=\{3,4\}$, $P_3=\{3,4,5,6\}$. Out of all prime components only P_1 and P_2 are cliques. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

1.2 Gaussian Graphical Models

Assume

$$\mathrm{y}|\Sigma_G, G \sim \mathrm{N}_q(0, \Sigma_G),$$

where $\mathrm{y}=(y_1,y_2,\ldots,y_q)$ and $\Sigma^{-1}=(\sigma^{ij})_{q\times q}$. The conditional dependencies lie in the precision matrix which is the inverse of the covariance matrix. Therefore, y_i and y_j are conditionally independent given the rest of the variables if and only if $\sigma^{ij}=0$, where $i\neq j$. This property induces a unique undirected graph corresponding to each multivariate Gaussian distribution. Thus, q random variables represent q nodes and if G is the adjacency graph pairing to the precision matrix, then the presence of an off-diagonal edge between two nodes implies non-zero partial correlation (i.e., conditional dependence) and the absence of an edge implies conditional independence.

1.3 Literature Review

Graphical models provide a framework for describing statistical dependencies in (possibly large) collections of random variables [1]. In this dissertation, we revisit the well known problem of inference on the underlying graph from observed data from a Bayesian point of view. Research on Bayesian inference for natural exponential families and associated conjugate priors (DY priors) is pioneered by [2] and has 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 p-variate normal distribution with mean vector 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 distribution [3] on $\Sigma = \Omega^{-1}$ and the corresponding induced class of distributions on Ω [4] are attractive choices of DY priors. A rich family of conjugate priors that subsumes the DY class is developed by [5]. Bayesian procedures corresponding to these Letac-Massam priors have been derived in a decision theoretic framework in the recent work of [6]. The key component of Bayesian structure learning is achieved through 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. In the interests of tractability and scalability, there has been a strong focus on the case where the true graph may be assumed to be decomposable. On the other hand, relatively few papers have considered non decomposable graphs in a Bayesian set-up; refer to HIW distributions for non-decomposable graphs [7, 8, 9, 10, 11, 12].

In this dissertation, we focus on the HIW distribution for decomposable graphs as this construction enjoys many advantages, such as computational efficiency due to its conjugate formulation and exact calculation of marginal likelihoods [13]. The use of HIW prior within a Bayesian framework for Gaussian graphical models has been well studied for the past decade, see [14, 15, 16, 17]. Although deemed as a restrictive model choice in the space of graphs, as long as the model for the data allows arbitrarily small interactions, the resulting model assuming decomposability is quite flexible. Stochastic search algorithms are empirically demonstrated to have good practical performance in these models. For detailed description and comparison of various Bayesian computation methods in this scenario, see [18, 19].

There has been a growing literature on model selection consistency in Gaussian graphical models from a frequentist point of view [20, 21, 22, 23]. Beyond the literature on Gaussian graphical models, there has been a incredible amount of frequentist work in the context of estimating high-dimensional covariance matrix estimation with rates of convergence of various regularized covariance estimators derived in [24, 25, 26, 27] among others. There is a relatively smaller literature on asymptotic properties of Bayesian procedures for covariance or precision matrices in graphical models; refer to [28, 29]. However, the literature on graph selection consistency in a Bayesian paradigm is surprisingly sparse. In the context of decomposable graphs, the only article we were aware of is [30] who 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 by an edge. For general graph selection consistency within a Bayesian framework, refer to the very recent article [31] in the context of Gaussian directed acyclic graph (DAG) models. The question of validity of using decomposable graphical models using the HIW prior when the true graph is in fact non-decomposable is unanswered till date despite its popularity and development of associated posterior computation techniques over the past 20 years.

1.4 Research Outline

In Chapter 2, we focus on developing a flexible Bayesian framework for simultaneous variable selection and graph learning. We address the variable selection consistency in the proposed Bayesian framework under moderate conditions. At the end, we combine the data sets from "The Cancer Genome Atlas (TCGA) project" and "The Cancer Proteome Atlas (TCPA)" project to demonstrate our proposed method.

In Chapter 3, we study the graph selection consistency theorems for model misspecification when using decomposable graphs only. We address the connection between sample partial correlations and graph selection consistency. Simulation studies are conducted to replicate the convergence rates for model misspecification along with well-specified case.

2. BAYESIAN VARIABLE SELECTION IN MULTIVARIATE NONLINEAR REGRESSION WITH GRAPH STRUCTURES

2.1 Introduction

Most of the existing Gaussian graphical models are used to infer the conditional dependency structure of stochastic variables ignoring any covariate effects on the variables. In this section we consider the situation when multiple sets of variables are assessed simultaneously. An example of such a data structure include various types of genomic, epigenomic, transcriptomic and proteomic data have become available using array and sequencing based techniques. The variables in these biological systems contain enormous numbers of genetic markers have been collected at various levels such as mRNA, DNA, microRNA and protein expressions from a common set of samples. The interrelations within and among these markers provide key insights into the disease etiology. One of the crucial questions is to integrate these diverse data-types to obtain more informative and interpretable graphs representing the interdependencies between the variables. For example, in the study used in this chapter we consider protein and mRNA expression levels from the same patient samples have been collected extensively under The Cancer Genome Atlas (TCGA) project. As the protein expression levels are correlated due to presence of complex biological pathways and interactions, hence we are interested to develop conditional dependence model for them. However, in addition to other proteins, it is well-established that transcriptomic-level mRNA expressions modify the downstream proteomic expressions. This integrating the mRNA expressions as covariates or predictors in the model will produce more precise and refined estimates of the protein-level graphical structure.

From a modeling standpoint, to incorporate covariates in this graphical modeling framework, we adopt seemingly unrelated regression (SUR) [32, 33] models where multiple predictors affect multiple responses, just as in the case of a multivariate regression, with the additional complication that the response variables exhibit an unspecified correlation structure. Similar SUR model has been proposed in [34] which allows different responses to have different predictors. On the other hand, we assume that all responses have the same predictors. The model we propose has both theoretical and computational advantages over the SUR model in [34]. In stead of using approximation for the marginal likelihood as in [34] , regression parameters and error covariance matrices can be marginalized explicitly in our modeling framework. Furthermore, we develop an efficient MCMC sampling alogorithm based on the exact conditional posterior distributions. Indeed, this closed form marginal likelihood enables us to explore the theoretical results for variable selection. In addition our model is suitable for the problem of interest, identifying the influential gene expression based drivers for the entire network. We propose a joint sparse modeling approach for the responses (e.g. protein expressions) as well as covariates (e.g. mRNA expressions). This joint model simultaneously performs a Bayesian selection of significant covariates [35, 36] as well as the significant entries in the adjacency matrix of the correlated responses [17]. In the frequentist setting, similar joint modeling has been recently attempted by [37], [38] and [39] for linear models.

To our best knowledge, the literature on Bayesian estimation of joint covariate-dependent graphical models is sparse, with the exception of [40]. Our proposed method differs from [40] in many aspects. In their paper, they used a linear model (for the covariate effects) with independent priors on the regression coefficients. On the other hand, we develop a nonlinear spline based model and propose a multivariate version of the well-known Zellner's g-prior [41] for the regression parameters. This is a natural extension of the original g-prior from multiple linear regression models to have a matrix normal structure. In fact, it is also a conjugate prior in this multivariate setup, hence drastically reduces the computational burden. Moreover, we investigate the Bayesian variable selection consistency of this multivariate regression model with graphical structures. Indeed, there are a few papers which have considered Bayesian variable selection consistency in a multiple linear regression framework with univariate responses [42], [43]. However, to our best knowledge, none of the existing papers investigated these theoretical results for multivariate regression with or without graphical structures.

To demonstrate our joint model for both variable and graph selection, we conduct a simulation study on a synthetic data set. The spline based regression captures the nonlinear structure precisely and the graph estimation has identified the underlying true graph. We also illustrate the necessity of incorporating the correct covariate structure by comparing the graph selection with respect to the null (no covariate) model and the linear regression model. As a result, we discover that the graph estimator is highly dependent on specifying the correct covariate structure. At the end, we analyze a data set from The Cancer Proteome Atlas (TCPA) project to identify the mRNA driver based protein networks.

The rest of this chapter is organized as follows. We introduce our model and prior specification in Section 2.2. In Section 2.3, we present the stochastic search algorithm for the joint variable and graph selection. The variable selection consistency results have been presented in Section 2.4. Some simulation experiments are conducted in Section 2.5. We apply our method on the TCPA data in Section 2.6. Section 2.7 includes discussions. The detailed proofs of all consistency results can be found in Appendix A and Appendix B.

2.2 The Model

2.2.1 Hyper-inverse Wishart Distribution

The inverse Wishart distribution which is a class of conjugate priors for positive definite matrices does not have the conditional independencies using to impose graphs. By imposing the conditional independencies on the inverse Wishart distribution, [14] derived two classes of distributions – “local” and “global”. But only the “local” one induces sparse graphs. This is known as hyper-inverse Wishart distribution, proposed by [3]. It is the general set of conjugate priors for positive definite matrices which satisfies the hyper Markov law. Its definition is based on the junction tree representation. Let $J_G = (C_1, S_2, C_2, S_3, \dots, C_{k-1}, S_k, C_k)$ be the junction tree representation of a decomposable graph G, then the hyper-inverse Wishart prior for the corresponding covariance matrix Σ_G can be written as a ratio of products of cliques over products of separators [17],

$$p(\Sigma|G) = \frac{\prod_{C\in\mathscr{C}} p(\Sigma_C|b, D_C)}{\prod_{S\in\mathscr{S}} p(\Sigma_S|b, D_S)}, \tag{2.1}$$

where $\mathscr{C}$ and $\mathscr{S}$ are the sets of all cliques and all separators respectively. For each clique C (and separator S), $\Sigma_C \sim \mathrm{IW}(b, D_C)$ with density

$$p(\Sigma_C|b, D_C) = \frac{|D_C|^{\frac{b+|C|-1}{2}}}{2^{\frac{(b+|C|-1)|C|}{2}}\Gamma_{|C|}\big(\frac{b+|C|-1}{2}\big)}|\Sigma_C|^{-\frac{b+2|C|}{2}}exp\Big\{-\frac{1}{2}tr\big(\Sigma_C^{-1}D_C\big)\Big\}, \qquad (2.2)$$

where $\Gamma_p(\cdot)$ is the multivariate gamma function.

For a given graph G, let $\mathrm{y}_i \sim N_q(0, \Sigma_G), i = 1, \ldots, n$ and $\mathrm{Y} = (\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_n)^T$. If $\Sigma_G|G \sim \mathrm{HIW}_G(b, D)$, for some positive integer $b > 3$ and positive definite matrix D, we have $\Sigma_G|\mathrm{Y}, G \sim \mathrm{HIW}_G(b + n, D + \mathrm{Y}^T\mathrm{Y})$. Therefore, the posterior of Σ_G is still a HIW distribution. In the next section, we will incorporate covariate information in this model in a nonlinear regression framework.

2.2.2 Covariate Adjusted GGMs

We consider the following covariate adjusted Gaussian distribution $\mathrm{y} \sim N_q(f(\mathrm{x}), \Sigma_G)$, where $\mathrm{y} = (y_1, y_2, \ldots, y_q)^T$, $\mathrm{x} = (x_1, x_2, \ldots, x_p)^T$ and the function $f : \mathbb{R}^p \to \mathbb{R}^q$ performs a smooth, nonlinear mapping from the p-dimensional predictor space to the q-dimensional response space. Σ_G is the covariance structure of y corresponding to the graph G. Linear model developed by [40] is a particular case of this where $f(\mathrm{x}) = \mathrm{x}^T\boldsymbol{\beta}$. In the nonlinear setup, we choose to use spline to approximate the nonlinear function $f(\cdot)$. Without loss of generality, we assume all components of x share the same range, which means we can use the same knot points for all variables which simplifies the notations. And we also assume all covariates are centered so that the intercept terms are zero here. Given k knot points, $\mathrm{w} = (w_1, w_2, \ldots, w_k)^T$, the spline basis for x_i is $\{(x_i - w_1)_+, (x_i - w_2)_+, \ldots, (x_i - w_q)_+\}$. So $f(\mathrm{x})$ can be approximated by the linear form uB, where $\mathrm{u}_{1\times p(k+1)} = \{\mathrm{x}^T, (\mathrm{x} - w_1)_+^T, (\mathrm{x} - w_2)_+^T, \ldots, (\mathrm{x} - w_k)_+^T\}$ and $(\mathrm{x} - w_i)_+^T = \{(x_1 - w_i)_+, (x_2 -$

$w_i)_+, \ldots, (x_p - w_i)_+\}$ and B is the coefficient matrix, which has the structure below,

$$\mathrm{B}_{p(k+1)\times q} \quad = \quad \begin{bmatrix} \beta_{110} & \beta_{210} & \cdots & \beta_{q10} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{1p0} & \beta_{2p0} & \cdots & \beta_{qp0} \\ \beta_{111} & \beta_{211} & \cdots & \beta_{q11} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{1p1} & \beta_{2p1} & \cdots & \beta_{qp1} \\ \beta_{1pk} & \beta_{2pk} & \cdots & \beta_{qpk} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{1pk} & \beta_{2pk} & \cdots & \beta_{qpk} \end{bmatrix}.$$

We assume the knot points w's to be known and prespecified. That way, we have spline-adjusted model $\mathrm{y} \sim N_q(\mathrm{uB}, \Sigma_G)$ which has a linear model structure. Therefore, any variable selection method for linear regression can be used for the mean structure.

2.2.3 The Bayesian Hierarchical Model

Assuming we have a set of n independent samples $\mathrm{Y} = (\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_n)^T$, where $\mathrm{y}_i \sim N_q(f(\mathrm{x}_i), \Sigma_G)$ and let $f(\mathrm{X}) = (f(\mathrm{x}_1), f(\mathrm{x}_2), \ldots, f(\mathrm{x}_n))^T$. We have

$$\mathrm{Y} \sim \mathrm{MN}_{n\times q}(f(\mathrm{X}), I_n, \Sigma_G), \tag{2.3}$$

where $\mathrm{MN}_{n\times q}(f(\mathrm{X}), I_n, \Sigma_G)$ is the matrix normal distribution with mean $f(\mathrm{X})$, and I_n as the covariance matrix between n rows and Σ_G as the covariance matrix between q columns. We approximate $f(\cdot)$ by $f(\mathrm{X}) = \mathrm{UB}$, where U is the spline basis matrix which has the structure below. And it is equivalent to write out the model as multivariate linear regression, $\mathrm{Y} = \mathrm{UB} + \mathrm{E}$, where $\mathrm{E} \sim \mathrm{MN}_{n\times q}(0, I_n, \Sigma_G)$.

$$\mathrm{U}_{n\times p(k+1)} = \begin{bmatrix} \mathrm{x}_1 & \mathrm{x}_2 & \dots & \mathrm{x}_n \\ (\mathrm{x}_1 - w_1)_+ & (\mathrm{x}_2 - w_1)_+ & \dots & (\mathrm{x}_n - w_k)_+ \\ \vdots & \vdots & \vdots & \ddots \\ (\mathrm{x}_1 - w_k)_+ & (\mathrm{x}_2 - w_k)_+ & \dots & (\mathrm{x}_n - w_k)_+ \end{bmatrix}^T$$

To introduce the notion of redundant variables for the variable selection in the mean structure, we define a binary vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^T$, where $\gamma_i = 0$ if and only if $\beta_{jis} = 0$, for all $j = 1, \dots, q, \;\; s = 0, 1, \dots, k$. By following this rule, the spline basis functions are related to each variable when performing the model selection. It means selecting one variable is equivalent to select all its related basis functions. Similarly, to introduce the notion of sparsity in the precision matrix, we define a binary variable G_l, where $l = 1, \dots, \frac{q(q-1)}{2}$, the lth off diagonal element in the adjacency matrix corresponding to the graph G. Diagonal elements of the adjacency matrix are restricted to one. The number of edges in the graph G is denoted as $|E| = \sum_l G_l$. The Bayesian hierarchical model is given by

$$\begin{aligned}
(\mathrm{Y} - \mathrm{U}_{\boldsymbol{\gamma}}\mathrm{B}_{\boldsymbol{\gamma},G})|\mathrm{B}_{\boldsymbol{\gamma},G}, \Sigma_G &\sim \mathrm{MN}_{n\times q}(0, I_n, \Sigma_G), && (2.4)\\
\mathrm{B}_{\boldsymbol{\gamma},G}|\boldsymbol{\gamma}, \Sigma_G &\sim \mathrm{MN}_{p_{\boldsymbol{\gamma}}(k+1)\times q}\big(0, g(\mathrm{U}_{\boldsymbol{\gamma}}^T\mathrm{U}_{\boldsymbol{\gamma}})^{-1}_{p_{\boldsymbol{\gamma}}(k+1)}, \Sigma_G\big), && (2.5)\\
\Sigma_G|G &\sim \mathrm{HIW}_G(b, dI_q), && (2.6)\\
\gamma_i &\overset{i.i.d.}{\sim} \text{Bernoulli}(\alpha_{\boldsymbol{\gamma}}) \text{ for } i = 1, \dots, p_{\boldsymbol{\gamma}}, && (2.7)\\
G_l &\overset{i.i.d.}{\sim} \text{Bernoulli}(\alpha_G) \text{ for } l = 1, \dots, \frac{q(q-1)}{2}, && (2.8)\\
\alpha_{\boldsymbol{\gamma}} &\sim \mathrm{U}(0, 1), && (2.9)\\
\alpha_G &= 2/(q-1), && (2.10)
\end{aligned}$$

where $\mathrm{U}_{\boldsymbol{\gamma}}$ is the spline basis matrix with regressors corresponding to $\boldsymbol{\gamma}$ and $b > 3$, g, d are fixed positive hyper parameters. $\alpha_{\boldsymbol{\gamma}}$ is used to control the sparsity of variable selection and α_G is responsible for the complexity of graph selection. Also, denote $p_\gamma = \sum_i \gamma_i$.

Equation (2.5) is the extended version of Zellner's g-prior [41] for multivariate regression. g-prior in this matrix normal form requires one more parameter than the usual multivariate normal form to allow the covariance structure between columns. Here, we use Σ_G as that parameter. There are a couple of reasons for this choice. First, it drastically decreases the complexity of marginalization. By using the same structure as the graph, it gives us the ability to integrate out the coefficient matrix $\mathrm{B}_{\gamma,G}$. That way, we derive the marginal of Y explicitly. Moreover, it allows the variable selection and the graph selection to borrow strength from each other. Next, we derive the marginal density of data Y given only $\boldsymbol{\gamma}$ and graph G in this modeling framework.

By using equation (2.4) and (2.5), we have

$$\mathrm{Y}|\boldsymbol{\gamma}, \Sigma_G \sim \mathrm{MN}_{n\times q}(0, I_n + gP_{\boldsymbol{\gamma}}, \Sigma_G),$$

where $P_\gamma = \mathrm{U}_\gamma(\mathrm{U}_\gamma^T\mathrm{U}_\gamma)^{-1}\mathrm{U}_\gamma^T$. In order to calculate the marginal of Y, we need to vectorize Y as follows,

$$vec(\mathrm{Y}^T)|\boldsymbol{\gamma}, \Sigma_G \sim N_{nq}(0, (I_n + gP_{\boldsymbol{\gamma}}) \otimes \Sigma_G),$$

where $\otimes$ is the Kronecker product. Next, using the equation (2.6), we integrate out the Σ_G to derive the marginal distribution of Y. The detailed calculation is in Appendix A. Let $\mathscr{C}$ and $\mathscr{S}$ be the sets of all cliques and all separators for the given graph G then

$$f(\mathrm{Y}|\boldsymbol{\gamma}, G) = M_{n,G} \times (g+1)^{-\frac{p_{\boldsymbol{\gamma}}(k+1)q}{2}} \frac{\prod_{C\in\mathscr{C}} |dI_C + S_C(\boldsymbol{\gamma})|^{-\frac{b+n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}} |dI_S + S_S(\boldsymbol{\gamma})|^{-\frac{b+n+|S|-1}{2}}}, \tag{2.11}$$

where $S(\boldsymbol{\gamma}) = \mathrm{Y}^T(I_n - \frac{g}{g+1}P_{\boldsymbol{\gamma}})\mathrm{Y}$, $S_C(\boldsymbol{\gamma})$ and $S_S(\boldsymbol{\gamma})$ denote the quadratic forms restricted to the clique $C \in \mathscr{C}$ and the separator $S \in \mathscr{S}$.

The normalizing constant $M_{n,G}$ has the following factorization which depends only on n and G, but it is the same for all $\boldsymbol{\gamma}$ under the same graph G. The advantage of this is when updating $\boldsymbol{\gamma}$

in the stochastic search, this term cancels out reducing the computational complexity.

$$M_{n,G} = (2\pi)^{-\frac{nq}{2}} \frac{\prod_{C \in \mathscr{C}} \frac{|dI_C|^{\frac{b+|C|-1}{2}}}{2^{-\frac{n|C|}{2}} \Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right) \Gamma_{|C|}^{-1}\left(\frac{b+n+|C|-1}{2}\right)}}{\prod_{S \in \mathscr{S}} \frac{|dI_S|^{\frac{b+|S|-1}{2}}}{2^{-\frac{n|S|}{2}} \Gamma_{|S|}\left(\frac{b+|S|-1}{2}\right) \Gamma_{|S|}^{-1}\left(\frac{b+n+|S|-1}{2}\right)}}.$$

2.2.4 Prior Specification for $\boldsymbol{\gamma}$ and G

We use beta-binomial priors [35] for both variable and graph selection. We control the sparsity by fixing α_G. [18] suggested to use $\frac{2}{|V|-1}$ as the hyper parameter for the Bernoulli distribution. For an undirected graph, it has peak around $|V|$ edges and it will be lower when applying to decomposable graphs. Additionally, we have other ways to control the number of edges which will be stated in the next section.

2.3 The Stochastic Search Algorithm

2.3.1 Searching for $\boldsymbol{\gamma}$

From equation (2.7), we obtain the prior $p(\boldsymbol{\gamma}|\alpha_{\boldsymbol{\gamma}}) = \prod_{i=1}^{p} p(\gamma_i|\alpha_{\boldsymbol{\gamma}}) = \alpha_{\boldsymbol{\gamma}}^{p_{\boldsymbol{\gamma}}}(1-\alpha_{\boldsymbol{\gamma}})^{p-p_{\boldsymbol{\gamma}}}$. Next, using equation (2.9), we integrate out $\alpha_{\boldsymbol{\gamma}}$, so that the marginal prior for $\boldsymbol{\gamma}$ is $p(\boldsymbol{\gamma}) \propto p_{\boldsymbol{\gamma}}!(p-p_{\boldsymbol{\gamma}})!$. The searching for $\boldsymbol{\gamma}$ proceeds as follows:

- Given $\boldsymbol{\gamma}$, propose $\boldsymbol{\gamma}^*$ by the following procedure. With equal probabilities, randomly choose one entry in $\boldsymbol{\gamma}$, say γ_{s^*}. If $\gamma_{s^*} = 0$, then with probability δ change it to 1 and with probability $1-\delta$ remain the same; if $\gamma_{s^*} = 1$, then with probability $1-\delta$ change it to 0 and with probability δ remain the same. Under this setting, δ is the probability of adding one variable when $\gamma_{s^*} = 0$ and $1-\delta$ is the probability of deleting one variable when $\gamma_{s^*} = 1$. If $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}$, then $\frac{q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})}$=1. If one variable has been added to the model, $\frac{q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})} = \frac{1-\delta}{\delta}$; if one variable has been deleted from the model, $\frac{q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})} = \frac{\delta}{1-\delta}$.

- Calculate the marginal densities under both models $p(\mathrm{Y}|\boldsymbol{\gamma}, G)$ and $p(\mathrm{Y}|\boldsymbol{\gamma}^*, G)$.

- Accept $\boldsymbol{\gamma}^*$ with probability

$$r(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = \min\left\{1, \frac{p(\mathrm{Y}|\boldsymbol{\gamma}^*, G)p(\boldsymbol{\gamma}^*)q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{p(\mathrm{Y}|\boldsymbol{\gamma}, G)p(\boldsymbol{\gamma})q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})}\right\}.$$

Notice, under the same graph the normalizing constant $M_{n,G}$ cancels out. Another thing is by using the parameter δ, we can further control the sparsity of variable selection.

2.3.2 Searching for G

Similar to the calculation for $\boldsymbol{\gamma}$, the prior over the graph space is $p(G|\alpha_G) = \prod_{l=1}^{q(q-1)/2} p(G_l|\alpha_G) = \alpha_G^{|E|}(1-\alpha_G)^{q(q-1)/2-|E|}$, where $|E|$ is the total number of edges in the graph G and $\alpha_G = 2/(|V|-1)$. The searching for G works as follows:

- Given the current decomposable graph G, propose a new decomposable graph G^* by the following procedure. With equal probabilities, randomly select an off-diagonal entry from the adjacency matrix of graph G, say G_{s^*}. If $G_{s^*} = 0$, then with probability η change it to 1 and with probability $1-\eta$ remain the same; if $G_{s^*} = 1$, with probability $1-\eta$ change it to 0 and with probability η remain the same. So the probability of adding an edge is η when $G_{s^*} = 0$ and $1-\eta$ is the probability of deleting an edge when $G_{s^*} = 1$. We discard all proposed graphs which are non-decomposable. In those cases, the chain remains in the same graph for that iteration. If an edge has been added to the graph, $\frac{p(G|G^*)}{p(G^*|G)} = \frac{1-\eta}{\eta}$; if an edge has been removed from the graph, $\frac{p(G|G^*)}{p(G^*|G)} = \frac{\eta}{1-\eta}$.

- Calculate the marginal densities under both graphs $p(\mathrm{Y}|\boldsymbol{\gamma}, G)$ and $p(\mathrm{Y}|\boldsymbol{\gamma}, G^*)$.

- Accept G^* with probability

$$r(G, G^*) = \min\left\{1, \frac{p(\mathrm{Y}|\boldsymbol{\gamma}, G^*)p(G^*)q(G|G^*)}{p(\mathrm{Y}|\boldsymbol{\gamma}, G)p(G)q(G^*|G)}\right\}.$$

This procedure is called add-delete Metropolis-Hastings sampler [18]. Another tool for sparsity is η. By choosing its value to be less than 0.5, it can reinforce sparsity on the graph.

2.3.3 Conditional Distributions of $\mathrm{B}_{\gamma,G}$ and Σ_G

We integrate out $\mathrm{B}_{\gamma,G}$ and Σ_G to make the stochastic search more efficient. But the conditional distributions of them both have simple closed forms. In [40], the conditional distribution of Σ_G depends on the coefficient matrix $\mathrm{B}_{\gamma,G}$, but by using Zellner's g-prior it only requires $\boldsymbol{\gamma}$ and G,

$$\Sigma_G|\mathrm{Y}, \boldsymbol{\gamma}, G \sim \mathrm{HIW}_G(b+n, dI_q + S(\boldsymbol{\gamma})).$$

At each iteration, given $\boldsymbol{\gamma}$ and Σ_G, using the following conditional distribution we can simulate $\mathrm{B}_{\gamma,G}$,

$$\mathrm{B}_{\gamma,G}|\mathrm{Y}, \boldsymbol{\gamma}, \Sigma_G \sim \mathrm{MN}_{p_\gamma(k+1)\times q}\left(\frac{g}{g+1}\left(\mathrm{U}_\gamma^T\mathrm{U}_\gamma\right)^{-1}\mathrm{U}_\gamma^T\mathrm{Y}, \frac{g}{g+1}\left(\mathrm{U}_\gamma^T\mathrm{U}_\gamma\right)^{-1}, \Sigma_G\right).$$

2.3.4 Choices of Hyperparameters

For choosing hyperparameters, we need to specify g, b, d. [43] summarized some choices for g in the g-prior, like $g = n$ [44], $g = p^2$ [45], $g = \max(n, p^2)$ [42] and other empirical Bayes methods to choose g. Based on simulations the choice of g is not very critical in our approach. As long as g satisfies the basic condition $g = O(n)$, there is no significant effect on the results. This condition is to keep the variances of the prior of coefficients not to be too small as n goes to infinity. But when the dimension of the predictor space p is large, one can consider to use $g = \max(n, p^2)$.

The hyperparameter b and d are the two constants which control the hyper-inverse Wishart distribution. The common choice for the degree of freedom b is 3 which provides a finite moment for the HIW prior [18, 17]. Based on our experiments d has a big impact on the graph selection results. Large d results in more sparse graphs. On the other hand, large d also contributes to large variances for coefficients. After standardizing the variances of responses to be 1, [18] suggested to use $1/(b+1)$ as a default choice of d, since the marginal prior mode for each variance term is $d(b+1)$. In our approach we are basically using the residuals after variable selection to fit the graphical model, hence it is impossible to know the variances. But we find $d = 1$ works well in our simulations.

The common choice for δ and η in the stochastic search is 0.5. Unless strong parsimony is required, we suggest to use this value. On the other hand, α_G can be set to $1/(|V|-1)$ or $0.5/(|V|-1)$ to achieve more sparsity on graph selection for noisy data.

2.4 Variable Selection Consistency

In this section, we first show the Bayes Factor consistency of the variable selection method for a given graphical structure. To our best knowledge, there are no results on Bayesian variable selection consistency for this case. We first define the pairwise Bayes factor consistency for a given graph, subsequently under moderate conditions, we prove the pairwise Bayes factor consistency. For some related development in multiple linear regression model with univariate response, see [43] and [42]. For simplicity, from now on we refer the multivariate regression model as the regression model or just the model. Without further specification, the model we refer implies the regression model, not the graphical model.

Let binary vector $\boldsymbol{t} = (t_1, \ldots, t_p)^T$ denote the regression model with respect to the true set of covariates of size $p_{\boldsymbol{t}} = \sum_{i=1}^{p} t_i$ and binary vector $\boldsymbol{a} = (a_1, \ldots, a_p)^T$ denote an alternative regression model of size $p_{\boldsymbol{a}} = \sum_{i=1}^{p} a_i$. We use $\boldsymbol{\gamma}$ to represent any subset of the regression model space for being consistent with the notation in the early section. Next, we introduce the definition of pairwise Bayes factor consistency with graph structures.

Definition 2.4.1. (pairwise Bayes factor consistency under a given graph) *Let* $\mathrm{BF}(\boldsymbol{a};\boldsymbol{t}|G)$ *be the Bayes factor in favor of an alternative model* $\boldsymbol{a}$ *for a given graph* G*, such that* $\mathrm{BF}(\boldsymbol{a};\boldsymbol{t}|G) = \frac{P(\mathrm{Y}|\boldsymbol{a},G)}{P(\mathrm{Y}|\boldsymbol{t},G)}$*. If* $\mathrm{p}\lim_{n\to\infty}\mathrm{BF}(\boldsymbol{a};\boldsymbol{t}|G) = 0$*, for any* $\boldsymbol{a} \neq \mathbf{t}$*, then we have pairwise Bayes factor consistency with respect to the true regression model* $\mathbf{t}$ *and the graph* G*.*

Here, “$\mathrm{p}\lim_{n\to\infty}$" denotes convergence in probability and the probability measure is the sampling distribution under the true data generating model [43]. Notice that the alternative model and the true model used in the Bayes factor calculation have the same graph G which may not be the true underlying graph. To clarify, the Bayes factor in the definition above is for a given graph G, where as the actual Bayes factor for the joint model is defined as $\mathrm{BF}(\boldsymbol{a};\boldsymbol{t}) = \frac{P(\mathrm{Y}|\boldsymbol{a})}{P(\mathrm{Y}|\boldsymbol{t})} = \frac{\int P(\mathrm{Y}|\boldsymbol{a},G)\pi(G)dG}{\int P(\mathrm{Y}|\boldsymbol{t},G)\pi(G)dG}$, where $\pi(G)$ is the prior on the graph space. In this paper, the graph G is restricted to the set of decomposable graphs and the number of nodes q in the graph is finite. Before the main result, some regularization conditions need to be introduced.

2.4.1 Conditions

Condition 2.4.1. *The set of graphs we consider is restricted to decomposable graphs with the number of nodes q is finite. The number of knots k for the spline basis is also finite.*

Condition 2.4.2. *Let $\lambda_{min} \le \cdots \le \lambda_i \le \cdots \le \lambda_{max}$ be the eigenvalues of $\left(\mathrm{U}_{\gamma}^T\mathrm{U}_{\gamma}\right)_{p_{\gamma}(k+1)\times p_{\gamma}(k+1)}$. Assume $0 < c_{\mathrm{U}} < \frac{\lambda_{min}}{n} \le \frac{\lambda_{max}}{n} < d_{\mathrm{U}} < \infty$, where c_{U} and d_{U} are two positive finite constants.*

Condition 2.4.3. *Let $\mathrm{E}_y = \mathrm{U}_{\boldsymbol{t}}\mathrm{B}_{\boldsymbol{t},G}$. Assume $\inf_{\boldsymbol{a}\neq\boldsymbol{t}} tr\{\mathrm{E}_y^T(I_n - P_{\boldsymbol{a}})\mathrm{E}_y\} > C_0 n$, where $P_{\boldsymbol{a}}$ is the projection matrix of an alternative model $\boldsymbol{a}$ and C_0 is some fixed constant.*

Condition 2.4.4. *For the g-prior $\mathrm{B}_{\gamma,G}|\boldsymbol{\gamma},\Sigma_G \sim \mathrm{MN}\big(0, g\left(\mathrm{U}_{\gamma}^T\mathrm{U}_{\gamma}\right)^{-1}, \Sigma_G\big)$ as in equation (2.6), assume $g = O(n)$.*

Condition 2.4.5. *Let $\hat{\Sigma}_{\gamma,G}^{-1}$ be the MLE of Σ_G^{-1} under any regression model $\boldsymbol{\gamma}$ and any decomposable graph G. Assume $\hat{\Sigma}_{\gamma,G}^{-1}$ converges to some positive definite matrix ${\Sigma_{\gamma,G}^0}^{-1}$ which has all eigenvalues bounded away from zero and infinity. Later, we drop the subscript $\boldsymbol{\gamma}$ and only use $\hat{\Sigma}_G^{-1}$ and ${\Sigma_G^0}^{-1}$.*

Condition 2.4.6. *The number of total covariates satisfies $\lim_{n\to\infty} \frac{p}{n} = 0$, i.e. $p = o(n)$.*

Condition 2.4.2 is needed to avoid singularity when the dimension of the model space p increases to infinity as n goes to infinity. Condition 2.4.3 indicates that no true covariate can be fully explained by the rest of the covariates, which implies that regressing any true covariate on all others, the coefficient of determination R^2 is less than 1. Condition 2.4.4 makes sure we assign a non-degenerated prior on the coefficient matrix. Condition 2.4.5 imposes restriction on the limit of $\hat{\Sigma}_G^{-1}$ or equivalently on the corresponding quadratic forms. It is needed for the given clique and separator decomposition of the hyper-inverse Wishart prior. For inverse Wishart prior, this condition can be relaxed if the corresponding graph is complete. We assume that the MLE converges to a positive definite matrix ${\Sigma_G^0}^{-1}$. For the true graph this statement holds trivially. The explicit calculation of the MLE can be done by calculating the MLEs for each clique and separator, then combining them to form $\hat{\Sigma}_G^{-1}$. Given any model $\boldsymbol{\gamma}$, for $C \in \mathscr{C}$, define $\hat{\Sigma}_C^{-1} = \left\{\frac{1}{n}\mathrm{Y}_C^T(I_n - \frac{g}{g+1}P_{\boldsymbol{\gamma}})\mathrm{Y}_C\right\}^{-1}$ and $\hat{\Sigma}_S^{-1}$ is defined similarly. The MLE is given by $\hat{\Sigma}_G^{-1} = \sum_{C\in\mathscr{C}} \hat{\Sigma}_C^{-1}\big|_0 - \sum_{\mathscr{S}} \hat{\Sigma}_S^{-1}\big|_0$ where the suffix '0' implies that the elements corresponding to the vertices which are not in that subgraph are filled with zeros to create a $q \times q$ matrix [1].

2.4.2 Consistency Results

Lemma 2.4.1. *Let $(\boldsymbol{t}, G)$ be the model with true covariates and any finite dimensional graph G. Assume $(\boldsymbol{a}, G)$ is any alternative model $(\boldsymbol{a} \neq \boldsymbol{t})$ with the same graph G. Then, for the model given by (2.4)-(2.6), under Condition 2.4.1-2.4.6,* $\mathrm{p}\lim_{n\to\infty} \mathrm{BF}(\boldsymbol{a}; \boldsymbol{t}|G) = 0$ *for* ***any*** *graph G and any model $\boldsymbol{a} \neq \boldsymbol{t}$.*

Proof. See Appendix B for details. □

That is to say, in the Lemma 2.4.1, we conclude that the pairwise Bayes factor for variable selection is consistent for any given graph, which is a quite strong result considering the magnitude of the graph space (here it is restricted to decomposable graph space). Next we show that with finite dimension graph, the result we have from Lemma 2.4.1 is equivalent to the traditional Bayes factor for regression models.

Theorem 2.4.1. (pairwise Bayes factor consistency) *For model given by (2.4)-(2.10), under Condition 2.4.1-2.4.6,* $\mathrm{p}\lim_{n\to\infty} \mathrm{BF}(\boldsymbol{a}; \boldsymbol{t}) = 0$ *for any model $\boldsymbol{a} \neq \boldsymbol{t}$.*

Proof. Since the number of nodes q in the graph is finite, then let $N_G(q) < \infty$ denote the number of all possible graphs. Therefore, for any alternative model $\boldsymbol{a} \neq \boldsymbol{t}$, we have

$$\begin{aligned}\mathrm{BF}(\boldsymbol{a}; \boldsymbol{t}) &= \frac{P(\mathrm{Y}|\boldsymbol{a})}{P(\mathrm{Y}|\boldsymbol{t})} = \frac{\int P(\mathrm{Y}|\boldsymbol{a}, G)\pi(G)dG}{\int P(\mathrm{Y}|\boldsymbol{t}, G)\pi(G)dG} \\ &= \frac{\sum_{i=1}^{N_G(q)} P(\mathrm{Y}|\boldsymbol{a}, G_i)\pi(G_i)}{\sum_{i=1}^{N_G(q)} P(\mathrm{Y}|\boldsymbol{t}, G_i)\pi(G_i)}\end{aligned}$$

$$\leq \frac{N_G(q)P(\mathrm{Y}|\boldsymbol{a}, G_M)\pi(G_M)}{P(\mathrm{Y}|\boldsymbol{t}, G_M)\pi(G_M)} = N_G(q)\frac{P(\mathrm{Y}|\boldsymbol{a}, G_M)}{P(\mathrm{Y}|\boldsymbol{t}, G_M)} \to 0,$$

where $P(\mathrm{Y}|\boldsymbol{a}, G_M) = \max\{P(\mathrm{Y}|\boldsymbol{a}, G_i), i = 1, \ldots, N_G(q)\}$ and $\pi(G_M) = \max\{\pi(G_i), i = 1, \ldots, N_G(q)\}$. Then following the result in Lemma 2.4.1, we have a pairwise Bayes factor consistency for regression models. □

Notice that we do not need the number of covariates p to be finite here. As long as the conditions are satisfied, the result of Lemma 2.4.1 and Theorem 2.4.1 hold accordingly. Next, we discuss the variable selection consistency. For finite dimensional covariate space, the variable selection consistency is an immediate result.

Corollary 2.4.1. *For model given by (2.4)-(2.10), under Condition 2.4.1-2.4.5, if the number of covariates p is finite, then* $\mathrm{p}\lim_{n\to\infty} P(\boldsymbol{t}|\mathrm{Y}) = 1$.

Proof.

$$P(\boldsymbol{t}|\mathrm{Y}) = \frac{\pi(\boldsymbol{t})P(\mathrm{Y}|\boldsymbol{t})}{\sum_{\boldsymbol{a}} \pi(\boldsymbol{a})P(\mathrm{Y}|\boldsymbol{a})} = \left(\sum_{\boldsymbol{a}} \frac{\pi(\boldsymbol{a})p(\mathrm{Y}|\boldsymbol{a})}{\pi(\boldsymbol{t})p(\mathrm{Y}|\boldsymbol{t})}\right)^{-1} = \left(1 + \sum_{\boldsymbol{a}\neq\boldsymbol{t}} \frac{\pi(\boldsymbol{a})}{\pi(\boldsymbol{t})}\mathrm{BF}(\boldsymbol{a};\boldsymbol{t})\right)^{-1} \to 1,$$

where $\pi(\cdot)$ is the prior on the regression model. Notice, the last summation has finite number of terms, since the covariate space is finite. Then the rest follows directly form Theorem 2.4.1. □

Therefore, the variable selection is consistent when the model space is finite. Corollary 2.4.1 does not depend on the graph, which means it holds even if we do not identify the true graph.

2.5 Simulation Study

In this section, we present the simulation study for our method considering the hierarchical model from Section 2.2. In order to justify the necessity for a covariate adjusted mean structure, we compare the graph estimation among spline regression, linear regression and no covariates (graph only) assuming an underlying nonlinear covariate structure as the true model.

Considering the hierarchical model in Section 2.2, we choose $p = 30$, $q = 40$, $n = 700$. All predictors x_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, p$ are simulated from uniform distribution $(-1, 1)$. For the nonlinear regression structure, we use a relatively simple and smooth function, similar to the function used in [46],

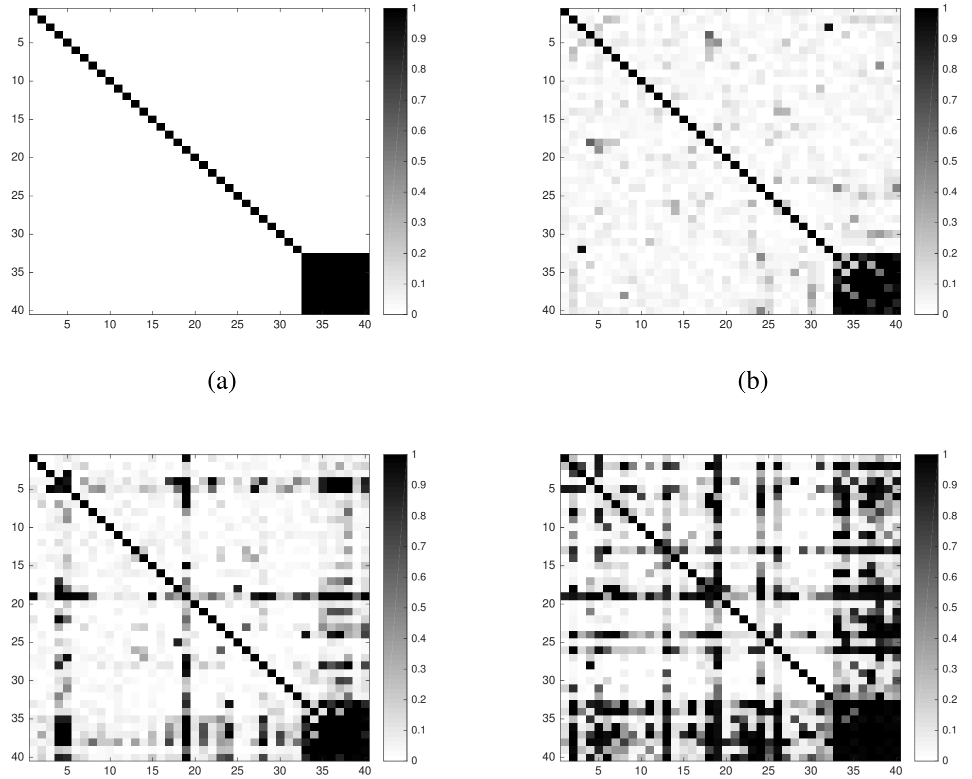
$$f_i(\mathrm{x}_j) = h_{i1}\sin(x_{j5}) + h_{i2}\sin(x_{j11}) + h_{i3}x_{j17} + h_{i4}e^{x_{j24}}, i = 1, 2, \ldots, q, \tag{2.12}$$

where $\mathrm{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ is the jth sample of data $\mathrm{X} = (\mathrm{x}_1^T, \mathrm{x}_2^T, \ldots, \mathrm{x}_q^T)^T$. The true set of predictors is $\{5, 11, 17, 24\}$. The coefficients in (2.12) are simulated from exponential distribution $\exp(1)$. Figure 2.1(a) shows the true adjacency matrix for the true graph G. The true covariance matrix Σ_G is generated from $\mathrm{HIW}_G(3, I_q)$. And the columns of error matrix E are n random draws from multivariate normal distribution $N_q(0, \Sigma_G)$. Thus $\mathrm{Y} = F(\mathrm{X}) + \mathrm{E}$, where $F(\mathrm{X}) = \big(f_j(\mathrm{x}_i)\big)_{ij}$. For hyperparameters, we use $g = n = 700$, $b = 3$, $d = 1$ and $\delta = \eta = 1/2$ in the stochastic search of graphs.

For spline basis functions, we use 10 fixed knot points which divide $(-1, 1)$ evenly into 11 intervals. 100,000 MCMC iterations are performed after 10,000 burn-in steps. We use a similar true graph as in [40] in the simulation. The results are quite fascinating. For variable selection, after burn-in iterations, it quickly converges to the true set of predictors. Furthermore for the variable selection, if we only use the linear regression model to estimate the nonlinear structure, it misses some important predictors. As shown in the simulation study, the exponential term has not been identified without the spline regression. Although the linear case selects most of the correct variables, estimates of mean functions are completely wrong, which misleads the graph estimation completely as we show next.

Here, we use marginal posterior probability for each edge to choose our final estimation of the graph. Marginal probabilities are calculated by averaging all adjacency matrices in the MCMC chain. The cut-off point is 0.5, which means we only select the edge with posterior probability more than half. The cut-off point can also be varied to accomplish different degrees of sparsity for the estimated graph.

Figure 2.1(b) shows when using spline regression to capture the nonlinear mean structure, the



(a)

(b)

(c)

(d)

Figure 2.1: Plots of graph selection. (a) The adjacency matrix of the true graph G. In the adjacency matrix, 1 indicates an edge and 0 indicates no edge. So edges are represented by black bricks. The diagonal entries are 1 by default. (b), (c) and (d) are the marginal posterior probabilities of each edge in the estimated graph on a gray scale under spline regression, linear regression and without covariates, respectively.

major parts of the true graph can be recovered. On the other hand, from Figure 2.1(c), we can see that the linear regression model fails to estimate the residual terms properly, hence the estimated graph is completely wrong. It may still capture a few true edges, however a large number of false edges have been added. Thus, modeling the true mean structure is essential for estimating the graph. That way, specification of an incorrect mean structure (e.g. using linear function to estimate nonlinear function, choosing a wrong set of covariates) always leads to a wrong graph estimation. This can also be illustrated by ignoring the covariates to estimate the graph. Figure 2.1(d) shows in this scenario the estimated graph is again completely wrong. We plot the receiver operating characteristic (ROC) curves for the above three cases in Figure 2.2. As we can see the joint models (i.e. spline and linear regression model) perform better than no covariates. Furthermore, the ROC curve of spline regression model is nearly perfect.

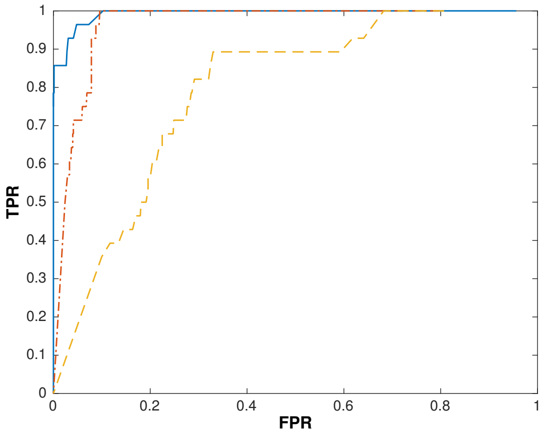


Figure 2.2: Plot of ROC curves for graph selection. The blue, red and yellow lines are spline regression model, linear regression model and no covariates (graph only) model, respectively.

2.6 Protein-mRNA data

In this section, we apply our method to a protein-mRNA data set from The Cancer Proteome Atlas (TCPA) project. The major goals of this analysis are (i) to identify the influential gene expression based drivers, i.e. mRNAs which influence the protein activities and (ii) to estimate the protein network, simultaneously. The central dogma to our model is the well-known fact that mRNA which is the messenger of DNA from transcription plays a critical role in proteomics by a process called translation. Consequently, the protein expressions play a crucial role for the development of tumor cells. Therefore, to identify which mRNAs dominate this process is the key component in this oncology study. This also motivates us to regress the protein level data on the mRNA based data. Multivariate regression is a powerful tool for combining information across all regressions. One can use an univariate regression model on each protein. However, there are multiple advantages of our model to apply in this scenario. First, it combines the information across all responses, i.e. protein expressions. As we know, proteins tend to work together as a network. Performing single regression separately on each of them will lose information which lies in this protein network. Second, by jointly modeling all proteins, we obtain a graph of all proteins after correcting the mRNA effects from the mean structure. In our analysis, we choose the breast cancer data which has the largest sample size of 844 among other tumors. Based on different DNA functions, proteins are categorized into 12 pathways with their corresponding mRNAs. For more details about these pathways, see [47] and [48]. We apply our model for each pathway since proteins from the same pathway exhibit similar behaviors. We use spline based regression to further investigate the different nonlinear relationships among proteins and mRNAs. The results of the covariate selection and the graph estimation are summarized below.

For the standardized design matrix, we use ten evenly distributed knot points for spline basis to capture the nonlinearities between proteins and mRNAs. Since the observations of mRNAs are not uniformly distributed across their ranges, we use the penalized spline regression proposed by [49] to solve the rank deficiency problem in the g-prior. The selection results along with the number of proteins and mRNAs used in each pathway are summarized in Table 2.2 below. Four of those pathways don't have any influential mRNA which controls the proteins. The rest seven pathways all have one or more related mRNAs according to the results. For example, the pathway of Apoptosis is about programmed cell death. The model selects only the mRNA corresponding to gene BCL2. BCL2 is an anti-cell death gene [50]. Proteins in BCL2 family play an important role in control of apoptosis. Studies have found that they constitute a life or death decision point for cells in a common pathway involved in various forms of Apoptosis [50]. In this sense, our model identifies the correct mRNA that dominates this Apoptosis process. BCL2 also contributes to the pathway about hormone receptor and signaling. CCNE1 has been selected in the pathway of cell cycle. It has been found that there is an association between CCNE1 amplification and breast cancer treatment [51]. CCNE1 is also related to the endometrioid endometrial carcinomas [52, 53]. CDH1 has been selected in the pathway of core reactive and EMT. Mutations in CDH1 have been observed to be associated with increased susceptibility to develop lobular breast cancer

[54, 55]. From the selection results, INPP4B is related to PI3K/AKT and hormone receptor and signaling pathways. Interestingly, there are emerging evidences that INPP4B identified as a tumor suppressor regulates PI3K/AKT signaling in breast cancer cell lines [56]. For more information about the relationship between INPP4B and PI3K/ATK pathway, see [57], [58] and [59]. Both ERBB2 and GATA3 genes have strong influence in breast cancer, see [60], [61], [62], [63] and [64]. The list of gene names related to each of 12 pathways is in Table 2.1, see [48] for more details. For the plots of estimated nonlinear functions of each node in all seven pathways, see Appendix C.

Table 2.1: Gene names corresponding to each of 12 pathways

#	Pathway	Genes
1	Apoptosis	BAK1, BAX, BID, BCL2L11, CASP7, BAD, BCL2, BCL2L1, BIRC2
2	Breast reactive	CAV1, MYH11, RAB11A, RAB11B, CTNNB1, GAPDH, RBM15
3	Cell cycle	CDK1, CCNB1, CCNE1, CCNE2, CDKN1B, PCNA, FOXM1
4	Core reactive	CAV1, CTNNB1, RBM15, CDH1, CLDN7
5	DNA damage response	TP53BP1, ATM, BRCA2, CHEK1, CHEK2, XRCC5, MRE11A, TP53, RAD50, RAD51, XRCC1
6	EMT	FN1, CDH2, COL6A1, CLDN7, CDH1, CTNNB1, SERPINE1
7	PI3K/AKT	AKT1, AKT2, AKT3, GSK3A, GSK3B, CDKN1B, AKT1S1, TSC2, INPP4B, PTEN
8	RAS/MAPK	ARAF, JUN, RAF1, MAPK8, MAPK1, MAPK3, MAP2K1, MAPK14, RPS6KA1, YBX1
9	RTK	EGFR, ERBB2, ERBB3, SHC1, SRC
10	TSC/mTOR	EIF4EBP1, RPS6KB1, MTOR, RPS6, RB1
11	Hormone receptor	ESR1, PGR, AR
12	Hormone signaling	INPP4B, GATA3, BCL2

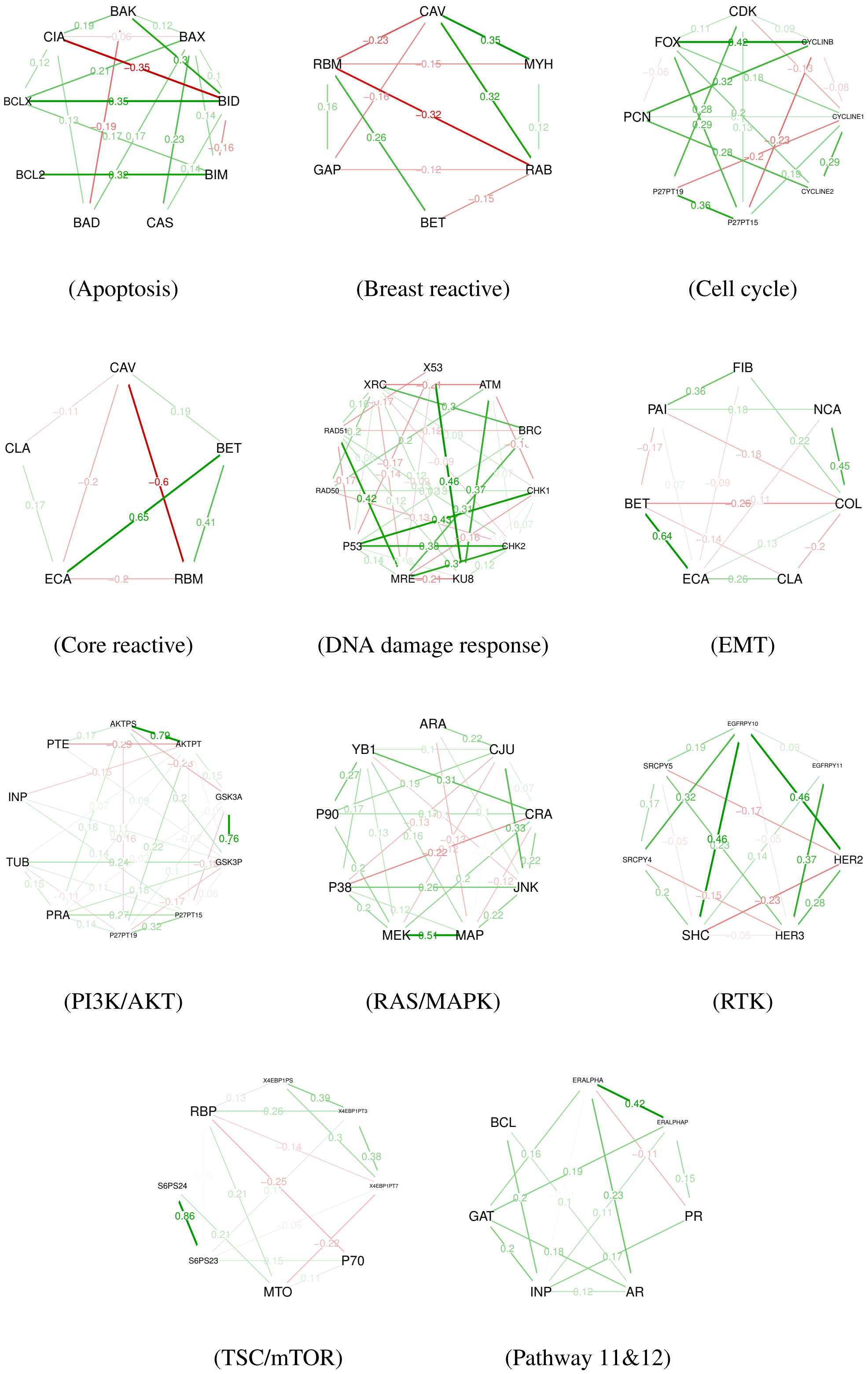
Table 2.2: 12 pathways and mRNA selection results

#	pathway names	# of proteins	# of mRNAs	mRNA selected
1	Apoptosis	9	9	BCL2
2	Breast reactive	6	7	-
3	Cell cycle	8	7	CCNE1
4	Core reactive	5	5	CDH1
5	DNA damage response	11	11	-
6	EMT	7	7	CDH1
7	PI3K/AKT	10	10	INPP4B
8	RAS/MAPK	9	10	-
9	RTK	7	5	ERBB2
10	TSC/mTOR	8	5	-
11&12	Hormone receptor&signaling	7	4	INPP4B, GATA3, BCL2

The protein networks for all 12 pathways are shown in Figure 2.3. The number on each edge is the estimated partial correlation between two proteins it connects. Green edge means positively partial correlated; red edge means negatively partial correlated. The thickness of the edge represents the magnitude of the absolute value of partial correlation. Within each network, majority of the proteins tends to be positively correlated, which means most of the proteins are working together within each pathway. Proteins which are related to the same gene family have high positive correlation in the graph. For example, AKTPS and AKTPT (AKT gene family) in PI3K/AKT pathway, GSK3A and GSK3P (GSK gene family) in PI3K/AKT pathway, and S6PS24 and S6PS23 (RPS6 gene family) in TSC/mTOR pathway. We define the degree of freedom for nodes as the number of edges connected to them. Then hub nodes are the nodes which have the largest degree of freedom in each pathway. These are the proteins which have the maximum connectivities and interact heavily with other proteins. The summary of hub nodes are shown in Table 2.3 below (the number in the bracket is the degree of freedom of the hub node).

Table 2.3: Hub nodes in each pathway

pathway	hub nodes	pathway	hub nodes
Apoptosis	BAX(7)	Breast reactive	RAB(5), RBM(5)
Cell cycle	CYCLINE1(7)	Core reactive	CAV(4), ECA(4)
DNA damage response	XRC(9)	EMT	ECA(6), COL(6)
PI3K/AKT	ATKPT(9), GSK3P(9)	RAS/MAPK	CJU(8)
RTK	EGFRPY10(6), HER3(6), SHC(6)	TSC/mTOR	S6PS23(6)
Hormone receptor&signaling	INPP4B(6)		



(Apoptosis)

(Breast reactive)

(Cell cycle)

(Core reactive)

(DNA damage response)

(EMT)

(PI3K/AKT)

(RAS/MAPK)

(RTK)

(TSC/mTOR)

(Pathway 11&12)

Figure 2.3: Protein networks for all 12 pathways

2.7 Further Discussion

Our model provides a framework for jointly learning about the nonlinear covariate structure as well as the dependence graph structures for the Gaussian graphical model. We used fixed knot splines for estimation of nonlinear functions. This can be extended for adaptive splines or other adaptive basis functions [46]. We have introduced our model for decomposable graphs but it can be extended for more general settings. Apart from genomic applications, there are numerous problems that arise in finance, econometrics, and biological sciences where nonlinear graph models can be a useful approach to modeling and therefore, we expect our inference procedure to be effective in those applications. Extension from Gaussian to non-Gaussian models is an interesting topic for future research.

3. BAYESIAN GRAPH SELECTION CONSISTENCY UNDER MODEL MISSPECIFICATION [*]

3.1 Introduction

In this chapter, focusing on the hyper-inverse g-Wishart (g-HIW) distribution 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 $\alpha < 1/3$. The key conditions relate to precise upper and lower bounds on the partial correlation and a suitably complexity prior on the space of graphs. We emphasize here that we do not need conditions to be verified on all subgraphs, i.e. all assumptions are easy to understand and relatively straightforward to verify. Regarding our findings, we discover that g-HIW prior places heavy penalty on missing true edges (false negatives), but comparatively smaller penalty on adding false edges (false positives). Henceforth in 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 [31] since hyper-inverse Wishart prior is a special case of DAG-Wishart prior proposed in [65] under perfect DAGs. However, the assumptions in this chapter are distinctly different from those stated in [31]. In particular, our assumptions are on the magnitude of the elements of partial correlation matrix rather than on the eigen values of covariance matrix as in [31]. 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, we are the first to show the strong selection consistency under HIW prior 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

[*] Reprinted with permission from arXiv, “Bayesian Graph Selection Consistency Under Model Misspecification”, arXiv preprint arXiv:1901.04134, 2019, by Niu, Yabo and Pati, Debdeep and Mallick, Bani K. In accordance arXiv copyright no modifications have been made except formatting.

is inspired by [30], but extends to the case when p is growing with n and provides a rigorous proof 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 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, i.e. addition and deletion of one edge using decomposable graph chain rule 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 and tail bounds for sample partial correlation, we obtain bounds for ratios of local marginal likelihoods which are then combined to yield strong selection consistency results.

The remaining part of this chapter is organized as follows. In Section 3.2, we introduce the necessary background and notations. Section 3.3 introduces the model with the HIW prior. Section 3.4 describes the main results on pairwise posterior ratio consistency and consistent graph selection when the true graph is decomposable. Section 3.5 states the main results on consistent graph selection under model misspecification and results on equivalence of minimal triangulations. In each of Sections 3.4 and 3.5, the results are presented progressively as follows: First we provide a non-asymptotic sharp upper bound for pairwise Bayes factor. Next, we state the main theorem for posterior ratio consistency when p diverges with n with p of the order n^{α} for $\alpha < 1/2$. Finally, we state the main theorem on strong graph selection consistency which further requires $\alpha < 1/3$. Numerical experiments are presented in Section 3.6 followed by a discussion in Section 3.7.

3.2 Preliminaries

In this section, we define a collection of notations required to describe the model and the prior. Section 3.2.1 introduces sample and population correlations and partial correlations. Section 3.2.2 contains matrix abbreviations and notations used throughout the dissertation. Section 3.2.3 addresses other notations that are necessary for theorems and proofs. **Notice, in this chapter and**

in Appendix D to Appendix G, we use p to denote the dimension of any graph.

3.2.1 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) samples observations from $\boldsymbol{X}_p$. Clearly, the $n \times p$ matrix formed by augmenting the n-dimensional column vectors x_i, denoted $(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_p)$ 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 an $n \times n$ identity matrix.

Definition 3.2.1. (Population correlation coefficient). *The population correlation coefficient between X_i and X_j, $1 \leq i, j \leq p$, 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 3.2.2. (Sample/Pearson correlation coefficient). *The sample correlation coefficient between X_i and X_j, $1 \leq i, j \leq p$, 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 3.2.3. (Population partial correlation coefficient). *Let $S = \{i_1, i_2, \ldots, i_{|S|}\}$, where $1 \leq i_1, i_2, \ldots, i_{|S|} \leq p$ and $|S|$ is the cardinality of set S. Define $X_S = (\mathrm{X}_{i_1}, \mathrm{X}_{i_2}, \ldots, \mathrm{X}_{i_{|S|}})^T$. The population partial correlation coefficient between X_i and X_j, where $i, j \notin S$ and $1 \leq i, j \leq p$, 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 3.2.4. (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* $1 \leq i, j \leq p$*, 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* $\hat{\sigma}_{Si} = (\mathrm{x}_S - \bar{\mathrm{x}}_S)^T(\mathrm{x}_i - \bar{\mathrm{x}}_i)/n$*,* $\hat{\sigma}_{SS} = \big\{(\mathrm{x}_S - \bar{\mathrm{x}}_S)^T(\mathrm{x}_S - \bar{\mathrm{x}}_S)/n\big\}^{-1}$*,* $\bar{\mathrm{x}}_S = (\bar{\mathrm{x}}_{i_1}\mathbb{1}_n, \ldots, \bar{\mathrm{x}}_{i_{|S|}}\mathbb{1}_n)$*.*

3.2.2 Matrix Notation

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 $1 \leq i_1 < i_2 < \ldots < i_{|C|} \leq p$, 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.

3.2.3 Miscellaneous

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 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 $C_1 a \leq b \leq C_2 a$ for constants C_1, C_2. $a \precsim b$ denotes $a \leq C_3 b$ for a constant C_3. For set relations, $A \subset B$ means A is a subset of B; $A \subsetneq B$ means $A \subset B$ and $A \neq B$; $A \not\subset B$ means A is not a subset of B. $|\cdot|$ determined by context can be absolute value, cardinality of sets or determinant of matrices. $\pi(\cdot)$ and $\pi(\cdot \mid \mathrm{Y})$ are the prior distribution and posterior distribution of graphs, respectively.

3.3 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 1.1. Assume $Y_i \mid \Sigma_G, G \sim \mathrm{N}_p(0, \Sigma_G)$. In matrix notations,

$$\mathrm{Y}_{n\times p} \mid \Sigma_G, G \sim \mathrm{MN}_{n\times p}(\mathbf{0}_{n\times p}, I_n, \Sigma_G), \tag{3.1}$$

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 in Section 2.2.1.

Since the joint density factorizes over cliques and separators,

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

in the same way as in Section 2.2.3, and

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

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

$$h(G,b,D) = \frac{\prod_{C\in\mathcal{C}} \left|\frac{1}{2}D_C\right|^{\frac{b+|C|-1}{2}} \Gamma^{-1}_{|C|}\left(\frac{b+|C|-1}{2}\right)}{\prod_{S\in\mathcal{S}} \left|\frac{1}{2}D_S\right|^{\frac{b+|S|-1}{2}} \Gamma^{-1}_{|S|}\left(\frac{b+|S|-1}{2}\right)},\ w(C) = \frac{|D_C|^{\frac{b+|C|-1}{2}} \left|D_C + \mathrm{Y}_C^T\mathrm{Y}_C\right|^{-\frac{b+n+|C|-1}{2}}}{2^{-\frac{n|C|}{2}}\Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right)\Gamma^{-1}_{|C|}\left(\frac{b+n+|C|-1}{2}\right)}.$$

Throughout the remainder of this dissertation, we shall be working with the hyper-inverse Wishart g-prior [17], denoted as

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

where g is some suitably small fraction in $(0, 1)$ and $b > 0$ is a fixed constant. Following the recommendation in [17], we choose $g = 1/n$ through the remainder of this dissertation. Intuitively, this choice of g avoids overwhelming the likelihood asymptotically as well as arbitrarily diffusing the prior. In that case,

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

The choice of focusing on the hyper-inverse Wishart g-prior in this dissertation is driven by the following two reasons. First, we can simplify the edge/signal strength assumption in terms of the smallest nonzero entries 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 3.4 and 3.5 continue to hold for any choice of HIW prior. The proof techniques under HIW g-prior serve as representations to the principle ideas in the article and can be easily adapted to other variations of HIW prior.

To complete a fully Bayesian specification, we place a prior distribution $\pi(\cdot)$ on the decomposable graph G. Our theoretical results in Section 3.4 and 3.5 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 [18, 66, 17, 13, 31] we consider in the dissertation. Considering an undirected decomposable graph G, we assume the edges are independently drawn from a Bernoulli distribution with a common probability q:

$$\pi(G \mid q) \propto \left[\prod_{r<s} q^{e_{rs}}(1-q)^{1-e_{rs}}\right] \cdot \mathbb{1}_{\mathcal{D}}(G), \tag{3.4}$$

where $\mathcal{D}$ is the set of all decomposable graphs with $|V| = p$ vertices and q is the prior edge inclusion probability. We control the parameter q to induce sparsity on the number of edges. [18] recommends using $2/(|V|-1)$ as the hyper-parameter for the Bernoulli distribution. For an undirected graph, it has peak around $|V|$ edges and the mode is smaller for decomposable graphs. We outline specific choices in Section 3.4 and Section 3.5 below.

3.4 Theoretical Results In The Well-specified Case

In this section, we present our main consistency results. The proofs of the results are deferred to Appendix D to Appendix G. Before introducing the assumptions, we need to adapt previous notations to the high-dimensional graph selection problem. Let $\mathrm{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ and $\Omega_0 = \Sigma_0^{-1}$ the corresponding precision matrix. Without loss of generality, we assume all column means of Y are zero. Let $G_t = (V, E_t)$ denote the true decomposable graph induced by Ω_0, $\rho_{ij|V\backslash\{i,j\}}$ denote the true partial correlation between node i and j given the rest of the nodes $V\backslash\{i,j\}$. Assume ρ_L and ρ_U are the smallest and largest in absolute value of the *non-zero* population partial correlations, i.e.

$$\rho_L = \min_{\substack{1\le i<j\le p \\ (i,j)\in E_t}} \left|\rho_{ij|V\backslash\{i,j\}}\right|, \quad \rho_U = \max_{\substack{1\le i<j\le p \\ (i,j)\in E_t}} \left|\rho_{ij|V\backslash\{i,j\}}\right|,$$

Let $G_a = (V, E_a)$ be any alternative decomposable graph other than the true graph G_t. Denote by $E_a^1 = E_t \cap E_a^n$ the set of true edges in G_a. Notice, when $E_t \subsetneq E_a$, we have $E_a^1 = E_t$. Denoting by $|\cdot|$ the cardinality of a set, $|E_t|$ is the number of edges in G_t, $|E_a^1|$ is the number of true edges in G_a. Define $G_c = (V, E_c)$, where $E_c = \{(i,j) : e_{ij} = 1, 1 \le i < j \le p\}$, to be the complete graph such that $|E_c| = p(p-1)/2$. By definition, G_c is a decomposable graph. We use $G_a \neq G_t$ to denote $E_a \neq E_t$; $G_a^n \not\subset G_t$ to denote $E_a \not\subset E_t$; $G_a \subsetneq G_t$ to denote $E_a \subsetneq E_t$. In the following, we state the main assumptions for graph selection consistency.

Note: The Appendix D introduces a set of auxiliary results related to the concentration and tail behavior of partial correlations, following by Appendix E which states bounds for Bayes factor for local moves required to prove Theorem 3.4.1. Then we provide a proof of Theorem 3.4.1 followed by the proofs of Theorem 3.4.2, Theorem 3.4.3, Corollary 3.4.2, the minimal triangulation Theorems 3.5.1 and 3.5.2 and Corollary 3.5.1.

3.4.1 Assumptions

Assumption 3.4.1. (Graph size)

$$p \precsim n^{\alpha}, \textit{ where } 0 < \alpha < 1.$$

Assumption 3.4.2. (Edge sensitivity and identifiability)

$$\rho_L \asymp n^{-\lambda}, \textit{ where } 0 \leq \lambda < \frac{1}{2}.$$

Assumption 3.4.3. (Number of maximum edges in G_t^n)

$$|E_t| \precsim n^{\sigma}, \textit{ where } 0 \leq \sigma \leq 2\alpha.$$

Assumption 3.4.4. (Prior edge inclusion probability)

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

Assumption 3.4.5. (Imperfect linear relationship)

$$1 - \rho_U \asymp n^{-k}, \textit{ where } k \geq 0 \textit{ and } \rho_U \neq 1.$$

The main results will have additional restrictions on the parameters (α, λ), but it is important to note that we require ρ_L to not decrease to 0 too quickly in order to ensure that the graph is identifiable. On the other hand, ρ_U can be allowed to be sufficiently close to 1.

3.4.2 Pairwise Bayes Factor Consistency for Fixed p

In this section, we assume p, ρ_U and ρ_L are all fixed constants. As a first step towards model selection, we investigate the behavior of the pairwise Bayes factor

$$\mathrm{BF}(G_a; G_t) = \frac{f(\mathrm{Y} \mid G_a)}{f(\mathrm{Y} \mid G_t)}, \tag{3.5}$$

where G_t is the decomposable true graph and G_a is any other decomposable graph. In this section, we shall investigate sufficient conditions on the likelihood (3.2) and the prior on (Σ_G, G) given by (3.3) and (3.4) such that the Bayes factor (3.5) converges to 0 as $n \to \infty$ for any graph $G_a \neq G_t$.

Theorem 3.4.1. (Upper bound for pairwise Bayes factor). *Assume the graph dimension* p *is a fixed constant and* $\rho_U \neq 1$. *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* $G_t \not\subset G_a$,
$$\mathrm{BF}(G_a; G_t) < \exp\Big\{ -\frac{n\rho_L^2}{2} + \delta(n) \Big\}, \tag{3.6}$$

2. *when* $G_t \subsetneq G_a$,
$$\mathrm{BF}(G_a; G_t) < \big(e^{p^2}\big) \cdot n^{-\frac{1}{2}(|E_a| - |E_t|)(1 - 2/\tau^*)}, \tag{3.7}$$

and

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

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* $n \to \infty$.

Proof. See Appendix F.2. □

The next corollary is the direct result from Theorem 3.4.1.

Corollary 3.4.1. (Finite graph pairwise Bayes factor consistency). *Let* G_a *be any decomposable graph and* $G_a \neq G_t$. *The graph dimension* p *is a fixed constant. If* $\rho_U \neq 1$, *then* $\mathrm{BF}(G_a; G_t) \xrightarrow{\mathbb{P}} 0$, *as* $n \to \infty$.

When p is fixed, the likelihood is strong enough to consistently recover the graph. One key aspect of the proof is that Bayes factor in favor of adding a true edge versus the lack of it is exponentially small, while the Bayes factor in favor for adding a false edge decreases to zero only at a polynomial rate.

We emphasize here that exponential rate for deletion (of true edges) is only true when the corresponding population partial correlation or correlation is non-zero. From the *global Markov property*, we know if two nodes are adjacent then any partial correlation between them is non-zero but their correlation can be zero. The polynomial rate for addition (of false edges) is only true when the corresponding population partial correlation or correlation is zero. When two nodes are not adjacent, then only the set that separates them will results in a zero partial correlation. We choose the path of $G_t \to G_c \to G_a$ which ensures us the exponential decay when missing true edges and polynomial decay when adding false edges.

3.4.3 Posterior Ratio Consistency for Growing p

Next we examine the convergence of posterior ratio,

$$\mathrm{PR}(G_a; G_t) = \frac{f(\mathrm{Y} \mid G_a)\pi(G_a)}{f(\mathrm{Y} \mid G_t)\pi(G_t)}, \tag{3.8}$$

when the dimension of graphs grows with sample size.

Theorem 3.4.2. (High-dimensional graph posterior ratio consistency). *Let G_a be any decomposable graph and $G_a \neq G_t$ and Assumptions 3.4.1-3.4.5 are satisfied with*

$$0 < \alpha < \frac{1}{2}, \quad 0 \leq \lambda < \min\Big\{\alpha, \frac{1}{2} - \alpha\Big\}.$$

By choosing γ in the interval $(\max\{0, 1-4\alpha\}, 1-\sigma-2\lambda)$ *we have* $\mathrm{PR}(G_a; G_t) \xrightarrow{\mathbb{P}} 0$, *as* $n \to \infty$.

Proof. See Appendix F.3. □

When the graph size grows with n, the partial correlation is no longer a constant. The HIW prior does not naturally favor parsimonious graphs, so a penalty on the number of edges in the graph in needed by restricting γ in the above interval. Note also that we do not need any further restriction on σ in Assumption 3.4.3 meaning that the true graph is allowed to be the complete graph for the posterior ratio consistency to hold.

3.4.4 Strong Graph Selection Consistency

In this section, we examine the behavior of

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

as $n, p \to \infty$.

Theorem 3.4.3. (Strong graph selection consistency). *Let G_a be any decomposable graph and $G_a \neq G_t$ and Assumptions 3.4.1-3.4.5 are satisfied with*

$$0 < \alpha < \frac{1}{3}, \quad 0 \leq \lambda < \min\Big\{\alpha, \frac{1-3\alpha}{2}\Big\}.$$

By choosing γ in the interval $(\max\big\{\alpha, 1-4\alpha\big\}, 1-\sigma-2\lambda)$, we have

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

Proof. See Appendix F.4. □

Strong selection consistency demands all posterior ratio 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 sum convergent, we require further assumptions on the model complexity and an accompanying stronger penalty π. We achieve this by shrinking the dimension of graph space ($\alpha < 1/3$) and inducing a slightly stronger sparsity (by selecting larger γ) on the prior over the graph space.

In the proofs of Theorem 3.4.1-3.4.3, by using the decomposable graph chain rule, we traverse to any decomposable graph from the true graph and thus break down the Bayes factor into local moves, i.e. addition and deletion of a single edge. The local moves then can be associated with sample partial correlations and sample correlations, which are the natural criterion of edge selection by definition. This enables us to transform the problem into a more understandable manner.

In practice, one might be interested in a consistent point estimate rather than the entire posterior distribution. In Bayesian inference for discrete configurations, a posterior mode provides a natural surrogate for the MLE. In the following, we investigate the consistency of the posterior mode obtained from our hierarchical Bayesian model as a simple bi-product of Theorems 3.4.2 and 3.4.3. Define $\hat{G}$ to be the posterior mode in the decomposable graph space, i.e.

$$\hat{G} = \text{argmax}_{G \in \mathcal{D}} \pi(G \mid \text{Y}).$$

Then the following in true.

Corollary 3.4.2. (Consistency of posterior mode when G_t is decomposable). *Under the assumptions of Theorem 3.4.3, the probability which the posterior mode $\hat{G}$ is equal to the true graph G_t goes to one, i.e.*

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

Proof. See Appendix F.5. □

3.5 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.

3.5.1 Minimal Triangulations

We begin with some definitions on triangulation and minimal triangulations of a graph. A triangulation of graph $G = (V, E)$ is a decomposable graph $G^{\Delta} = (V, E \cup F)$. The edges in F are called *fill-in* edges. A triangulation $G^{\Delta} = (V, E \cup F)$ of $G = (V, E)$ is minimal if $(V, E \cup F')$ is non-decomposable for every $F' \subsetneq F$ [67]. A triangulation is minimal if and only if the removal of any single fill-in edge from it results in a non-decomposable graph [68, 67]. This property captures the important aspect of minimal triangulations. For a summary of minimal triangulations of graphs, see [67] for more details. Next, we state two theorems graph selection consistency under a true non-decomposable graph.

3.5.2 Consistency Results Under Model Misspecification

Theorem 3.5.1. (Convergence and equivalence 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_U, ρ_U are fixed constants), we have the following:*

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 $\mathrm{BF}(G_a; G_m) \xrightarrow{\mathbb{P}} 0$, as $n \to \infty$.*

2. *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). Then the Bayes factor between them are stochastically bounded, i.e. for any $0 < \epsilon < 1$, there exist two positive finite constants $A_1(\epsilon) < 1$ and $A_2(\epsilon) > 1$, such that*

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

3. *If $\rho_U \neq 1$, 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.*

Proof. See Appendix G.2. □

Theorem 3.5.2. (Convergence and equivalence 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

$$0 < \alpha < \frac{1}{2}, \quad 0 \leq \lambda < \min\Big\{\alpha, \frac{1}{2} - \alpha\Big\}, \quad 0 < \sigma < \min\Big\{2(\alpha - \lambda), 2(\frac{1}{2} - \alpha - \lambda)\Big\}.$$

Choose γ in the interval $(\max\big\{2\alpha, 1-2\alpha\big\}, 1-\sigma-2\lambda)$. Then under Assumptions 3.4.1-3.4.5, we have $\mathrm{PR}(G_a; G_m) \overset{\mathbb{P}}{\to} 0$, as $n \to \infty$.

2. *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.*

3. *If*

$$0 < \alpha < \frac{1}{3}, \quad 0 \leq \lambda < \min\Big\{\alpha, \frac{1-3\alpha}{2}\Big\}, 0 \leq \sigma < \min\Big\{2(\alpha - \lambda), 2\Big(\frac{1-3\alpha}{2} - \lambda\Big)\Big\}.$$

And we choose γ in the interval $(\max\big\{3\alpha, 1-2\alpha\big\}, 1-\sigma-2\lambda)$, then under Assumptions 3.4.1-3.4.5, we have $\sum_{G_m \in \mathcal{M}_t} \pi(G_m \mid \mathrm{Y}) \overset{\mathbb{P}}{\to} 1$, as $n \to \infty$, where $\mathcal{M}_t$ is the minimal triangulation space of G_t.

Proof. See Appendix G.3. □

Based on the theorems presented above, 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 the true graph. In that case, there are too many possibilities among the minimal triangulations such that they can be vastly different for each other. It is worth mentioning that any decomposable subgraph of the true graph is not a good posterior estimate of the true graph. This is simply due to the fact that such a graph is associated with at least one edge deletion step following by reciprocal of addition steps from a minimal triangulation. Since deletion of any true edge results in an exponential decay of the Bayes factor in favor of the deletion and the reciprocal of additions will be in favor of additions (the minimal triangulations) or neutral depending on whether the corresponding population partial correlation is zero. Thus, pairwise speaking, the posterior mode is among minimal triangulation class.

Analogous to Corollary 3.4.2, when the true graph G_t is not decomposable, we state the behavior of posterior mode in the following corollary under model misspecification.

Corollary 3.5.1. (Consistency of posterior mode when G_t is non-decomposable). *Under the assumptions of Theorem 3.5.2, 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, i.e.*

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

Proof. See Appendix G.4. □

3.6 Simulations

We conduct two sets of simulations for the demonstrate the convergence of Bayes factors in the well-specified case (Theorem 3.4.1) and in the misspecified case (Theorem 3.5.1) for fixed p.

3.6.1 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. Let $\mathcal{G}_k$ be the k-dimensional graph space. 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 3.1.

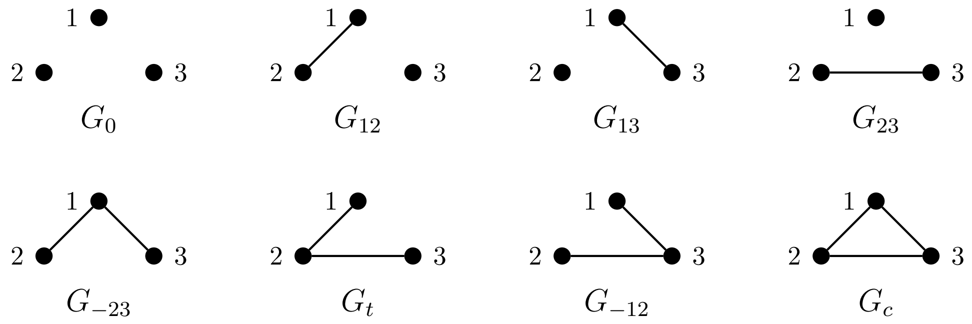


Figure 3.1: Enumerating all 3-node decomposable graphs in $\mathcal{D}_3$ with G_t as the true graph, G_0 as the null graph and G_c as the complete graph. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

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 independent and identically 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 degree of freedom b in the HIW g-prior is chosen to be 3. The first six pairwise Bayes factors in logarithmic scale is shown in Figure 3.2 (a) and the logarithm of $\mathrm{BF}(G_c; G_t)$ is shown separately in Figure 3.2 (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 3.1. To calculate the leading terms in the logarithm of Bayes factors, the sample partial correlations or sample correlations are replaced with their population counterparts that do not depend on n. The leading terms are obtained by following the route we have used in the proof, i.e. $G_t \to G_c \to G_a$. The slopes of logarithms of the first six Bayes factors in Figure 3.2 (a) are calculated in Table 3.1 based on linear regression fit on n. The last slope in Table 3.1 is calculated based on linear regression on $\log n$; refer to Figure 3.2 (b). Table 3.1 shows that the theoretical asymptotic leading terms match well with the empirical values.

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

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

Table 3.1: Asymptotic leading terms and simulation slopes of Bayes factors in logarithmic scale. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

Bayes factors	**asymptotic leading term**	**simulation slope**
BF$(G_0; G_t)$	$\left\{\log\left(1-\rho_{12}^2\right)+\log\left(1-\rho_{23}^2\ \ \right)\right\}\cdot n/2 = \boldsymbol{-0.2967\cdot n}$	$\mathbf{-0.2963}$
BF$(G_{13}; G_t)$	$\left\{\log\left(1-\rho_{12}^2\right)+\log\left(1-\rho_{23\mid 1}^2\right)\right\}\cdot n/2 = \boldsymbol{-0.2639\cdot n}$	$\mathbf{-0.2637}$
BF$(G_{23}; G_t)$	$\log\left(1-\rho_{12}^2\ \ \right)\cdot n/2 = \boldsymbol{-0.1767\cdot n}$	$\mathbf{-0.1765}$
BF$(G_{-12}; G_t)$	$\log\left(1-\rho_{12\mid 3}^2\right)\cdot n/2 = \boldsymbol{-0.1438\cdot n}$	$\mathbf{-0.1439}$
BF$(G_{12}; G_t)$	$\log\left(1-\rho_{23}^2\ \ \right)\cdot n/2 = \boldsymbol{-0.1120\cdot n}$	$\mathbf{-0.1198}$
BF$(G_{-23}; G_t)$	$\log\left(1-\rho_{23\mid 1}^2\right)\cdot n/2 = \boldsymbol{-0.0872\cdot n}$	$\mathbf{-0.0873}$
BF$(G_c; G_t)$	$\boldsymbol{-0.5\cdot\log n}$	$\mathbf{-0.5106}$

From the simulation results, we can see missing at least one true edge of G_t in G_a will result in the Bayes factor converging to zero exponentially. This is perfectly illustrated by all six Bayes factors in Figure 3.2 (a). On the other hand, adding false edges in G_a results in a Bayes factor going to zero at a polynomial rate which is much slower than missing a true edge, see Figure 3.2 (b). These discoveries are consistent with Table 3.1 and our proofs.

Next we compare the different types of rates in the convergence of the first six Bayes factors. The convergence rate associated with missing two edges of G_t is faster than missing only one edge, i.e. BF$(G_0; G_t)$ vs. BF$(G_{23}; G_t)$ and BF$(G_0; G_t)$ vs. BF$(G_{12}; G_t)$. The convergence rate is faster when the missing edge of G_t corresponds to a larger partial correlation (or correlation) in absolute value, i.e. BF$(G_{-12}; G_t)$ vs. BF$(G_{-23}; G_t)$ and BF$(G_{23}; G_t)$ vs. BF$(G_{12}; G_t)$. One interesting fact is although G_0 and G_{13} are both missing two edges of G_t, with G_{13} having an additional false edge of G_t 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 3.1. As the absolute value of correlation between node 2 and 3 ($|\rho_{23}| = 0.4619$) is larger than the absolute value of partial correlation between them given node

1 ($|\rho_{23|1}| = 0.4$), the leading term of BF$(G_0; G_t)$ is smaller than that of BF$(G_{13}; G_t)$. The effect due to false edges (polynomial rate) is overwhelmed by the leading term (exponential rate). It is evident that HIW prior places higher penalties on false negative edges compared to false positive edges. Hence in the high-dimensional case, a prior on graph space is needed for penalizing false positive edges. Similar conclusions can be made comparing BF$(G_{23}; G_t)$ and BF$(G_{-12}; G_t)$, also from comparing BF$(G_{12}; G_t)$ and BF$(G_{-23}; G_t)$.

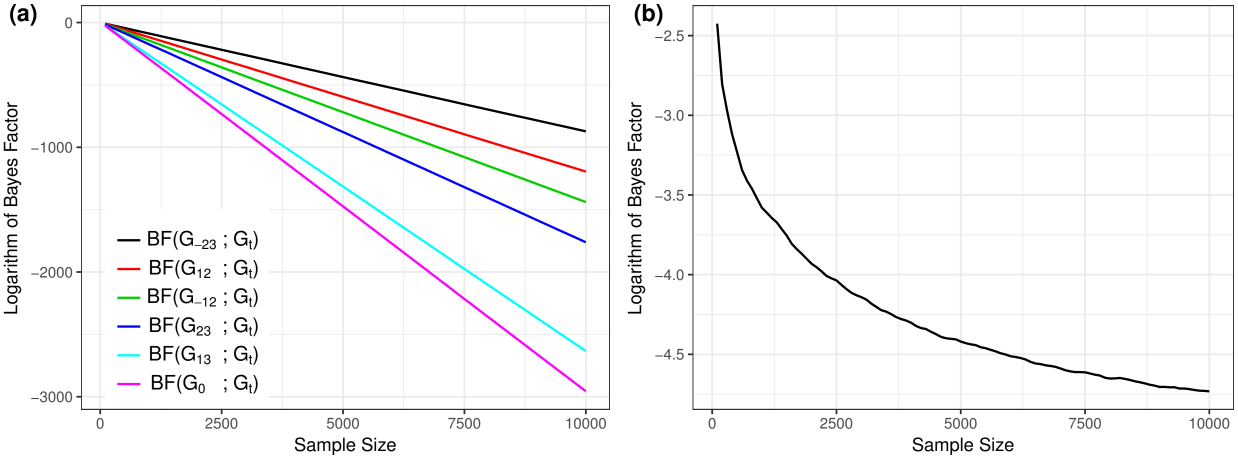


Figure 3.2: Simulation results of pairwise Bayes factors of $\mathcal{D}_3$ in logarithmic scale. (a) Six Bayes factors where $G_t \not\subset G_a$ (at least missing one edge in G_t). (b) When $G_t \subsetneq G_a = G_c$ (only addition). Reprinted with permission from arXiv preprint, arXiv:1901.04134.

3.6.2 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 graph increases exponentially with the dimension of graphs, we only select 5 alternative graphs in $\mathcal{D}_4$ other than the minimal triangulations, see Figure 3.3. 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 in the simulation of $\mathcal{D}_3$.

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

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

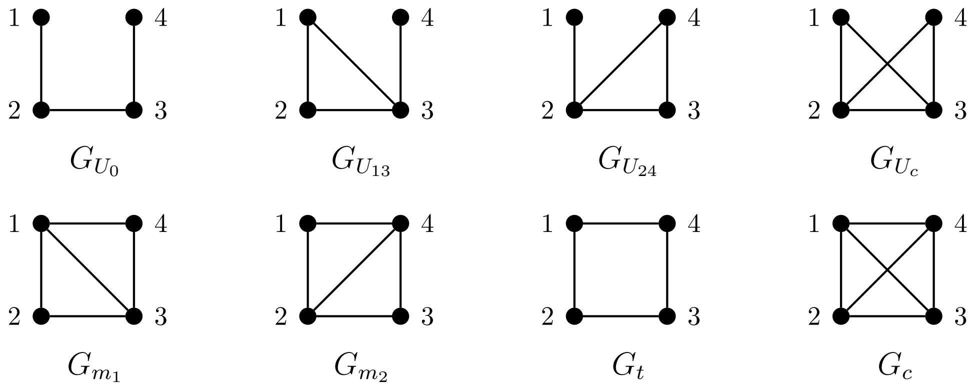


Figure 3.3: Some selected graphs in $\mathcal{G}_4$, including G_t as the true graph which is non-decomposable. G_{m_1} and G_{m_2} are two minimal triangulations of G_t. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

Since the true graph G_t is non-decomposable, the two minimal triangulations of G_t act like the pseudo-true graphs. So we plot the first four pairwise Bayes factors where $G_{m_i} \not\subset G_a$, $i = 1, 2$ for G_{m_1} and G_{m_2} in logarithmic scale together in Figure 3.4 (a) and (b), respectively. The logarithm of Bayes factor between two minimal triangulations is in Figure 3.4 (c). Finally, we plot the Bayes factors of one triangulation (i.e. G_c, not minimal) of G_t against both minimal triangulations in Figure 3.4 (d).

From Figure 3.4 (a) and (b), we can see the behavior of two minimal triangulations is the same as what we observed in the case where Bayes factors against the true decomposable graph, i.e. missing true edges causes exponential decay of pairwise Bayes factors. And in the case of false positive edges, i.e. Figure 3.4 (c), the rate is what we expected if G_{m_1} and G_{m_2} are the true graph, polynomial rate. Based on the simulation result in Figure 3.4 (c), we can see the Bayes factor between two minimal triangulations neither converges to zero nor diverges to infinity. And they are stochastically bounded. In this case, it is closely to 1 which means these two minimal triangulations of G_t are almost the same in this case (in terms of posterior probability). It is also demonstrated by Figure 3.4 (a), (b) and (d) where the curves between G_{m_1} and G_{m_2} are almost identical.

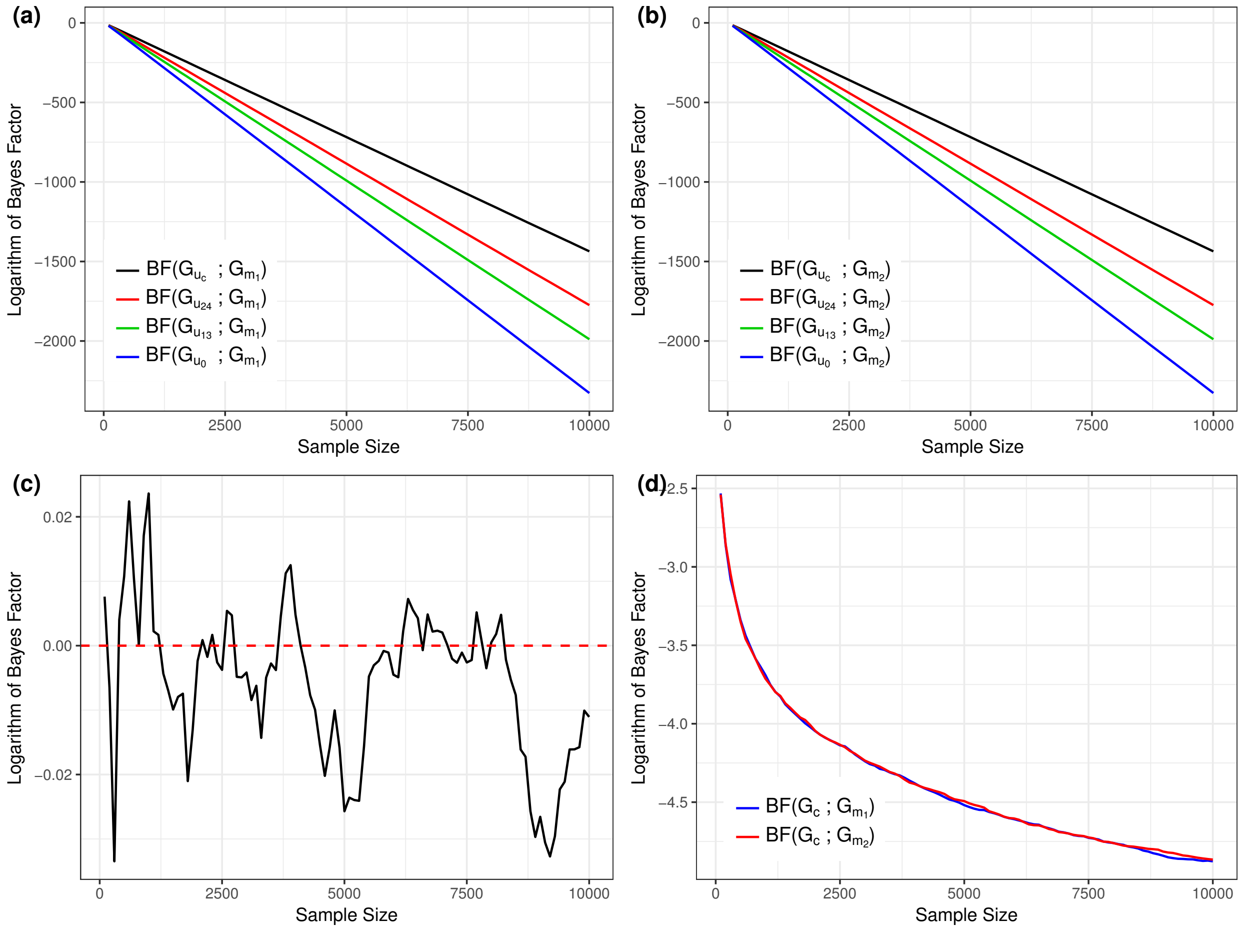


Figure 3.4: Simulation results of pairwise Bayes factors of $\mathcal{D}_4$ in logarithmic scale. (a) When $G_{m_1} \not\subset G_a$ (missing true edges). (b) When $G_{m_2} \not\subset G_a$ (missing true edges). (c) The Bayes factor between two minimal triangulations of G_t, i.e. BF$(G_{m_2}; G_{m_1})$. (d) When $G_{m_i} \subsetneq G_a = G_c$, $i = 1, 2$ (only addition). Reprinted with permission from arXiv preprint, arXiv:1901.04134.

3.7 Discussion

In this chapter, 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 [30] 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 scenario. To the best of our knowledge, these are the first complete results on this topic so far.

In our results, the graph dimension can not be equal to or exceed $n^{1/2}$ and $n^{1/3}$ for pairwise 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 i.i.d. case without any sparsity assumptions, it is well-known that the MLE is consistent under "p/n small", the Fisher expansion for the MLE is valid under "p^2/n small" while the Wilks and asymptotic normality results apply under "p^3/n small" [69, 70]. 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 the HIW prior. This is simply because HIW does not penalize false edges significantly enough so that in high dimension a prior on graph space is needed to achieve both pairwise and strong selection consistency. Also any other sparsity restriction on the elements of the precision matrix is not supported by the HIW prior due to its inability to enforce sufficient shrinkage conditional on the graph. This limits extending the technical results to ultra-high-dimensional case by enforcing additional sparsity assumptions on the elements of the precision matrix. This apparent "flaw" lies in the construction of the HIW prior itself and can not be improved by adding any reasonable penalty on the graph space.

For technical simplicity, our results are based on HIW g-prior only. We conjecture that the consistency results continue to hold for general HIW prior. Moreover, extensions to non-decomposable graphical models can be done by using G-Wishart prior, but major bottlenecks are expected stemming from the lack of a closed form for the normalizing constant for the general HIW prior. Recent work [71] on the development of approximation results for the normalizing constant may prove to be useful in this regard.

4. SUMMARY AND ONGOING RESEARCH

4.1 Summary

In this dissertation, I first developed a Bayesian method to incorporate covariate information in Gaussian graphical models by adding a linear or nonlinear framework in the mean structure of the Gaussian distribution. In order to select the important covariates, I applied a Bayesian variable selection scheme to the covariate structure assumed. It enables us to simultaneously estimate the graph structure and select the influential variable to the same graph. To examine the property of variable selection in this scenario, I studied the consistency of variable selection. The theorems conclude that under moderate conditions the consistency can be achieved with graph selection even if the underlying graph chosen is not the true graph. This guarantees the convergence of the stochastic search algorithm. I also developed an efficient collapsed Gibbs sampler algorithm to search the joint model space, i.e. covariate space and graph space. The simulation results confirm the theoretical finding which is that variable selection is not affected by graph selection and it converges fast. This method can be applied to estimate protein networks with the ability to identify influential mRNAs. I applied the proposed method to analyze gene and protein expression data acquired from The Cancer Genome Atlas (TCGA) and The Cancer Proteome Atlas (TCPA). The results are consistent with some biological properties of the selected mRNAs.

In the second part of this dissertation, I studied the graph selection consistency for model misspecification when using decomposable graphs only. By unveiling the connection between sample partial correlations and single edge selection consistency, I was able to show the selection consistency when the true graph is decomposable. Using minimal triangulation graphs as a bridge in the model misspecification case, theoretical results can be derived. By proving the equivalence between minimal triangulations of any nondecomposable graph under certain assumptions, we are able to uncover the structure of minimal triangulation space. One crucial character of HIW priors is that it does not enforce heavy penalty on false positive edges which means it also does not induce any sparsity on the decomposable graph space. Therefore, a sparse prior on the decomposable graph space or on edges is the necessary regularization to induce sparsity. Although only HIW-g Wishart prior is considered in this dissertation due to its simplicity, all theorems in Chapter 3 can be extended to any form of HIW priors with little changes in the assumptions. Simulation studies are conducted to replicate the convergence rates for model misspecification along with well-specified case. The results are consistent with the theorems I proposed. The theoretical convergence rates are almost the same as we calculated from the simulation studies.

4.2 Future Topics

There are many ways to incorporate covariate information in the graph. The way I proposed in this dissertation is to let only the covariates affect the mean structure of the Gaussian graphical models. The covariance structure of Gaussian distributions can also be affected by the covariates. Factor models are designed to accomplish this goal in a parametric way. But a fundamental limitation is that it is not very flexible and it is easy to mis-specify the underlying true structure. Future work can be done by discovering nonparametric methods to achieve this such as partition methods, for example, classification and regression trees. The nonparametric model is flexible in its nature and can deal with model misspecification. Variable selection consistency can be studied as well, but more advanced tools are needed to solve the proof of corresponding consistency theorems. Furthermore, I only used HIW prior for the covariance matrix. It is not hard to extend this to G-Wishart priors in the future. One crucial bottle neck is that the normalizing constant of G-Wishart distributions does not have a analytic form. This fact alone causes problems in marginalization and computation.

For graph selection consistency, although I studied the property of HIW priors thoroughly, it is impossible to apply the techniques presented here onto the G-Wishart priors. The non-analytic form of the normalizing constants of the G-Wishart priors creates a challenging problem to the current proving techniques. Besides this, there are more scenarios to be considered when the graph travels outside the decomposable space. The enumerations become extremely complicated due to the loss of decomposability. The above two reasons are the main causes why developing a result of selection consistency is nearly impossible for the G-Wishart priors.

Another way to study the selection consistency for all undirected graphs is to abandon the existing framework. Future research can focus on using a pseudo-likelihood functions for multivariate Gaussian distributions. The graph selection consistency can be transformed into a problem related to the traditional variable selection consistency. Pseudo-likelihood functions also enjoy some good properties, such as more flexible and computationally efficient. Other current existing Bayesian shrinkage methods can also be incorporated into the pseudo-likelihood framework. Since pseudo-likelihood functions are approximations to Gaussian likelihood functions, model misspecification must be studied comprehensively.

REFERENCES

[1] S. L. Lauritzen, Graphical models, vol. 17. Clarendon Press, 1996.

[2] P. Diaconis and D. Ylvisaker, "Conjugate priors for exponential families," The Annals of statistics, pp. 269–281, 1979.

[3] A. P. Dawid and S. L. Lauritzen, "Hyper markov laws in the statistical analysis of decomposable graphical models," The Annals of Statistics, pp. 1272–1317, 1993.

[4] A. Roverato, "Cholesky decomposition of a hyper inverse wishart matrix," Biometrika, vol. 87, no. 1, pp. 99–112, 2000.

[5] G. Letac, H. Massam, et al., "Wishart distributions for decomposable graphs," The Annals of Statistics, vol. 35, no. 3, pp. 1278–1323, 2007.

[6] B. Rajaratnam, H. Massam, C. M. Carvalho, et al., "Flexible covariance estimation in graphical gaussian models," The Annals of Statistics, vol. 36, no. 6, pp. 2818–2849, 2008.

[7] A. Roverato, "Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models," Scandinavian Journal of Statistics, vol. 29, no. 3, pp. 391–411, 2002.

[8] A. Atay-Kayis and H. Massam, "A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models," Biometrika, vol. 92, no. 2, pp. 317–335, 2005.

[9] P. Dellaportas, P. Giudici, and G. Roberts, "Bayesian inference for nondecomposable graphical gaussian models," Sankhyā: The Indian Journal of Statistics, pp. 43–55, 2003.

[10] B. Moghaddam, E. Khan, K. P. Murphy, and B. M. Marlin, "Accelerating bayesian structural inference for non-decomposable gaussian graphical models," in Advances in Neural Information Processing Systems, pp. 1285–1293, 2009.

[11] H. Wang, C. M. Carvalho, et al., “Simulation of hyper-inverse wishart distributions for non-decomposable graphs,” Electronic Journal of Statistics, vol. 4, pp. 1470–1475, 2010.

[12] K. Khare, B. Rajaratnam, and A. Saha, “Bayesian inference for gaussian graphical models beyond decomposable graphs,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 80, no. 4, pp. 727–747, 2018.

[13] J. G. Scott and C. M. Carvalho, “Feature-inclusion stochastic search for gaussian graphical models,” Journal of Computational and Graphical Statistics, vol. 17, no. 4, pp. 790–808, 2008.

[14] P. Giudici, “Learning in graphical gaussian models,” Bayesian Statistics, vol. 5, pp. 621–628, 1996.

[15] P. Giudici, Green, and PJ, “Decomposable graphical gaussian model determination,” Biometrika, vol. 86, no. 4, pp. 785–801, 1999.

[16] C. M. Carvalho, H. Massam, and M. West, “Simulation of hyper-inverse wishart distributions in graphical models,” Biometrika, vol. 94, no. 3, pp. 647–659, 2007.

[17] C. M. Carvalho and J. G. Scott, “Objective bayesian model selection in gaussian graphical models,” Biometrika, vol. 96, no. 3, pp. 497–512, 2009.

[18] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West, “Experiments in stochastic computation for high-dimensional graphical models,” Statistical Science, pp. 388–400, 2005.

[19] S. Donnet and J.-M. Marin, “An empirical bayes procedure for the selection of gaussian graphical models,” Statistics and Computing, vol. 22, no. 5, pp. 1113–1123, 2012.

[20] G. Raskutti, B. Yu, M. J. Wainwright, and P. K. Ravikumar, “Model selection in gaussian graphical models: High-dimensional consistency of lregularized mle,” in Advances in Neural Information Processing Systems, pp. 1329–1336, 2009.

[21] N. Meinshausen, P. Bühlmann, et al., “High-dimensional graphs and variable selection with the lasso,” The annals of statistics, vol. 34, no. 3, pp. 1436–1462, 2006.

[22] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," Biometrika, vol. 94, no. 1, pp. 19–35, 2007.

[23] M. Drton, M. D. Perlman, et al., "Multiple testing and error control in gaussian graphical model selection," Statistical Science, vol. 22, no. 3, pp. 430–449, 2007.

[24] P. Bickel and E. Levina, "Regularized estimation of large covariance matrices," The Annals of Statistics, vol. 36, no. 1, pp. 199–227, 2008.

[25] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," Annals of statistics, vol. 37, no. 6B, p. 4254, 2009.

[26] N. El Karoui, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," The Annals of Statistics, vol. 36, no. 6, pp. 2717–2756, 2008.

[27] T. Cai and W. Liu, "Adaptive thresholding for sparse covariance matrix estimation," Journal of the American Statistical Association, vol. 106, no. 494, pp. 672–684, 2011.

[28] S. Banerjee, S. Ghosal, et al., "Posterior convergence rates for estimating large precision matrices using graphical models," Electronic Journal of Statistics, vol. 8, no. 2, pp. 2111–2137, 2014.

[29] S. Banerjee and S. Ghosal, "Bayesian structure learning in graphical models," Journal of Multivariate Analysis, vol. 136, pp. 147–162, 2015.

[30] A. M. Fitch, M. B. Jones, H. Massam, et al., "The performance of covariance selection methods that consider decomposable models only," Bayesian Analysis, vol. 9, no. 3, pp. 659–684, 2014.

[31] X. Cao, K. Khare, and M. Ghosh, "Posterior graph selection and estimation consistency for high-dimensional bayesian dag models," arXiv preprint arXiv:1611.01205, 2016.

[32] A. Zellner, "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," Journal of the American statistical Association, vol. 57, no. 298, pp. 348–368, 1962.

[33] C. Holmes, D. T. Denison, and B. Mallick, “Accounting for model uncertainty in seemingly unrelated regressions,” Journal of Computational and Graphical Statistics, vol. 11, no. 3, pp. 533–551, 2002.

[34] H. Wang, “Sparse seemingly unrelated regression modelling: Applications in finance and econometrics,” Computational Statistics & Data Analysis, vol. 54, no. 11, pp. 2866–2877, 2010.

[35] E. I. George and R. E. McCulloch, “Variable selection via gibbs sampling,” Journal of the American Statistical Association, vol. 88, no. 423, pp. 881–889, 1993.

[36] L. Kuo and B. Mallick, “Variable selection for regression models,” Sankhyā: The Indian Journal of Statistics, Series B, pp. 65–81, 1998.

[37] J. Yin and H. Li, “A sparse conditional gaussian graphical model for analysis of genetical genomics data,” The annals of applied statistics, vol. 5, no. 4, p. 2630, 2011.

[38] W. Lee and Y. Liu, “Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood,” Journal of multivariate analysis, vol. 111, pp. 241–255, 2012.

[39] T. T. Cai, H. Li, W. Liu, and J. Xie, “Covariate-adjusted precision matrix estimation with an application in genetical genomics,” Biometrika, vol. 100, no. 1, pp. 139–156, 2012.

[40] A. Bhadra and B. K. Mallick, “Joint high-dimensional bayesian variable and covariance selection with an application to eqtl analysis,” Biometrics, vol. 69, no. 2, pp. 447–457, 2013.

[41] A. Zellner, “On assessing prior distributions and bayesian regression analysis with g-prior distributions,” Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti, vol. 6, pp. 233–243, 1986.

[42] C. Fernandez, E. Ley, and M. F. Steel, “Benchmark priors for bayesian model averaging,” Journal of Econometrics, vol. 100, no. 2, pp. 381–427, 2001.

[43] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, “Mixtures of g priors for bayesian variable selection,” Journal of the American Statistical Association, vol. 103, no. 481, pp. 410–423, 2008.

[44] R. E. Kass and L. Wasserman, “A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion,” Journal of the american statistical association, vol. 90, no. 431, pp. 928–934, 1995.

[45] D. P. Foster and E. I. George, “The risk inflation criterion for multiple regression,” The Annals of Statistics, pp. 1947–1975, 1994.

[46] D. Denison, B. Mallick, and A. Smith, “Automatic bayesian curve fitting,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 60, no. 2, pp. 333–350, 1998.

[47] R. Akbani, P. K. S. Ng, H. M. Werner, M. Shahmoradgoli, F. Zhang, Z. Ju, W. Liu, J.-Y. Yang, K. Yoshihara, J. Li, et al., “A pan-cancer proteomic perspective on the cancer genome atlas,” Nature communications, vol. 5, p. 3887, 2014.

[48] M. Ha, S. Banerjee, R. Akbani, H. Liang, G. Mills, K. Do, and V. Baladandayuthapani, “Personalized cancer-specific integrated network estimation,” Submitted, 2016.

[49] C. Crainiceanu, D. Ruppert, and M. P. Wand, “Bayesian analysis for penalized spline regression using winbugs,” 2005.

[50] Y. Tsujimoto, “Role of bcl-2 family proteins in apoptosis: apoptosomes or mitochondria?,” Genes to cells, vol. 3, no. 11, pp. 697–707, 1998.

[51] D. Etemadmoghadam, B. A. Weir, G. Au-Yeung, K. Alsop, G. Mitchell, J. George, S. Davis, A. D. D’Andrea, K. Simpson, W. C. Hahn, et al., “Synthetic lethality between ccne1 amplification and loss of brca1,” Proceedings of the National Academy of Sciences, vol. 110, no. 48, pp. 19489–19494, 2013.

[52] K. Nakayama, M. T. Rahman, M. Rahman, K. Nakamura, M. Ishikawa, H. Katagiri, E. Sato, T. Ishibashi, K. Iida, N. Ishikawa, et al., “Ccne1 amplification is associated with aggres-

sive potential in endometrioid endometrial carcinomas," International journal of oncology, vol. 48, no. 2, pp. 506–516, 2016.

[53] E. Cocco, S. Lopez, J. Black, S. Bellone, E. Bonazzoli, F. Predolini, F. Ferrari, C. L. Schwab, G. Menderes, L. Zammataro, et al., "Dual ccne1/pik3ca targeting is synergistic in ccne1-amplified/pik3ca-mutated uterine serous carcinomas in vitro and in vivo," British journal of cancer, 2016.

[54] K. A. Schrader, S. Masciari, N. Boyd, S. Wiyrick, P. Kaurah, J. Senz, W. Burke, H. T. Lynch, J. E. Garber, and D. G. Huntsman, "Hereditary diffuse gastric cancer: association with lobular breast cancer," Familial cancer, vol. 7, no. 1, pp. 73–82, 2008.

[55] K. Polyak and R. A. Weinberg, "Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits," Nature Reviews Cancer, vol. 9, no. 4, pp. 265–273, 2009.

[56] M. C. Bertucci and C. A. Mitchell, "Phosphoinositide 3-kinase and inpp4b in human breast cancer," Annals of the New York Academy of Sciences, vol. 1280, no. 1, pp. 1–5, 2013.

[57] P. F. McAuliffe, F. Meric-Bernstam, G. B. Mills, and A. M. Gonzalez-Angulo, "Deciphering the role of pi3k/akt/mtor pathway in breast cancer biology and pathogenesis," Clinical breast cancer, vol. 10, pp. S59–S65, 2010.

[58] T. W. Miller, B. N. Rexer, J. T. Garrett, and C. L. Arteaga, "Mutations in the phosphatidylinositol 3-kinase pathway: role in tumor progression and therapeutic implications in breast cancer," Breast cancer research, vol. 13, no. 6, p. 224, 2011.

[59] J. A. Gasser, H. Inuzuka, A. W. Lau, W. Wei, R. Beroukhim, and A. Toker, "Sgk3 mediates inpp4b-dependent pi3k signaling in breast cancer," Molecular cell, vol. 56, no. 4, pp. 595–607, 2014.

[60] D. Harari and Y. Yarden, "Molecular mechanisms underlying erbb2/her2 action in breast cancer," Oncogene, vol. 19, no. 53, p. 6102, 2000.

[61] F. Revillion, J. Bonneterre, and J. Peyrat, “Erbb2 oncogene in human breast cancer and its clinical significance,” European Journal of Cancer, vol. 34, no. 6, pp. 791–808, 1998.

[62] O.-P. Kallioniemi, A. Kallioniemi, W. Kurisu, A. Thor, L.-C. Chen, H. S. Smith, F. M. Waldman, D. Pinkel, and J. W. Gray, “Erbb2 amplification in breast cancer analyzed by fluorescence in situ hybridization.,” Proceedings of the National Academy of Sciences, vol. 89, no. 12, pp. 5321–5325, 1992.

[63] A. Dydensborg, A. Rose, B. Wilson, D. Grote, M. Paquet, V. Giguere, P. Siegel, and M. Bouchard, “Gata3 inhibits breast cancer growth and pulmonary breast cancer metastasis,” Oncogene, vol. 28, no. 29, pp. 2634–2642, 2009.

[64] W. Yan, Q. J. Cao, R. B. Arenas, B. Bentley, and R. Shao, “Gata3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition,” Journal of Biological Chemistry, vol. 285, no. 18, pp. 14042–14051, 2010.

[65] E. Ben-David, T. Li, H. Massam, and B. Rajaratnam, “High dimensional bayesian inference for gaussian directed acyclic graph models,” arXiv preprint arXiv:1109.4371, 2011.

[66] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, “Sparse graphical models for exploring gene expression data,” Journal of Multivariate Analysis, vol. 90, no. 1, pp. 196–212, 2004.

[67] P. Heggernes, “Minimal triangulations of graphs: A survey,” Discrete Mathematics, vol. 306, no. 3, pp. 297–317, 2006.

[68] D. J. Rose, R. E. Tarjan, and G. S. Lueker, “Algorithmic aspects of vertex elimination on graphs,” SIAM Journal on computing, vol. 5, no. 2, pp. 266–283, 1976.

[69] I. M. Johnstone, “High dimensional bernstein-von mises: simple examples,” Institute of Mathematical Statistics collections, vol. 6, p. 87, 2010.

[70] V. Spokoiny, “Bernstein-von mises theorem for growing parameter dimension,” arXiv preprint arXiv:1302.3430, 2013.

[71] C. Uhler, A. Lenkoski, D. Richards, et al., "Exact formulas for the normalizing constants of wishart distributions for graphical models," The Annals of Statistics, vol. 46, no. 1, pp. 90–118, 2018.

[72] Y. Niu, D. Pati, and B. Mallick, "Bayesian graph selection consistency for decomposable graphs," arXiv preprint arXiv:1901.04134, 2019.

[73] M. Singull and T. Koski, "On the distribution of matrix quadratic forms," Communications in Statistics-Theory and Methods, vol. 41, no. 18, pp. 3403–3415, 2012.

[74] T. W. Anderson, An introduction to multivariate statistical analysis. Wiley, 1984.

[75] H. Hotelling, "New light on the correlation coefficient and its transforms," Journal of the Royal Statistical Society. Series B (Methodological), vol. 15, no. 2, pp. 193–232, 1953.

[76] G. Watson, "A note on gamma functions," Edinburgh Mathematical Notes, vol. 42, pp. 7–9, 1959.

[77] C. Mortici, "New approximation formulas for evaluating the ratio of gamma functions," Mathematical and Computer Modelling, vol. 52, no. 1-2, pp. 425–433, 2010.

[78] J. Segura, "Sharp bounds for cumulative distribution functions," Journal of Mathematical Analysis and Applications, vol. 436, no. 2, pp. 748–763, 2016.

[79] M. Frydenberg and L. L. STEFFEN, "Decomposition of maximum likelihood in mixed graphical interaction models," Biometrika, vol. 76, no. 3, pp. 539–555, 1989.

[80] A. Thomas and P. J. Green, "Enumerating the decomposable neighbors of a decomposable graph under a simple perturbation scheme," Computational statistics & data analysis, vol. 53, no. 4, pp. 1232–1238, 2009.

[81] P. J. Green and A. Thomas, "Sampling decomposable graphs using a markov chain on junction trees," Biometrika, vol. 100, no. 1, pp. 91–110, 2013.

APPENDIX A

MARGINAL DISTRIBUTION OF Y GIVEN γ AND G

In this appendix, we provide the detail calculation for the conditionally marginal density of Y given only $\boldsymbol{\gamma}$ and graph G. Given the hierarchical model in Section 2.2,

$$\begin{aligned}(\mathrm{Y}-\mathrm{U}_{\gamma}\mathrm{B}_{\gamma,G})|\boldsymbol{\gamma},\Sigma_G &\sim \mathrm{MN}_{n\times q}(0, I_n, \Sigma_G),\\ \mathrm{B}_{\gamma,G}|\boldsymbol{\gamma},\Sigma_G &\sim \mathrm{MN}_{p_{\gamma}(k+1)\times q}(0, g(\mathrm{U}_{\gamma}^T\mathrm{U}_{\gamma})^{-1}_{p_{\gamma}(k+1)}, \Sigma_G),\\ \Sigma_G|G &\sim \mathrm{HIW}_G(b, dI_q),\end{aligned}$$

where Y is $n\times q$, U_{γ} is $n\times p_{\gamma}(k+1)$, $\mathrm{B}_{\gamma,G}$ is $p_{\gamma}(k+1)\times q$, Σ_G is $q\times q$. First, we can marginalize out the coefficient matrix $\mathrm{B}_{\gamma,G}$ due to the conjugacy of its prior to the likelihood of Y. We have

$$\mathrm{Y}|\boldsymbol{\gamma},\Sigma_G \sim \mathrm{MN}_{n\times q}(0, I_n + g\mathrm{U}_{\gamma}(\mathrm{U}_{\gamma}^T\mathrm{U}_{\gamma})^{-1}\mathrm{U}_{\gamma}^T, \Sigma_G).$$

Next, we vectorize Y to prepare for integrating out Σ_G. So,

$$vec(\mathrm{Y}^T)|\boldsymbol{\gamma},\Sigma_G \sim N_{nq}(0, (I_n + gP_{\gamma})\otimes\Sigma_G), \tag{A.1}$$

where $P_{\gamma} = \mathrm{U}_{\gamma}(\mathrm{U}_{\gamma}^T\mathrm{U}_{\gamma})^{-1}\mathrm{U}_{\gamma}^T$ and $\otimes$ is the Kronecker product operation.

We use the Sylvester's determinant theorem to further simplify the density of $vec(\mathrm{Y}^T)$. If A and B are matrices of size $m\times n$ and $n\times m$ respectively, then $|I_m + AB| = |I_n + BA|$. We have

$$|I_n+gP_{\gamma}| = |I_n+g\mathrm{U}_{\gamma}(\mathrm{U}_{\gamma}^T\mathrm{U}_{\gamma})^{-1}\mathrm{U}_{\gamma}^T| = |I_{p_{\gamma}(k+1)}+g(\mathrm{U}_{\gamma}^T\mathrm{U}_{\gamma})^{-1}(\mathrm{U}_{\gamma}^T\mathrm{U}_{\gamma})| = (g+1)^{p_{\gamma}(k+1)}. \tag{A.2}$$

By the Sherman-Morrison-Woodbury (SMW) identity, assuming A, C and $(C^{-1}+DA^{-1}B)$ to be nonsingular,

$$(A+BCD)^{-1}=A^{-1}-A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1},$$

we have

$$(I_n+gP_{\boldsymbol\gamma})^{-1}=\left\{I_n+g\mathrm{U}_{\boldsymbol\gamma}(\mathrm{U}_{\boldsymbol\gamma}^T\mathrm{U}_{\boldsymbol\gamma})^{-1}\mathrm{U}_{\boldsymbol\gamma}^T\right\}^{-1}=I_n-\frac{g}{g+1}P_{\boldsymbol\gamma}. \tag{A.3}$$

Simplifying the density of (A.1) using (A.2) and (A.3),

$$\begin{aligned}
f(vec(\mathrm{Y}^T)|\boldsymbol\gamma,\Sigma_G)&=(2\pi)^{-\frac{nq}{2}}|I_n+gP_{\boldsymbol\gamma}|^{-\frac{g}{2}}|\Sigma_G|^{-\frac{n}{2}}\\
&\qquad exp\Big[-\frac{1}{2}\{vec(\mathrm{Y}^T)\}^T\Big\{(I_n+gP_{\boldsymbol\gamma})^{-1}\otimes{\Sigma_G}^{-1}\Big\}\{vec(\mathrm{Y}^T)\}\Big]\\
&=(2\pi)^{-\frac{nq}{2}}(g+1)^{-\frac{p_{\boldsymbol\gamma}(k+1)q}{2}}|\Sigma_G|^{-\frac{n}{2}}\\
&\qquad exp\Big[-\frac{1}{2}\{vec(\mathrm{Y}^T)\}^T\Big\{\Big(I_n-\frac{g}{g+1}P_{\boldsymbol\gamma}\Big)\otimes{\Sigma_G}^{-1}\Big\}\{vec(\mathrm{Y}^T)\}\Big].
\end{aligned}$$

Matrix vectorization and trace operation have the following relationship. Suppose that A is an $r\times s$ matrix and B is $s\times r$, then $tr(AB)=\{vec(A)\}^Tvec(B^T)=\{vec(A^T)\}^Tvec(B)$. So we can further reduce the complexity of the exponential term in the density above.

$$\begin{aligned}
&\{vec(\mathrm{Y}^T)\}^T\Big\{\Big(I_n-\frac{g}{g+1}P_{\boldsymbol\gamma}\Big)\otimes{\Sigma_G}^{-1}\Big\}\{vec(\mathrm{Y}^T)\}\\
&=\{vec(\mathrm{Y}^T)\}^T\cdot\Big\{\Big(I_n-\frac{g}{g+1}P_{\boldsymbol\gamma}\Big)\otimes{\Sigma_G}^{-1}\cdot vec(\mathrm{Y}^T)\Big\}\\
&=\{vec(\mathrm{Y}^T)\}^T\cdot vec\Big\{\Sigma_G^{-1}\mathrm{Y}^T\Big(I_n-\frac{g}{g+1}P_{\boldsymbol\gamma}\Big)\Big\}\\
&=tr\Big\{\mathrm{Y}^T\Big(I_n-\frac{g}{g+1}P_{\boldsymbol\gamma}\Big)\mathrm{Y}\Sigma_G^{-1}\Big\}\\
&=tr\{S(\boldsymbol\gamma)\Sigma_G^{-1}\},
\end{aligned}$$

where $S(\boldsymbol\gamma)=\mathrm{Y}^T\big(I_n-\frac{g}{g+1}P_{\boldsymbol\gamma}\big)\mathrm{Y}$. Eventually, by factorizing the density $f(\mathrm{Y}|\boldsymbol\gamma,\Sigma_G)$ corresponding to the hyper-inverse Wishart prior, we have

$$\begin{aligned}
f(\mathrm{Y}|\boldsymbol{\gamma}, \Sigma_G) &= (2\pi)^{-\frac{nq}{2}}(g+1)^{-\frac{p_{\boldsymbol{\gamma}}(k+1)q}{2}}|\Sigma_G|^{-\frac{n}{2}} etr\Big\{-\frac{1}{2}S(\boldsymbol{\gamma})\Sigma_G^{-1}\Big\} \\
&= (2\pi)^{-\frac{nq}{2}}(g+1)^{-\frac{p_{\boldsymbol{\gamma}}(k+1)q}{2}} \frac{\prod_{C\in\mathscr{C}}|\Sigma_C|^{-\frac{n}{2}} etr\Big\{-\frac{1}{2}S_C(\boldsymbol{\gamma})\Sigma_C^{-1}\Big\}}{\prod_{S\in\mathscr{S}}|\Sigma_S|^{-\frac{n}{2}} etr\Big\{-\frac{1}{2}S_S(\boldsymbol{\gamma})\Sigma_S^{-1}\Big\}}. \qquad \text{(A.4)}
\end{aligned}$$

Since $\Sigma_G|G \sim \mathrm{HIW}_G(b, dI_q)$, then

$$\begin{aligned}
f(\Sigma_G|G) &= \frac{\prod_{C\in\mathscr{C}} \frac{|dI_C|^{\frac{b+|C|-1}{2}}}{2^{\frac{(b+|C|-1)|C|}{2}}\Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right)}|\Sigma_C|^{-(\frac{b}{2}+|C|)} etr(-\frac{1}{2}dI_C\Sigma_C^{-1})}{\prod_{S\in\mathscr{S}} \frac{|dI_S|^{\frac{b+|S|-1}{2}}}{2^{\frac{(b+|S|-1)|S|}{2}}\Gamma_{|S|}\left(\frac{b+|S|-1}{2}\right)}|\Sigma_S|^{-(\frac{b}{2}+|S|)} etr(-\frac{1}{2}dI_S\Sigma_S^{-1})} \\
&= \mathscr{H}(b, dI_q, G)\cdot\frac{\prod_{C\in\mathscr{C}}|\Sigma_C|^{-(\frac{b}{2}+|C|)} etr(-\frac{1}{2}dI_C\Sigma_C^{-1})}{\prod_{S\in\mathscr{S}}|\Sigma_S|^{-(\frac{b}{2}+|S|)} etr(-\frac{1}{2}dI_S\Sigma_S^{-1})}, \qquad \text{(A.5)}
\end{aligned}$$

where

$$\mathscr{H}(b, dI_q, G) = \frac{\prod_{C\in\mathscr{C}} \frac{|dI_C|^{\frac{b+|C|-1}{2}}}{2^{\frac{(b+|C|-1)|C|}{2}}\Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right)}}{\prod_{S\in\mathscr{S}} \frac{|dI_S|^{\frac{b+|S|-1}{2}}}{2^{\frac{(b+|S|-1)|S|}{2}}\Gamma_{|S|}\left(\frac{b+|S|-1}{2}\right)}}.$$

Next, we integrate out Σ_G by (A.4) and (A.5),

$$\begin{aligned}
&f(\mathrm{Y}|\boldsymbol{\gamma}, G) = \int f(\mathrm{Y}|\boldsymbol{\gamma}, \Sigma_G) f(\Sigma_G|G) d\Sigma_G \\
&= (2\pi)^{-\frac{nq}{2}}(g+1)^{-\frac{p_{\boldsymbol{\gamma}}(k+1)q}{2}}\mathscr{H}(b, dI_q, G)\int\frac{\prod_{C\in\mathscr{C}}|\Sigma_C|^{-(\frac{b+n}{2}+|C|)} etr[-\frac{1}{2}(dI_C+S_C(\gamma))\Sigma_C^{-1}]}{\prod_{S\in\mathscr{S}}|\Sigma_S|^{-(\frac{b+n}{2}+|S|)} etr[-\frac{1}{2}(dI_S+S_S(\gamma))\Sigma_S^{-1}]}d\Sigma_G \\
&= (2\pi)^{-\frac{nq}{2}}(g+1)^{-\frac{p_{\boldsymbol{\gamma}}(k+1)q}{2}}\mathscr{H}(b, dI_q, G)\mathscr{H}^{-1}(b+n, dI_q+S(\boldsymbol{\gamma}), G) \\
&= (2\pi)^{-\frac{nq}{2}}(g+1)^{-\frac{p_{\boldsymbol{\gamma}}(k+1)q}{2}}\mathscr{H}(b, dI_q, G)\frac{\prod_{C\in\mathscr{C}} \frac{|dI_C+S_C(\boldsymbol{\gamma})|^{-\frac{b+n+|C|-1}{2}}}{2^{-\frac{(b+n+|C|-1)|C|}{2}}\Gamma^{-1}_{|C|}\left(\frac{b+n+|C|-1}{2}\right)}}{\prod_{S\in\mathscr{S}} \frac{|dI_S+S_S(\boldsymbol{\gamma})|^{-\frac{b+n+|S|-1}{2}}}{2^{-\frac{(b+n+|S|-1)|S|}{2}}\Gamma^{-1}_{|S|}\left(\frac{b+n+|S|-1}{2}\right)}} \\
&= M_{n,G}\times(g+1)^{-\frac{p_{\boldsymbol{\gamma}}(k+1)q}{2}}\frac{\prod_{C\in\mathscr{C}}|dI_C+S_C(\boldsymbol{\gamma})|^{-\frac{b+n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}|dI_S+S_S(\boldsymbol{\gamma})|^{-\frac{b+n+|S|-1}{2}}}, \qquad \text{(A.6)}
\end{aligned}$$

where

$$M_{n,G} = (2\pi)^{-\frac{nq}{2}} \frac{\prod_{C \in \mathscr{C}} \frac{|dI_C|^{\frac{b+|C|-1}{2}}}{2^{-\frac{n|C|}{2}} \Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right) \Gamma_{|C|}^{-1}\left(\frac{b+n+|C|-1}{2}\right)}}{\prod_{S \in \mathscr{S}} \frac{|dI_S|^{\frac{b+|S|-1}{2}}}{2^{-\frac{n|S|}{2}} \Gamma_{|S|}\left(\frac{b+|S|-1}{2}\right) \Gamma_{|S|}^{-1}\left(\frac{b+n+|S|-1}{2}\right)}}$$

is a constant which depends only on n and G, but it is the same for all $\boldsymbol{\gamma}$ under the same graph G. This makes possible for the cancellation in Metroplis-Hasting step for variable selection, which leads to faster in computation. $S_C(\boldsymbol{\gamma})$ and $S_S(\boldsymbol{\gamma})$ are the corresponding quadratic form similar to $S(\boldsymbol{\gamma})$ but restricted to sub-graphs denoted by C and S (i.e. cliques and separators).

APPENDIX B

THE PROOF OF VARIABLE SELECTION CONSISTENCY

We now show the complete proof of Lemma 2.4.1 in this appendix. The main idea is, if the alternative model $\boldsymbol{a}$ does not contain the true model, the likelihood part drives the Bayes factor to zero exponentially, as $\boldsymbol{a}$ cannot fit the true mean adequately. If $\boldsymbol{a}$ contains true model then the difference in the likelihood becomes negligible but the prior penalizes for the extra dimensions and the Bayes factor goes to zero as n goes to infinity.

B.1 Preparation

By (A.6), the Bayes factor in favor of alternative model $\boldsymbol{a}$ under any graph G is

$$\begin{aligned} BF(\boldsymbol{a};\boldsymbol{t}|G) &= \frac{f(\mathrm{Y}|\boldsymbol{a},G)}{f(\mathrm{Y}|\boldsymbol{t},G)} \\ &= (g+1)^{-\frac{(p_{\boldsymbol{a}}-p_{\boldsymbol{t}})(k+1)q}{2}} \times \frac{\prod_{C\in\mathscr{C}}\left(\frac{|dI_C+S_C(\boldsymbol{a})|}{|dI_C+S_C(\boldsymbol{t})|}\right)^{-\frac{b+n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\left(\frac{|dI_S+S_S(\boldsymbol{a})|}{|dI_S+S_S(\boldsymbol{t})|}\right)^{-\frac{b+n+|S|-1}{2}}} \\ &:= (g+1)^{-\frac{(p_{\boldsymbol{a}}-p_{\boldsymbol{t}})(k+1)q}{2}} \times \frac{\prod_{C\in\mathscr{C}}\left\{\Delta_C(\boldsymbol{a},\boldsymbol{t})\right\}^{-\frac{b+n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\left\{\Delta_S(\boldsymbol{a},\boldsymbol{t})\right\}^{-\frac{b+n+|S|-1}{2}}} \\ &:= \mathrm{I}\times\mathrm{II}(\boldsymbol{a},\boldsymbol{t}), \end{aligned} \tag{B.1}$$

where $S(\boldsymbol{a}) = \mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}}\big)\mathrm{Y}$, $S(\boldsymbol{t}) = \mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{t}}\big)Y$, and S_C and S_S denote the quadratic forms restricted to clique $C \in \mathscr{C}$ and separator $S \in \mathscr{S}$; furthermore, we denote $\Delta_C(\boldsymbol{a},\boldsymbol{t}) = \frac{|dI_C+S_C(\boldsymbol{a})|}{|dI_C+S_C(\boldsymbol{t})|}$ and $\Delta_S(\boldsymbol{a},\boldsymbol{t}) = \frac{|dI_S+S_S(\boldsymbol{a})|}{|dI_S+S_S(\boldsymbol{t})|}$. Let $\Delta(\boldsymbol{a},\boldsymbol{t}) = \frac{|dI_q+S(\boldsymbol{a})|}{|dI_q+S(\boldsymbol{t})|}$ be the version of $\Delta(\boldsymbol{a},\boldsymbol{t})$ for the whole graph G.

Lemma B.1.1. *$\Delta(\boldsymbol{a},\boldsymbol{t}) = \left|I_q + \frac{1}{d}\frac{g}{g+1}\left(\frac{1}{n}A_{\boldsymbol{t}}\right)^{-\frac{1}{2}}\left\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{t}} - P_{\boldsymbol{a}}\big)\mathrm{Y}\right\}\left(\frac{1}{n}A_{\boldsymbol{t}}\right)^{-\frac{1}{2}}\right|$, where $A_{\boldsymbol{t}} = I_q + \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{t}}\big)Y$.*

Proof.

$$\begin{aligned}
\Delta(\boldsymbol{a},\boldsymbol{t}) &= \frac{|dI_q + S(\boldsymbol{a})|}{|dI_q + S(\boldsymbol{t})|} = \frac{|I_q + \frac{1}{d}S(\boldsymbol{a})|}{|I_q + \frac{1}{d}S(\boldsymbol{t})|} \\
&= \frac{|I_q + \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}}\big)\mathrm{Y}|}{|I_q + \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{t}}\big)\mathrm{Y}|} \\
&= \frac{|I_q + \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{t}}\big)\mathrm{Y} + \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}}\big)\mathrm{Y} - \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{t}}\big)\mathrm{Y}|}{|I_q + \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{t}}\big)\mathrm{Y}|} \\
&= \frac{|I_q + \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{t}}\big)\mathrm{Y} + \frac{1}{d}\frac{g}{g+1}\mathrm{Y}^T\big(P_{\boldsymbol{t}} - P_{\boldsymbol{a}}\big)\mathrm{Y}|}{|I_q + \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{t}}\big)\mathrm{Y}|} \\
&= \frac{|A_{\boldsymbol{t}} + \frac{1}{d}\frac{g}{g+1}\mathrm{Y}^T\big(P_{\boldsymbol{t}} - P_{\boldsymbol{a}}\big)\mathrm{Y}|}{|A_{\boldsymbol{t}}|} \\
&= \left|I_q + \frac{1}{d}\frac{g}{g+1}A_{\boldsymbol{t}}^{-\frac{1}{2}}\mathrm{Y}^T\big(P_{\boldsymbol{t}} - P_{\boldsymbol{a}}\big)\mathrm{Y}A_{\boldsymbol{t}}^{-\frac{1}{2}}\right| \\
&= \left|I_q + \frac{1}{d}\frac{g}{g+1}\Big(\frac{1}{n}A_{\boldsymbol{t}}\Big)^{-\frac{1}{2}}\Big\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{t}} - P_{\boldsymbol{a}}\big)\mathrm{Y}\Big\}\Big(\frac{1}{n}A_{\boldsymbol{t}}\Big)^{-\frac{1}{2}}\right|.
\end{aligned}$$

□

Remark B.1.1. *Lemma B.1.1 is with respect to the whole graph G, but the same result holds for every clique C and separator S. And similarly we have $A_{\boldsymbol{t}}^C$ and $A_{\boldsymbol{t}}^S$ for clique C and separator S, respectively. For simplicity, we will not show the results for cliques and separators. In the next several lemmas, we only show the results with respect to the whole graph G, but they all hold for any subgraphs of G, i.e. cliques and separators.*

Next we split Bayes factor $\mathrm{BF}(\boldsymbol{a};\boldsymbol{t}|G)$ into two parts $\mathrm{BF}(\boldsymbol{a};\boldsymbol{a}\cup\boldsymbol{t}|G)$ and $\mathrm{BF}(\boldsymbol{a}\cup\boldsymbol{t};\boldsymbol{t}|G)$ and show them both converge to zero as $n \to \infty$. But before that, we need to introduce several lemmas.

Lemma B.1.2. *Under Condition 2.4.1, $\operatorname{p\lim}_{n\to\infty} \frac{\mathrm{Y}^T(I_n - P_{\boldsymbol{t}})\mathrm{Y}}{n} = \Sigma_{G^*}$, where Σ_{G^*} is the true covariance matrix with respect to the true graph G^*.*

Proof. Since $Y|\boldsymbol{t}, \Sigma_{G^*} \sim \mathrm{MN}_{n\times q}(\mathrm{U}_{\boldsymbol{t}}\mathrm{B}_{\boldsymbol{t},G^*}, I_n, \Sigma_{G^*})$ and $I_n - P_{\boldsymbol{t}}$ is symmetric and idempotent, by Corollary 2.1 in [73], we have $\frac{\mathrm{Y}^T(I_n - P_{\boldsymbol{t}})\mathrm{Y}}{n} \sim W_q(n - r_{\boldsymbol{t}}, \frac{1}{n}\Sigma_{G^*})$ and the non-central parameter is zero here. Let $\tilde{y}_{ij}(n), i,j = 1,\ldots,q$ denote the entries of $\frac{\mathrm{Y}^T(I_n - P_{\boldsymbol{t}})\mathrm{Y}}{n}$ and $\tilde{y}_{ij}(n) = \tilde{y}_{ji}(n)$.

Further more, let $\sigma^*_{ij}, i,j = 1,\ldots,q$ be the entries of Σ_{G^*}, and $\sigma^*_{ij} = \sigma^*_{ji}$. Since $\mathbb{E}\big\{\frac{\mathrm{Y}^T(I_n-P_{\boldsymbol{t}})\mathrm{Y}}{n}\big\} = (1-\frac{r_{\boldsymbol{t}}}{n})\Sigma_{G^*} \to \Sigma_{G^*}$, then $\mathbb{E}\big\{\tilde{y}_{ij}(n)\big\} = (1-\frac{r_{\boldsymbol{t}}}{n})\sigma^*_{ij} \to \sigma^*_{ij}$ and $Var\big\{\tilde{y}_{ij}(n)\big\} = \frac{1}{n}(\sigma^{*2}_{ij}+\sigma^*_{ii}\sigma^*_{jj}) \to 0$. Thus for any $\epsilon > 0$, there exist M_ϵ, when $n > M_\epsilon$, such that $|(1-\frac{r_{\boldsymbol{t}}}{n})\sigma^*_{ij} - \sigma^*_{ij}| < \epsilon/2$, then $Pr(|\tilde{y}_{ij}(n) - \sigma^*_{ij}| > \epsilon) \le Pr(|\tilde{y}_{ij}(n) - \mathbb{E}\{\tilde{y}_{ij}(n)\}| > \epsilon/2) \le \frac{4Var\{\tilde{y}_{ij}(n))\}}{\epsilon^2} = \frac{4(\sigma^{*2}_{ij}+\sigma^*_{ii}\sigma^*_{jj})}{n\epsilon^2} \to 0$. So $\tilde{y}_{ij}(n) \xrightarrow{p} \sigma^*_{ij}$ in probability, for all $i,j = 1,\ldots,q$. Therefore, $\frac{\mathrm{Y}^T(I_n-P_t)\mathrm{Y}}{n} \xrightarrow{p} \Sigma_{G^*}$ as $n \to \infty$. □

Lemma B.1.3. *Under Condition 2.4.1, 2.4.2, 2.4.4,* $\mathrm{p}\lim_{n\to\infty} \frac{1}{n}A_{\boldsymbol{t}} = \Sigma_{G^*}$*, where* Σ_{G^*} *is the true covariance matrix with respect to the true graph* G^**.*

Proof. First, we show $\mathrm{p}\lim_{n\to\infty} \frac{1}{n}\frac{1}{g+1}\mathrm{Y}^T P_{\boldsymbol{t}}\mathrm{Y} = \mathbf{0}_{q\times q}$. Let y_i be the ith column of Y. Note that $\lim_{n\to\infty}\mathbb{E}(\frac{1}{n}\frac{1}{g+1}\mathrm{y}_i^T P_{\boldsymbol{t}}\mathrm{y}_i) = 0$ and $\lim_{n\to\infty} Var(\frac{1}{n}\frac{1}{g+1}\mathrm{y}_i^T P_{\boldsymbol{t}}\mathrm{y}_i) = 0$, so $\frac{1}{n}\frac{1}{g+1}\mathrm{y}_i^T P_{\boldsymbol{t}}\mathrm{y}_i \to 0$ in probability. Hence, $\sum_{i=1}^q \frac{1}{n}\frac{1}{g+1}\mathrm{y}_i^T P_{\boldsymbol{t}}\mathrm{y}_i \to 0$ in probability. Therefore, the sum of eigenvalues of matrix $\frac{1}{n}\frac{1}{g+1}\mathrm{Y}^T P_{\boldsymbol{t}}\mathrm{Y}$ goes to zero in probability. Let λ_i^t be the ith eigenvalue of $\frac{1}{n}\frac{1}{g+1}\mathrm{Y}^T P_{\boldsymbol{t}}\mathrm{Y}$, $i = 1,2,\ldots,q$. So λ_i^t goes to zero in probability as the matrix is non-negative definite. Using spectral decomposition, $\frac{1}{n}\frac{1}{g+1}\mathrm{Y}^T P_{\boldsymbol{t}}\mathrm{Y} = \sum_{i=1}^q \lambda_i^t u_i u_i^T$, where u_i's are orthonormal eigenvectors, each of the entries of $\frac{1}{n}\frac{1}{g+1}\mathrm{Y}^T P_{\boldsymbol{t}}\mathrm{Y}$ goes to zero in probability and our claim follows.

Therefore,

$$\begin{aligned}
\mathrm{p}\lim_{n\to\infty}\frac{1}{n}A_{\boldsymbol{t}} &= \mathrm{p}\lim_{n\to\infty}\frac{1}{n}\Big\{I_q + \frac{1}{d}\mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{t}}\big)\mathrm{Y}\Big\}\\
&= \mathrm{p}\lim_{n\to\infty}\frac{1}{n}I_q + \mathrm{p}\lim_{n\to\infty}\frac{1}{d}\frac{\mathrm{Y}^T\big(I_n - P_{\boldsymbol{t}}\big)\mathrm{Y}}{n} + \mathrm{p}\lim_{n\to\infty}\frac{1}{d}\frac{1}{g+1}\frac{\mathrm{Y}^T P_{\boldsymbol{t}}\mathrm{Y}}{n}\\
&= \mathbf{0}_{q\times q} + \Sigma_{G^*} + \mathbf{0}_{q\times q} = \Sigma_{G^*}.
\end{aligned}$$

□

Lemma B.1.4. *Let* $\widetilde{\lambda}_i^{\boldsymbol{a}}, i = 1,\ldots,q$ *be the eigenvalues of* $\frac{1}{n}S(\boldsymbol{a})$ *and* $\widetilde{\lambda}_i^{\boldsymbol{a}\cup\boldsymbol{t}}, i = 1,\ldots,q$ *be the eigenvalues of* $\frac{1}{n}S(\boldsymbol{a}\cup\boldsymbol{t})$*, where* $S(\boldsymbol{a}) = \mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}}\big)\mathrm{Y}$ *and* $S(\boldsymbol{a}\cup\boldsymbol{t}) = \mathrm{Y}^T\big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}\cup\boldsymbol{t}}\big)\mathrm{Y}$*. Under Condition 2.4.1, 2.4.2, 2.4.4, 2.4.6,* $Pr(\widetilde{\lambda}_i^{\boldsymbol{a}} > \bar{C}) \to 0$ *and* $Pr(\widetilde{\lambda}_i^{\boldsymbol{a}\cup\boldsymbol{t}} > \bar{C}) \to 0$*,* $i = 1,\ldots,q$*, as* $n \to \infty$*, where* $\bar{C}$ *is some fixed positive constant.*

Proof. Let y_i be the ith column of Y and b_i be the ith column of $\mathrm{B}_{\boldsymbol{t},G^*}$. Then $\mathrm{v}_i := \mathrm{U}_{\boldsymbol{t}}\mathrm{b}_i$ is the ith column of $\mathrm{U}_{\boldsymbol{t}}\mathrm{B}_{\boldsymbol{t},G^*}$, $i = 1, \dots, q$. Next, we have

$$\begin{aligned} tr\Big\{\frac{S(\boldsymbol{a})}{n}\Big\} &= tr\Big\{\frac{1}{n}\mathrm{Y}^T\Big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}}\Big)\mathrm{Y}\Big\} \\ &= tr\Big\{\frac{1}{n}\mathrm{Y}^T\Big(I_n - P_{\boldsymbol{a}}\Big)\mathrm{Y}\Big\} + tr\Big\{\frac{1}{n}\frac{1}{g+1}\mathrm{Y}^T P_{\boldsymbol{a}}\mathrm{Y}\Big\} \\ &= \sum_{i=1}^{q} tr\Big\{\frac{1}{n}\mathrm{y}_i^T\Big(I_n - P_{\boldsymbol{a}}\Big)\mathrm{y}_i\Big\} + \sum_{i=1}^{q} tr\Big\{\frac{1}{n}\frac{1}{g+1}\mathrm{y}_i^T P_{\boldsymbol{a}}\mathrm{y}_i\Big\}. \end{aligned}$$

Let $r_{\boldsymbol{a}} = rank(\mathrm{U}_{\boldsymbol{a}})$. Since $\mathrm{y}_i \sim N_n(\mathrm{v}_i, I_n)$, $i = 1, \dots, q$, we have

$$\mathrm{y}_i^T\Big(I_n - P_{\boldsymbol{a}}\Big)\mathrm{y}_i \sim \chi^2_{n-r_{\boldsymbol{a}}}(\phi_i^{n-\boldsymbol{a}}),$$
$$\mathrm{y}_i^T P_{\boldsymbol{a}}\mathrm{y}_i \sim \chi^2_{r_{\boldsymbol{a}}}(\phi_i^{\boldsymbol{a}}),$$

where $\phi_i^{n-\boldsymbol{a}} = \frac{1}{2}\mathrm{v}_i^T(I_n - P_{\boldsymbol{a}})\mathrm{v}_i$, $\phi_i^{\boldsymbol{a}} = \frac{1}{2}\mathrm{v}_i^T P_{\boldsymbol{a}}\mathrm{v}_i$. By Condition 2.4.2, we have $\frac{1}{n}\phi_i^{n-\boldsymbol{a}} = \frac{1}{2n}\mathrm{v}_i^T(I_n - P_{\boldsymbol{a}})\mathrm{v}_i \leq \frac{1}{n}\mathrm{v}_i^T\mathrm{v}_i = \mathrm{b}_i^T\frac{\mathrm{U}_{\boldsymbol{t}}^T\mathrm{U}_{\boldsymbol{t}}}{n}\mathrm{b}_i \leq \frac{\lambda_{max}}{n}\|\mathrm{b}_i\|_2^2 < b_M d_{\mathrm{U}}$, where $b_M = max\{\|\mathrm{b}_i\|_2^2, i = 1, \dots, q\} < \infty$. Similarly, $\phi_i^{\boldsymbol{a}} = \frac{1}{2}\mathrm{v}_i^T P_{\boldsymbol{a}}\mathrm{v}_i \leq b_M d_{\mathrm{U}}$. Next,

$$\begin{aligned} \mathbb{E}\Big[\frac{1}{n}\mathrm{y}_i^T\Big(I_n - P_{\boldsymbol{a}}\Big)\mathrm{y}_i\Big] &= \frac{1}{n}(n - r_{\boldsymbol{a}} + \phi_i^{n-\boldsymbol{a}}) \leq 1 + \frac{1}{n}\phi_i^{n-\boldsymbol{a}} < 1 + b_M d_{\mathrm{U}}, \\ Var\Big[\frac{1}{n}\mathrm{y}_i^T\Big(I_n - P_{\boldsymbol{a}}\Big)\mathrm{y}_i\Big] &= \frac{1}{n^2}(2n - 2r_{\boldsymbol{a}} + 4\phi_i^{n-\boldsymbol{a}}) \leq \frac{1}{n}\Big(2 + \frac{4}{n}\phi_i^{n-\boldsymbol{a}}\Big) < \frac{1}{n}\Big(2 + 4b_M d_{\mathrm{U}}\Big) \to 0. \end{aligned}$$

Analogously, by Condition 2.4.4 and 2.4.6,

$$\begin{aligned} \mathbb{E}\Big[\frac{1}{g+1}\frac{1}{n}\mathrm{y}_i^T P_{\boldsymbol{a}}\mathrm{y}_i\Big] &= \frac{1}{g+1}\frac{1}{n}(r_{\boldsymbol{a}} + \phi_i^{\boldsymbol{a}}) < \frac{1}{g+1}b_M d_{\mathrm{U}} \to 0, \\ Var\Big[\frac{1}{g+1}\frac{1}{n}\mathrm{y}_i^T P_{\boldsymbol{a}}\mathrm{y}_i\Big] &= \frac{1}{(g+1)^2}\frac{1}{n^2}(2r_{\boldsymbol{a}} + 4\phi_i^{\boldsymbol{a}}) < \frac{1}{(g+1)^2}\frac{1}{n}\Big(2 + 4b_M d_{\mathrm{U}}\Big) \to 0, \end{aligned}$$

So for any $\bar{\epsilon} > 0$, we have $Pr\Big\{\frac{1}{n}\mathrm{y}_i^T\big(I_n - P_{\boldsymbol{a}}\big)\mathrm{y}_i > 1 + b_M d_{\mathrm{U}} + \bar{\epsilon}\Big\} \to 0$ and $Pr\Big\{\frac{1}{g+1}\frac{1}{n}\mathrm{y}_i^T P_{\boldsymbol{a}}\mathrm{y}_i > \bar{\epsilon}\Big\} \to 0$, $i = 1. \dots, q$, as $n \to 0$. By combining the two results together,

$$Pr\Big\{\frac{1}{n}\mathrm{y}_i^T\Big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}}\Big)\mathrm{y}_i > 1 + b_M d_{\mathrm{U}} + 2\bar{\epsilon}\Big\}$$

$$
\begin{aligned}
=&Pr\Big\{\frac{1}{n}\mathrm{y}_i^T\Big(I_n - P_{\boldsymbol{a}}\Big)\mathrm{y}_i + \frac{1}{g+1}\frac{1}{n}\mathrm{y}_i^T P_{\boldsymbol{a}}\mathrm{y}_i > 1 + b_M d_{\mathrm{U}} + \bar{\epsilon} + \bar{\epsilon}\Big\}\\
\leq&Pr\Big\{\frac{1}{n}\mathrm{y}_i^T\big(I_n - P_{\boldsymbol{a}}\big)\mathrm{y}_i > 1 + b_M d_{\mathrm{U}} + \bar{\epsilon}\Big\} + Pr\Big\{\frac{1}{g+1}\frac{1}{n}\mathrm{y}_i^T P_{\boldsymbol{a}}\mathrm{y}_i > \bar{\epsilon}\Big\} \to 0.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&Pr\Big\{\widetilde{\lambda}_i^{\boldsymbol{a}} > q(1 + b_M d_{\mathrm{U}} + 2\bar{\epsilon})\Big\} \leq Pr\Big\{\sum_{i=1}^{q}\widetilde{\lambda}_i^{\boldsymbol{a}} > q(1 + b_M d_{\mathrm{U}} + 2\bar{\epsilon})\Big\}\\
=&Pr\Big\{tr\Big(\frac{S(\boldsymbol{a})}{n}\Big) > q(1 + b_M d_{\mathrm{U}} + 2\bar{\epsilon})\Big\}\\
=&Pr\Big\{\sum_{i=1}^{q}\frac{1}{n}\mathrm{y}_i^T\Big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}}\Big)\mathrm{y}_i > q(1 + b_M d_{\mathrm{U}} + 2\bar{\epsilon})\Big\}\\
\leq&Pr\Big\{\cup_{i=1}^{q}\Big(\frac{1}{n}\mathrm{y}_i^T\Big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}}\Big)\mathrm{y}_i > 1 + b_M d_{\mathrm{U}} + 2\bar{\epsilon}\Big)\Big\}\\
\leq&\sum_{i=1}^{q}Pr\Big\{\frac{1}{n}\mathrm{y}_i^T\Big(I_n - \frac{g}{g+1}P_{\boldsymbol{a}}\Big)\mathrm{y}_i > 1 + b_M d_{\mathrm{U}} + 2\bar{\epsilon}\Big\} \to 0.
\end{aligned}
$$

Let $\bar{\epsilon} = 0.5$ and $\bar{C} = q(2 + b_M d_{\mathrm{U}})$, we have $Pr(\widetilde{\lambda}_i^{\boldsymbol{a}} > \bar{C}) \to 0$, $i = 1, \dots, q$, as $n \to 0$. Same as the proof above, we can show $Pr(\widetilde{\lambda}_i^{\boldsymbol{a}\cup\boldsymbol{t}} > \bar{C}) \to 0$, $i = 1, \dots, q$, as $n \to 0$. □

Lemma B.1.5. *Under Condition 2.4.1, 2.4.2 and 2.4.6, when* $\boldsymbol{a} \not\subseteq \boldsymbol{t}$*,*

1. *If* $p_{\boldsymbol{a}}$ *is bounded, the largest eigenvalue of* $\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}}\big)\mathrm{Y}$ *is* $O_p(1)$*;*
2. *If* $p_{\boldsymbol{a}}$ *is unbounded, the largest eigenvalue of* $\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}}\big)\mathrm{Y}$ *is at most* $O_p(r_{\boldsymbol{a}\cup\boldsymbol{t}})$*.*

Proof. As we know $P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}}$ is idempotent, then follow the same notations as in Lemma B.1.4, we have $tr\{\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}}\big)\mathrm{Y}\} = \sum_{i=1}^{q}\mathrm{y}_i^T(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}})\mathrm{y}_i$, and $\mathrm{y}_i^T(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}})\mathrm{y}_i \sim \chi^2_{r_{\boldsymbol{a}\cup\boldsymbol{t}} - r_{\boldsymbol{t}}}$. If $p_{\boldsymbol{a}}$ is bounded, then $r_{\boldsymbol{a}\cup\boldsymbol{t}} - r_{\boldsymbol{t}} = O(1)$. In this case, $tr\{\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}}\big)\mathrm{Y}\} = O_p(1)$, which means the largest eigenvalue of $\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}}\big)\mathrm{Y}$ is $O_p(1)$. By Condition 2.4.2, $U_{\boldsymbol{a}\cup\boldsymbol{t}}$ has full column rank, then $r_{\boldsymbol{a}\cup\boldsymbol{t}} = p_{\boldsymbol{a}\cup\boldsymbol{t}}(k+1)$. If $p_{\boldsymbol{a}}$ is unbounded, then $r_{\boldsymbol{a}\cup\boldsymbol{t}} - r_{\boldsymbol{t}} \preceq O_p(r_{\boldsymbol{a}\cup\boldsymbol{t}})$. So $tr\{\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}}\big)\mathrm{Y}\} \preceq O_p(r_{\boldsymbol{a}\cup\boldsymbol{t}})$, which means the largest eigenvalue of $\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}}\big)\mathrm{Y}$ is at most $O_p(r_{\boldsymbol{a}\cup\boldsymbol{t}})$. By Condition 2.4.6 we know $\frac{r_{\boldsymbol{a}\cup\boldsymbol{t}}}{n} = o_p(n)$. □

Lemma B.1.6. *Let* $\widetilde{\lambda}_M^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}}$ *denote the largest eigenvalue of* $\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{Y}$. *Under Condition 2.4.1, 2.4.3,* $Pr(\widetilde{\lambda}_M^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}} > \bar{\bar{C}}) \to 1$, *as* $n \to \infty$, *where* $\bar{\bar{C}}$ *is some fixed positive constant.*

Proof. Follow the same notations as in Lemma B.1.4, $tr\big\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{Y}\big\} = \sum_{i=1}^q \frac{1}{n}\mathrm{y}_i^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{y}_i$. Then,

$$\mathrm{y}_i^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{y}_i \sim \chi^2_{r_{\boldsymbol{a}\cup\boldsymbol{t}}-r_{\boldsymbol{a}}}(\phi_i^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}}),$$

where $\phi_i^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}} = \frac{1}{2}\mathrm{v}_i^T(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}})\mathrm{v}_i = \frac{1}{2}\mathrm{v}_i^T(I_n-P_{\boldsymbol{a}})P_{\boldsymbol{t}}\mathrm{v}_i = \frac{1}{2}\mathrm{v}_i^T(I_n-P_{\boldsymbol{a}})\mathrm{v}_i$ and $r_{\boldsymbol{a}\cup\boldsymbol{t}} - r_{\boldsymbol{a}} \le r_{\boldsymbol{t}} < \infty$. As in Lemma B.1.4, we know $\frac{1}{n}\phi_i^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}} \le \frac{1}{n}\mathrm{v}_i^T\mathrm{v}_i \le b_M d_{\mathrm{U}}$. Next, by Condotion 2.4.3,

$$\begin{aligned}\mathbb{E}\Big[tr\Big\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{Y}\Big\}\Big] &= \sum_{i=1}^q \mathbb{E}\Big[\frac{1}{n}\mathrm{y}_i^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{y}_i\Big] = \sum_{i=1}^q \frac{1}{n}(r_{\boldsymbol{a}\cup\boldsymbol{t}} - r_{\boldsymbol{a}} + \phi_i^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}}) \\ &\ge \frac{1}{2n}\sum_{i=1}^q \mathrm{v}_i^T(I_n-P_{\boldsymbol{a}})\mathrm{v}_i = \frac{1}{2n}tr\{\mathrm{E}_y^T(I_n-P_{\boldsymbol{a}})\mathrm{E}_y\} > C_0/2,\end{aligned}$$

$$\begin{aligned}Var\Big[\frac{1}{n}\mathrm{y}_i^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{y}_i\Big] &= \frac{1}{n^2}(2r_{\boldsymbol{a}\cup\boldsymbol{t}} - 2r_{\boldsymbol{a}} + 4\phi_i^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}}) \le \frac{1}{n}\Big(\frac{1}{n}2r_{\boldsymbol{t}} + \frac{1}{n}4\phi_i^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}}\Big) \\ &\le \frac{2r_{\boldsymbol{t}}}{n^2} + \frac{b_M d_{\mathrm{U}}}{n} \to 0, i = 1, \ldots, q.\end{aligned}$$

Then, $Var\Big[tr\Big\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{Y}\Big\}\Big] \le \sum_{i=1}^q\sum_{j=1}^q \sqrt{Var\Big[\frac{1}{n}\mathrm{y}_i^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{y}_i\Big]Var\Big[\frac{1}{n}\mathrm{y}_j^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{y}_j\Big]} \to 0$, as $n \to \infty$. So $Pr\Big\{tr\Big\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{Y}\Big\} > C_0/4\Big\} \to 1$. Let $\widetilde{\lambda}_i^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}}, i = 1, \ldots, q$ be the eigenvalues of $\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{Y}$, therefore

$$Pr\Big(\widetilde{\lambda}_M^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}} > \frac{C_0}{4q}\Big) \ge Pr(\sum_{i=1}^q \widetilde{\lambda}_i^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}} > C_0/4) = Pr\Big\{tr\Big\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{a}}\big)\mathrm{Y}\Big\} > C_0/4\Big\} \to 1.$$

Let $\bar{\bar{C}} = C_0/4q$, then we have $Pr(\widetilde{\lambda}_M^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}} > \bar{\bar{C}}) \to 1$, as $n \to \infty$. □

B.2 Combining Two Cases

Lemma B.2.1. *Under Condition 2.4.1, 2.4.2, 2.4.4 and 2.4.6,* $\mathrm{p}\lim_{n\to\infty}\mathrm{BF}(\boldsymbol{a}\cup\boldsymbol{t};\boldsymbol{t}|G) = 0$, *when* $\boldsymbol{a} \nsubseteq \boldsymbol{t}$.

Proof. Case 1: If $p_{\boldsymbol{a}}$ is bounded, by Lemma B.1.3, we know $\frac{1}{n}A_{\boldsymbol{t}}$ converges in probability to a positive definite constant matrix. So for large n, all eigenvalues of $\big(\frac{1}{n}A_{\boldsymbol{t}}\big)^{-\frac{1}{2}}$ are positive and $O_p(1)$. By Lemma B.1.5, the largest eigenvalue of $\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}}-P_{\boldsymbol{t}}\big)\mathrm{Y}$ is positive and $O_p(1)$. Since $g = O(n)$ and $d = O(1)$, we have the largest eigenvalue of $\frac{1}{d}\frac{g}{g+1}\big(\frac{1}{n}A_{\boldsymbol{t}}\big)^{-\frac{1}{2}}\big\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{a}\cup\boldsymbol{t}} - P_{\boldsymbol{t}}\big)\mathrm{Y}\big\}\big(\frac{1}{n}A_{\boldsymbol{t}}\big)^{-\frac{1}{2}}$ is positive $O_p(\frac{1}{n})$. Therefore,

$$\Delta(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t}) = \left|I_q + \frac{1}{d}\frac{g}{g+1}\left(\frac{1}{n}A_{\boldsymbol{t}}\right)^{-\frac{1}{2}}\left\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{t}} - P_{\boldsymbol{a}\cup\boldsymbol{t}}\big)\mathrm{Y}\right\}\left(\frac{1}{n}A_{\boldsymbol{t}}\right)^{-\frac{1}{2}}\right| \succeq \left\{1 - O_p\left(\frac{1}{n}\right)\right\}^h,$$

where h is the number of nonzero eigenvalues of matrix $\frac{1}{d}\frac{g}{g+1}\left(\frac{1}{n}A_{\boldsymbol{t}}\right)^{-\frac{1}{2}}\left\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{t}}-P_{\boldsymbol{a}\cup\boldsymbol{t}}\big)\mathrm{Y}\right\}\left(\frac{1}{n}A_{\boldsymbol{t}}\right)^{-\frac{1}{2}}$. Since the result also holds for every clique and separator, we have

$$\mathrm{II}(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t}) = \frac{\prod_{C\in\mathscr{C}}\big\{\Delta_C(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t})\big\}^{-\frac{b+n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\big\{\Delta_S(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t})\big\}^{-\frac{b+n+|S|-1}{2}}} \preceq \frac{\prod_{C\in\mathscr{C}}\big\{1 - O_p\big(\frac{1}{n}\big)\big\}^{-O(n)}}{\big\{1 - O_p\big(\frac{1}{n}\big)\big\}^{O(n)}} = \left\{1-O_p\left(\frac{1}{n}\right)\right\}^{-O(n)},$$

where θ is a constant. So $\mathrm{II}(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t}) \to$ some constant, as $n\to\infty$. Then

$$\mathrm{BF}(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t}|G) = \mathrm{I}\times\mathrm{II}(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t}) = \{O(n)+1\}^{-\frac{(p_{\boldsymbol{a}\cup\boldsymbol{t}}-p_{\boldsymbol{t}})(k+1)q}{2}} \times \text{constant} \to 0.$$

Case 2: If $p_{\boldsymbol{a}}$ is unbounded, similarly, we have

$$\Delta(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t}) = \left|I_q + \frac{1}{d}\frac{g}{g+1}\left(\frac{1}{n}A_{\boldsymbol{t}}\right)^{-\frac{1}{2}}\left\{\frac{1}{n}\mathrm{Y}^T\big(P_{\boldsymbol{t}} - P_{\boldsymbol{a}\cup\boldsymbol{t}}\big)\mathrm{Y}\right\}\left(\frac{1}{n}A_{\boldsymbol{t}}\right)^{-\frac{1}{2}}\right| \succeq \left\{1 - O_p\left(\frac{r_{\boldsymbol{a}\cup\boldsymbol{t}}}{n}\right)\right\}^h$$

and

$$\mathrm{II}(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t}) = \frac{\prod_{C\in\mathscr{C}}\big\{\Delta_C(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t})\big\}^{-\frac{b+n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\big\{\Delta_S(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t})\big\}^{-\frac{b+n+|S|-1}{2}}} \preceq \frac{\prod_{C\in\mathscr{C}}\big\{1 - O_p\big(\frac{r_{\boldsymbol{a}\cup\boldsymbol{t}}}{n}\big)\big\}^{-O(n)}}{\big\{1 - O_p\big(\frac{r_{\boldsymbol{a}\cup\boldsymbol{t}}}{n}\big)\big\}^{O(n)}} = \left\{1-O_p\left(\frac{r_{\boldsymbol{a}\cup\boldsymbol{t}}}{n}\right)\right\}^{-O(n)}.$$

Then,

$$log\{\mathrm{BF}(\boldsymbol{a}\cup\boldsymbol{t},\boldsymbol{t}|G)\} \preceq -\frac{(p_{\boldsymbol{a}\cup\boldsymbol{t}} - p_{\boldsymbol{t}})(k+1)q}{2}log\{O(n)+1\} - O(n)log\left\{1 - O_p\left(\frac{r_{\boldsymbol{a}\cup\boldsymbol{t}}}{n}\right)\right\}$$

$$
\begin{aligned}
&\preceq -O_p(r_{\boldsymbol{a}\cup\boldsymbol{t}})log\{O(n)+1\} - O_p(r_{\boldsymbol{a}\cup\boldsymbol{t}})O_p\left(\frac{n}{r_{\boldsymbol{a}\cup\boldsymbol{t}}}\right)log\left\{1 - O_p\left(\frac{r_{\boldsymbol{a}\cup\boldsymbol{t}}}{n}\right)\right\} \\
&= -O_p(r_{\boldsymbol{a}\cup\boldsymbol{t}})log\{O(n)+1\} \to -\infty(\text{as } log(1+x)/x \to 1, \text{ as } x \to 0).
\end{aligned}
$$

Therefore, $\mathrm{BF}(\boldsymbol{a}\cup\boldsymbol{t};\boldsymbol{t}|G) \to 0$, as $n \to \infty$. □

Lemma B.2.2. *Under Condition 2.4.1-2.4.6,* $\mathrm{p}\lim_{n\to\infty}\mathrm{BF}(\boldsymbol{a};\boldsymbol{a}\cup\boldsymbol{t}|G) = 0$*, when* $\boldsymbol{t} \nsubseteq \boldsymbol{a}$.

Proof. Let $\hat{\Sigma}_G^{-1}$ be the MLE of Σ_G^{-1} under model $\boldsymbol{a}$, then $f(\mathrm{Y}|\boldsymbol{a},\Sigma_G) \leq f(\mathrm{Y}|\boldsymbol{a},\hat{\Sigma}_G)$ for any positive definite matrix Σ_G under the given graph G. The explicit calculation of the MLE can be done by calculating the MLEs of cliques, separators and combining them. We assume that the MLE converges to a positive definite matrix ${\Sigma_G^0}^{-1}$. For the true graph G^*, this statement holds trivially. Under supremum norm for each clique and separator, given $0 < \epsilon < 1$, we have a ϵ'-neighborhood $Nb(\epsilon')$ of ${\Sigma_G^0}^{-1}$, where $0 < \epsilon' < \epsilon$, which satisfies $Nb(\epsilon') = \{\Sigma_G^{-1} : \|\Sigma_G^{-1} - {\Sigma_G^0}^{-1}\|_\infty < \epsilon'\}$ and $Pr\{Nb(\epsilon')\} > \delta' > 0$ under HIW prior, such that $|\Sigma_G^{-1}\Sigma_G^0| < 1+\epsilon$, $|\Sigma_G{\Sigma_G^0}^{-1}| < 1+\epsilon$. For large n, we also have $\hat{\Sigma}_G^{-1} \in Nb(\epsilon')$. So $|\hat{\Sigma}_G^{-1}\Sigma_G^0| < 1+\epsilon$, $|\hat{\Sigma}_G{\Sigma_G^0}^{-1}| < 1+\epsilon$ and $\|\hat{\Sigma}_G^{-1} - {\Sigma_G^0}^{-1}\|_\infty < \epsilon'$.

Now dividing numerator and denominator of $\mathrm{BF}(\boldsymbol{a};\boldsymbol{a}\cup\boldsymbol{t}|G)$ by $f(\mathrm{Y}|\boldsymbol{a},\hat{\Sigma}_G)$, the likelihood at MLE under model $\boldsymbol{a}$,

$$
\begin{aligned}
\mathrm{BF}(\boldsymbol{a};\boldsymbol{a}\cup\boldsymbol{t}|G) &= \frac{\int f(\mathrm{Y}|\boldsymbol{a},\Sigma_G)f(\Sigma_G|G)d\Sigma_G}{\int f(\mathrm{Y}|\boldsymbol{a}\cup\boldsymbol{t},\Sigma_G)f(\Sigma_G|G)d\Sigma_G} \\
&= \frac{\int \frac{f(\mathrm{Y}|\boldsymbol{a},\Sigma_G)}{f(\mathrm{Y}|\boldsymbol{a},\hat{\Sigma}_G)}f(\Sigma_G|G)d\Sigma_G}{\int \frac{f(\mathrm{Y}|\boldsymbol{a}\cup\boldsymbol{t},\Sigma_G)}{f(\mathrm{Y}|\boldsymbol{a},\hat{\Sigma}_G)}f(\Sigma_G|G)d\Sigma_G} \\
&< \frac{\int f(\Sigma_G|G)d\Sigma_G}{\int_{Nb(\epsilon')} \frac{f(\mathrm{Y}|\boldsymbol{a}\cup\boldsymbol{t},\Sigma_G)}{f(\mathrm{Y}|\boldsymbol{a},\hat{\Sigma}_G)}f(\Sigma_G|G)d\Sigma_G} \\
&= \frac{(g+1)^{\frac{(p_{\boldsymbol{a}\cup\boldsymbol{t}}-p_{\boldsymbol{a}})(k+1)q}{2}}}{\int_{Nb(\epsilon')} |\Sigma_G\hat{\Sigma}_G^{-1}|^{-\frac{n}{2}} exp\big[-\frac{1}{2}tr\{S(\boldsymbol{a}\cup\boldsymbol{t})\Sigma_G^{-1} - S(\boldsymbol{a})\hat{\Sigma}_G^{-1}\}\big]f(\Sigma_G|G)d\Sigma_G} \\
&= \frac{(g+1)^{\frac{(p_{\boldsymbol{a}\cup\boldsymbol{t}}-p_{\boldsymbol{a}})(k+1)q}{2}}}{\int_{Nb(\epsilon')} |\Sigma_G{\Sigma_G^0}^{-1}|^{-\frac{n}{2}}|\Sigma_G^0\hat{\Sigma}_G^{-1}|^{-\frac{n}{2}} exp\big[-\frac{1}{2}tr\{S(\boldsymbol{a}\cup\boldsymbol{t})\Sigma_G^{-1} - S(\boldsymbol{a})\hat{\Sigma}_G^{-1}\}\big]f(\Sigma_G|G)d\Sigma_G} \\
&= \frac{(g+1)^{\frac{(p_{\boldsymbol{a}\cup\boldsymbol{t}}-p_{\boldsymbol{a}})(k+1)q}{2}}(1+\epsilon)^n}{\int_{Nb(\epsilon')} exp\big[-\frac{1}{2}tr\{S(\boldsymbol{a}\cup\boldsymbol{t})\Sigma_G^{-1} - S(\boldsymbol{a})\hat{\Sigma}_G^{-1}\}\big]f(\Sigma_G|G)d\Sigma_G}.
\end{aligned}
$$

Next, let α be a $q \times 1$ vector, where $\alpha \in \mathbb{R}^q$, such that $\alpha^T \frac{S(\boldsymbol{a})-S(\boldsymbol{a}\cup\boldsymbol{t})}{n} \alpha = \widetilde{\lambda}_M^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}}$. Let $\beta = {\Sigma_G^0}^{-1/2}\alpha$ and $b_\beta = \|\beta\|_2^2 < \infty$. Denote λ'_M be the largest eigenvalue of ${\Sigma_G^0}^{-1/2} \frac{S(\boldsymbol{a})-S(\boldsymbol{a}\cup\boldsymbol{t})}{n} {\Sigma_G^0}^{-1/2}$, then $\beta^T {\Sigma_G^0}^{-1/2} \frac{S(\boldsymbol{a})-S(\boldsymbol{a}\cup\boldsymbol{t})}{n} {\Sigma_G^0}^{-1/2} \beta = \widetilde{\lambda}_M^{\boldsymbol{a}\cup\boldsymbol{t}-\boldsymbol{a}} \leq \lambda'_M \|\beta\|_2^2$. By Lemma B.1.6, $Pr(\lambda'_M > \bar{\bar{C}}/b_\beta) \to 1$, as $n \to \infty$.

By Lemma B.1.4 and Lemma B.1.6, we have

$$Pr\Big(\lambda'_M - \Big|tr\Big\{\frac{S(\boldsymbol{a}\cup\boldsymbol{t})}{n}(\hat{\Sigma}_G^{-1} - {\Sigma_G^0}^{-1})\Big\}\Big| - \Big|tr\Big\{\frac{S(\boldsymbol{a})}{n}({\Sigma_G^0}^{-1} - \hat{\Sigma}_G^{-1})\Big\}\Big| > \bar{\bar{C}}/b_\beta - 2q\epsilon\bar{C}\Big) \to 1.$$

Then, by choosing $\epsilon < \frac{\bar{\bar{C}}}{2qb_\beta\bar{C}}$, we know $\lambda'_M - \Big|tr\Big\{\frac{S(\boldsymbol{a}\cup\boldsymbol{t})}{n}(\hat{\Sigma}_G^{-1} - {\Sigma_G^0}^{-1})\Big\}\Big| - \Big|tr\Big\{\frac{S(\boldsymbol{a})}{n}({\Sigma_G^0}^{-1} - \hat{\Sigma}_G^{-1})\Big\}\Big| > \bar{\bar{C}}/(2b_\beta)$ in probability.

So, we have

$$\begin{aligned}
&-\frac{1}{2}tr\{S(\boldsymbol{a}\cup\boldsymbol{t})\Sigma_G^{-1} - S(\boldsymbol{a})\hat{\Sigma}_G^{-1}\} \\
&= \frac{n}{2}\Big[tr\Big\{{\Sigma_G^0}^{-1}\frac{S(\boldsymbol{a}) - S(\boldsymbol{a}\cup\boldsymbol{t})}{n}\Big\} - tr\Big\{\frac{S(\boldsymbol{a}\cup\boldsymbol{t})}{n}(\hat{\Sigma}_G^{-1} - {\Sigma_G^0}^{-1})\Big\} - tr\Big\{\frac{S(\boldsymbol{a})}{n}({\Sigma_G^0}^{-1} - \hat{\Sigma}_G^{-1})\Big\}\Big] \\
&\geq \frac{n}{2}\Big[tr\Big\{{\Sigma_G^0}^{-1/2}\frac{S(\boldsymbol{a}) - S(\boldsymbol{a}\cup\boldsymbol{t})}{n}{\Sigma_G^0}^{-1/2}\Big\} - \Big|tr\Big\{\frac{S(\boldsymbol{a}\cup\boldsymbol{t})}{n}(\hat{\Sigma}_G^{-1} - {\Sigma_G^0}^{-1})\Big\}\Big| - \Big|tr\Big\{\frac{S(\boldsymbol{a})}{n}({\Sigma_G^0}^{-1} - \hat{\Sigma}_G^{-1})\Big\}\Big|\Big] \\
&\geq \frac{n}{2}\Big[\lambda'_M - \Big|tr\Big\{\frac{S(\boldsymbol{a}\cup\boldsymbol{t})}{n}(\hat{\Sigma}_G^{-1} - {\Sigma_G^0}^{-1})\Big\}\Big| - \Big|tr\Big\{\frac{S(\boldsymbol{a})}{n}({\Sigma_G^0}^{-1} - \hat{\Sigma}_G^{-1})\Big\}\Big|\Big].
\end{aligned}$$

Then $Pr\Big\{-\frac{1}{2}tr\{S(\boldsymbol{a}\cup\boldsymbol{t})\Sigma_G^{-1} - S(\boldsymbol{a})\hat{\Sigma}_G^{-1}\} > \widetilde{C}n\Big\} \to 1$, as $n \to \infty$, where $\widetilde{C}$ is some fixed constant.

Since $BF_{\boldsymbol{a},\boldsymbol{a}\cup\boldsymbol{t}} < \frac{(g+1)^{\frac{(p_{\boldsymbol{a}\cup\boldsymbol{t}} - p_{\boldsymbol{a}})(k+1)q}{2}}(1+\epsilon)^n}{\int_{Nb(\epsilon')} exp\Big[-\frac{1}{2}tr\{S(\boldsymbol{a}\cup\boldsymbol{t})\Sigma_G^{-1} - S(\boldsymbol{a})\hat{\Sigma}_G^{-1}\}\Big] f(\Sigma_G|G)d\Sigma_G}$, then $Pr\Big\{BF_{\boldsymbol{a},\boldsymbol{a}\cup\boldsymbol{t}} < (g+1)^{\frac{p_{\boldsymbol{t}}(k+1)q}{2}}(1+\epsilon)^n e^{-\widetilde{C}n}\Big\} \to 1$. Therefore, $\mathrm{p}\lim_{n\to\infty} \mathrm{BF}(\boldsymbol{a};\boldsymbol{a}\cup\boldsymbol{t}|G) = 0$. □

By combining the results from Lemma B.2.1 and B.2.2, we have

$$\mathrm{p}\lim_{n\to\infty} \mathrm{BF}(\boldsymbol{a};\boldsymbol{t}|G) = \mathrm{p}\lim_{n\to\infty} \mathrm{BF}(\boldsymbol{a};\boldsymbol{a}\cup\boldsymbol{t}|G) \cdot \mathrm{p}\lim_{n\to\infty} \mathrm{BF}(\boldsymbol{a}\cup\boldsymbol{t};\boldsymbol{t}|G) = 0,$$

for any model $\boldsymbol{a} \neq \boldsymbol{t}$.

APPENDIX C

PLOTS OF ESTIMATED NONLINEAR FUNCTIONS

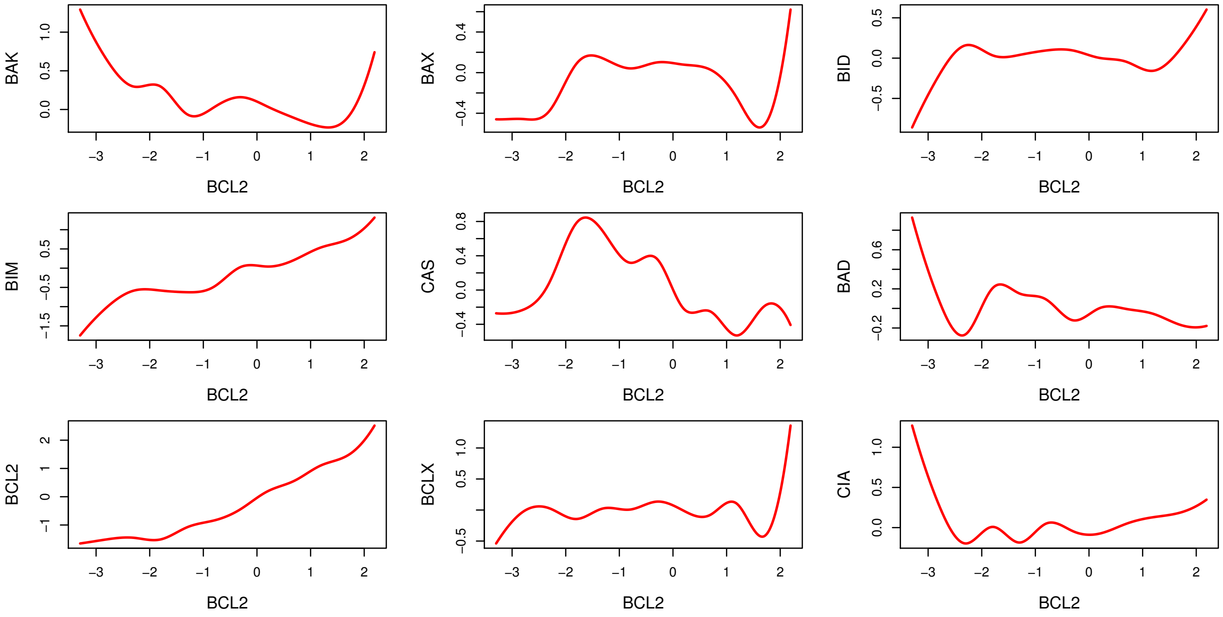


Figure C.1: Posterior mean of the nonlinear functions for proteins in apoptosis pathway, mRNA selected is BCL2.

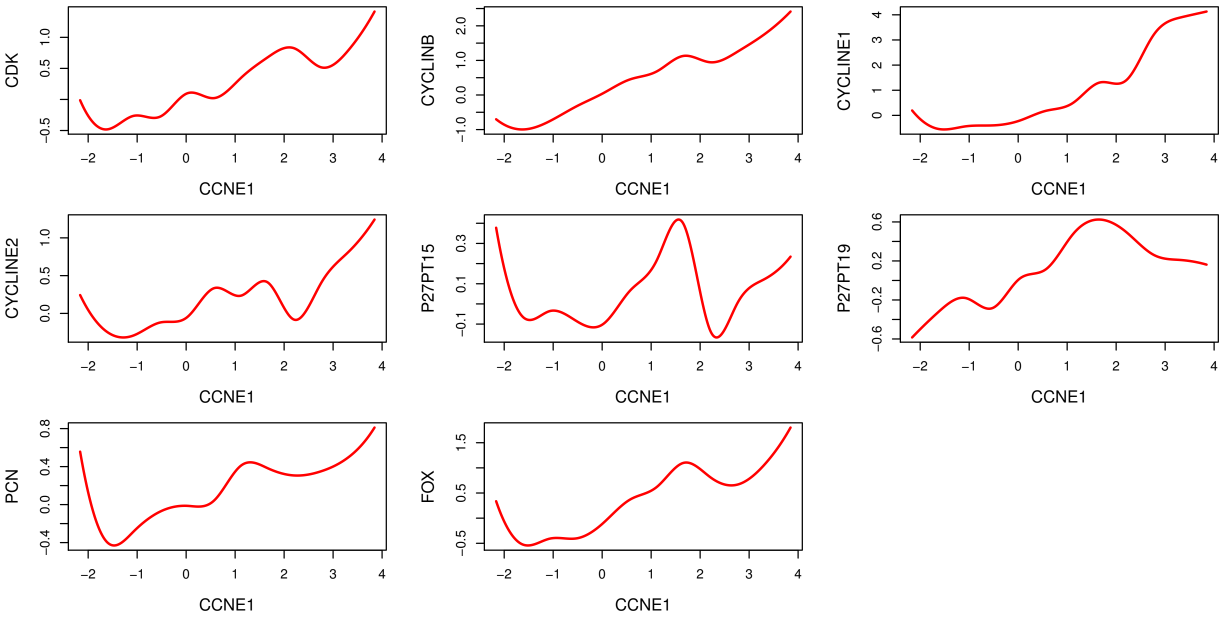


Figure C.2: Posterior mean of the nonlinear functions for proteins in cell cycle pathway, mRNA selected is CCNE1.

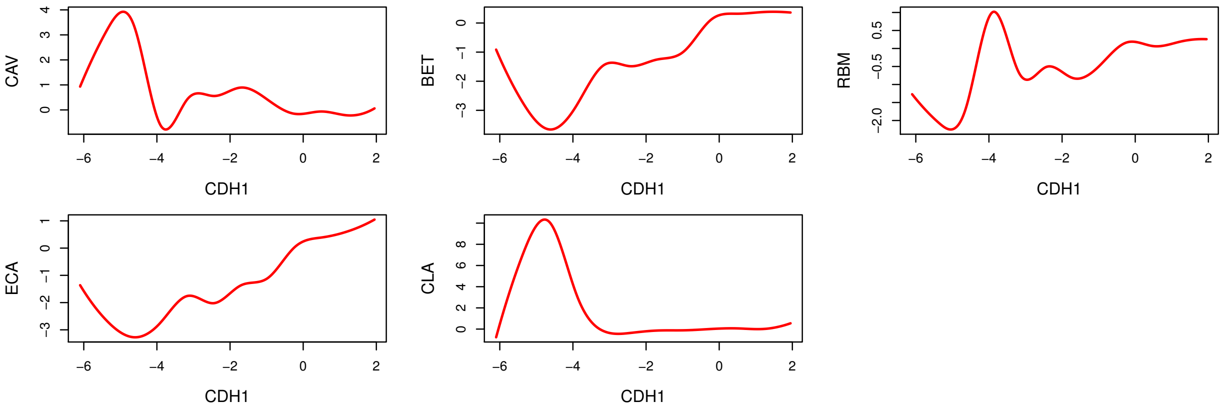


Figure C.3: Posterior mean of the nonlinear functions for proteins in core reactive pathway, mRNA selected is CDH1.

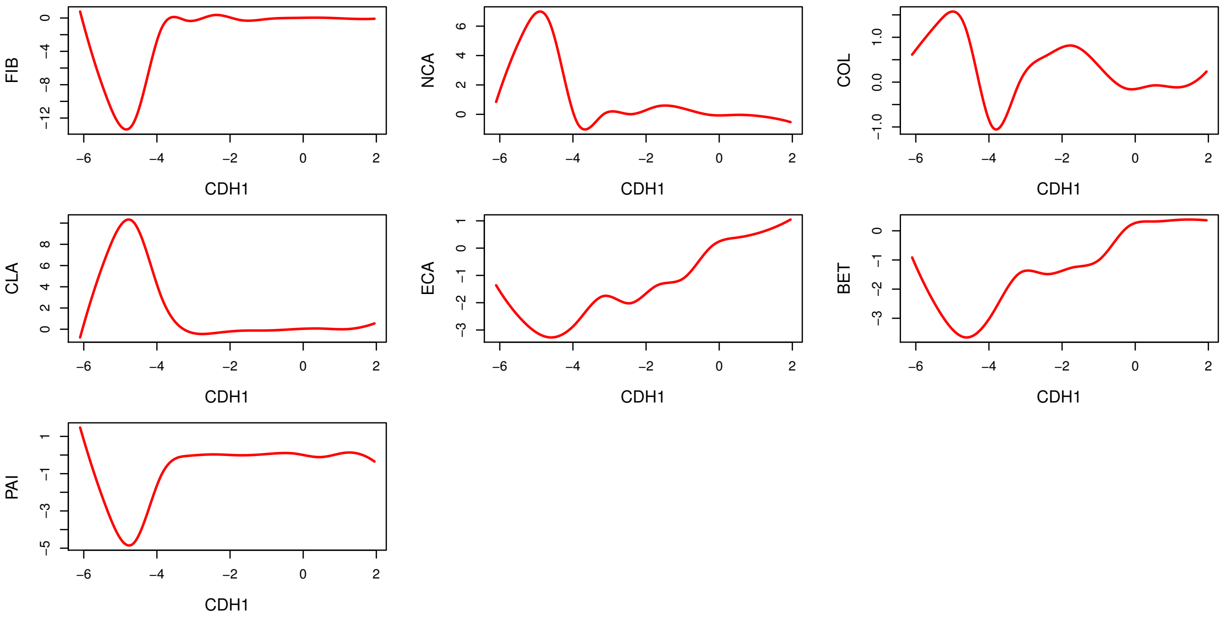


Figure C.4: Posterior mean of the nonlinear functions for proteins in EMT pathway, mRNA selected is CDH1.

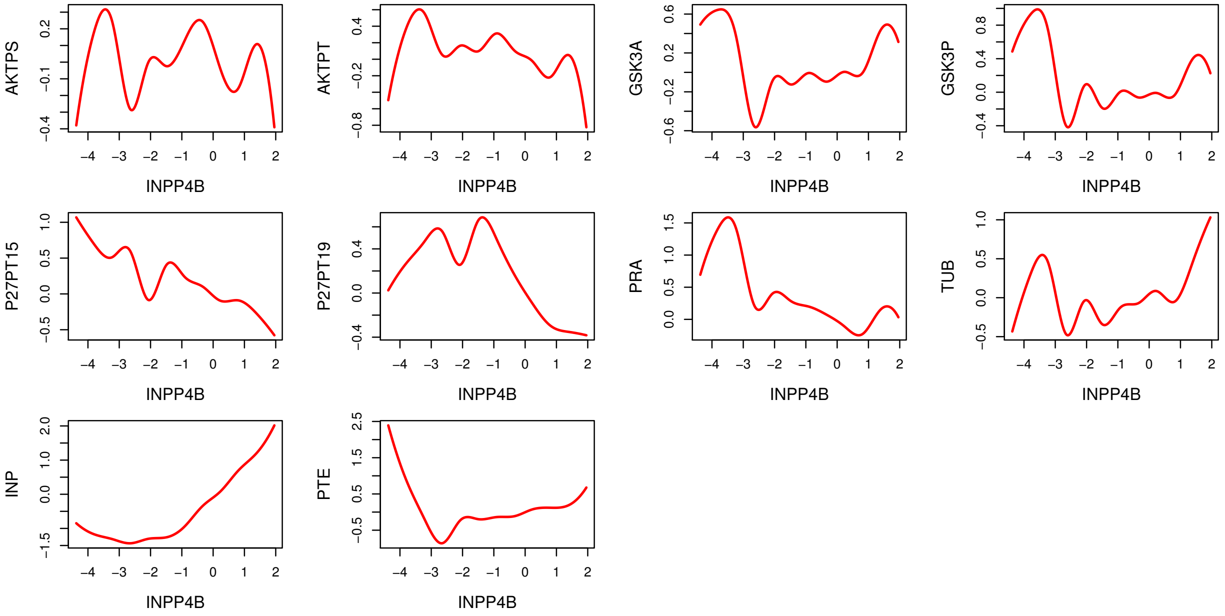


Figure C.5: Posterior mean of the nonlinear functions for proteins in PI3K/AKT pathway, mRNA selected is INPP4B.

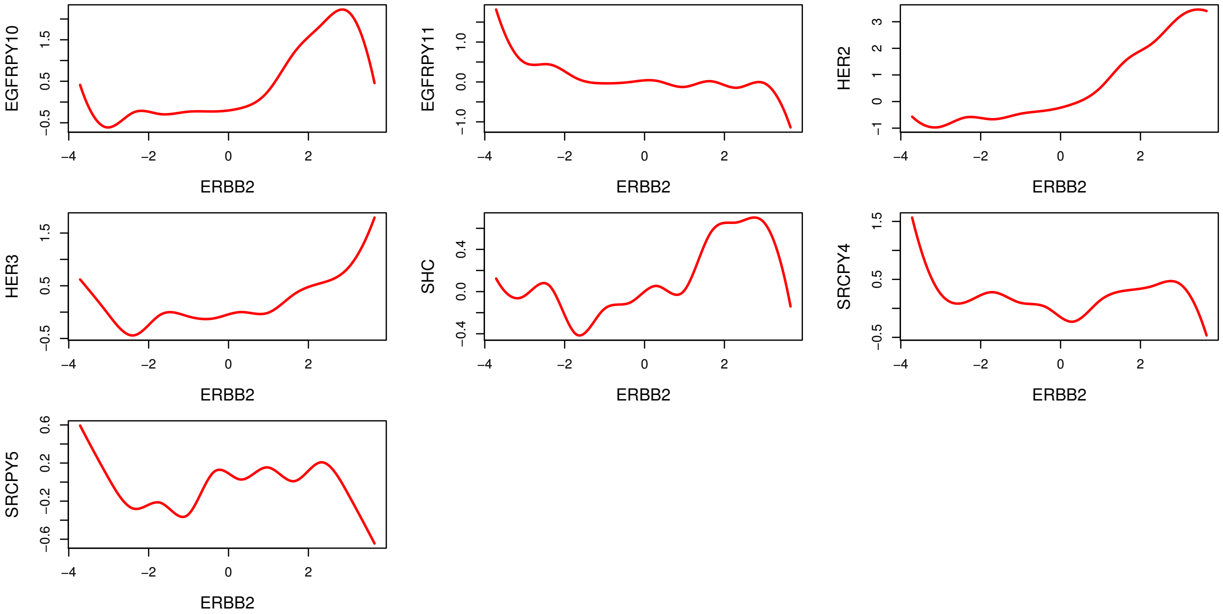


Figure C.6: Posterior mean of the nonlinear functions for proteins in RTK pathway, mRNA selected is ERBB2.

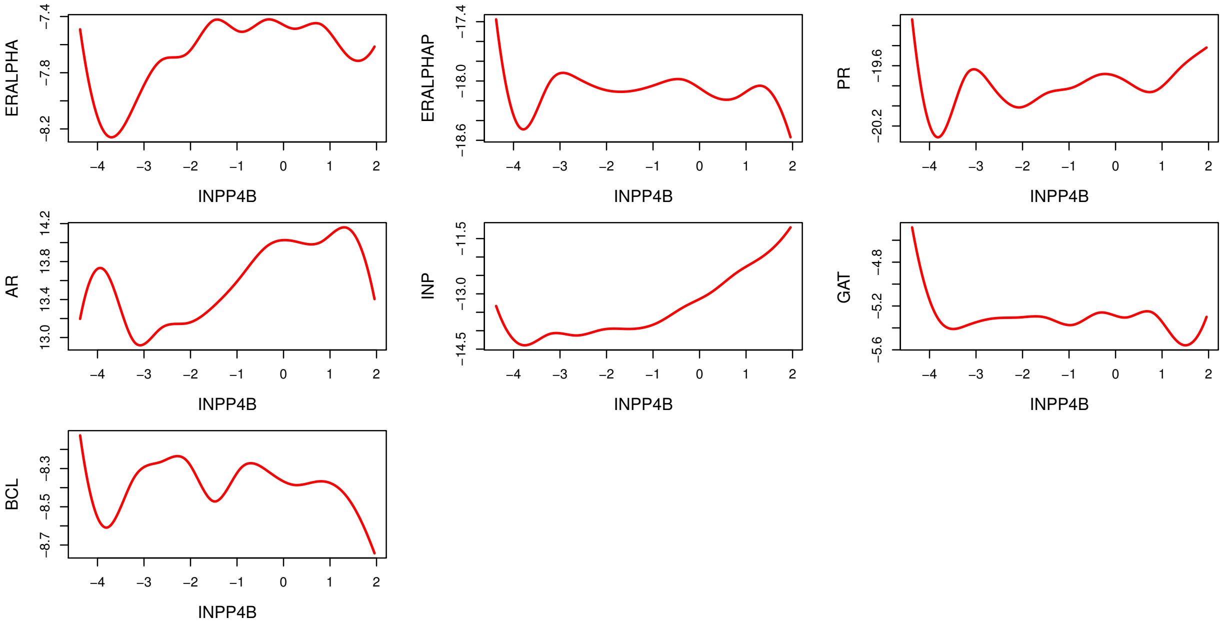


Figure C.7: Posterior mean of the nonlinear functions for proteins in hormone receptor&signaling pathway, mRNA selected is INPP4B.

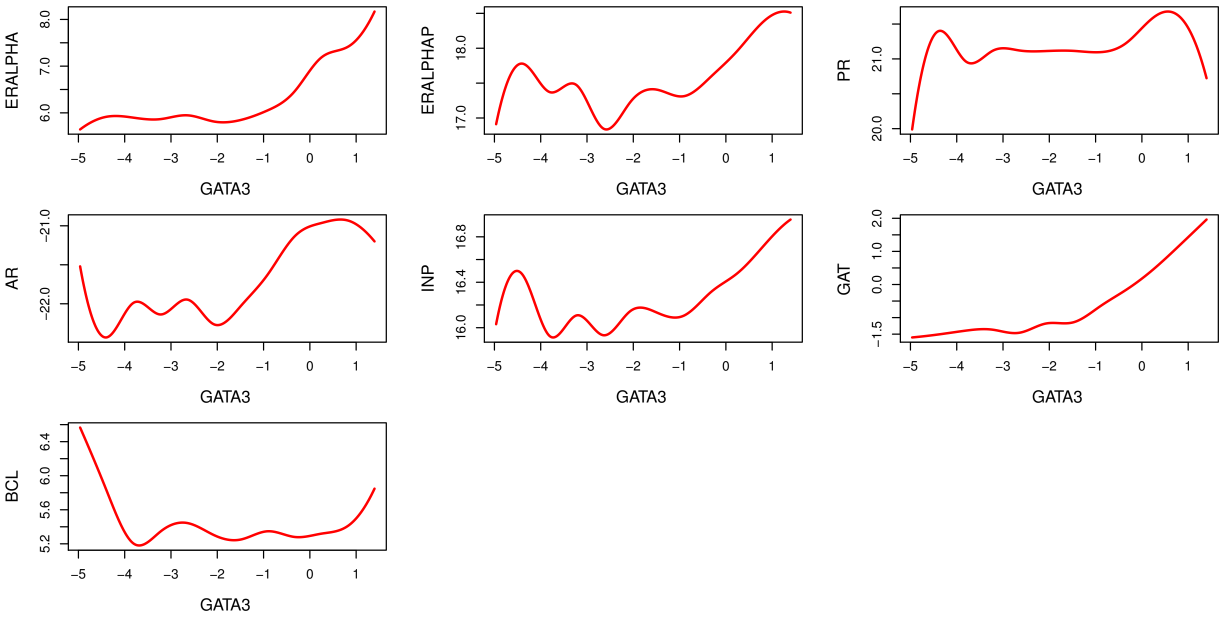


Figure C.8: Posterior mean of the nonlinear functions for proteins in hormone receptor&signaling pathway, mRNA selected is GATA3.

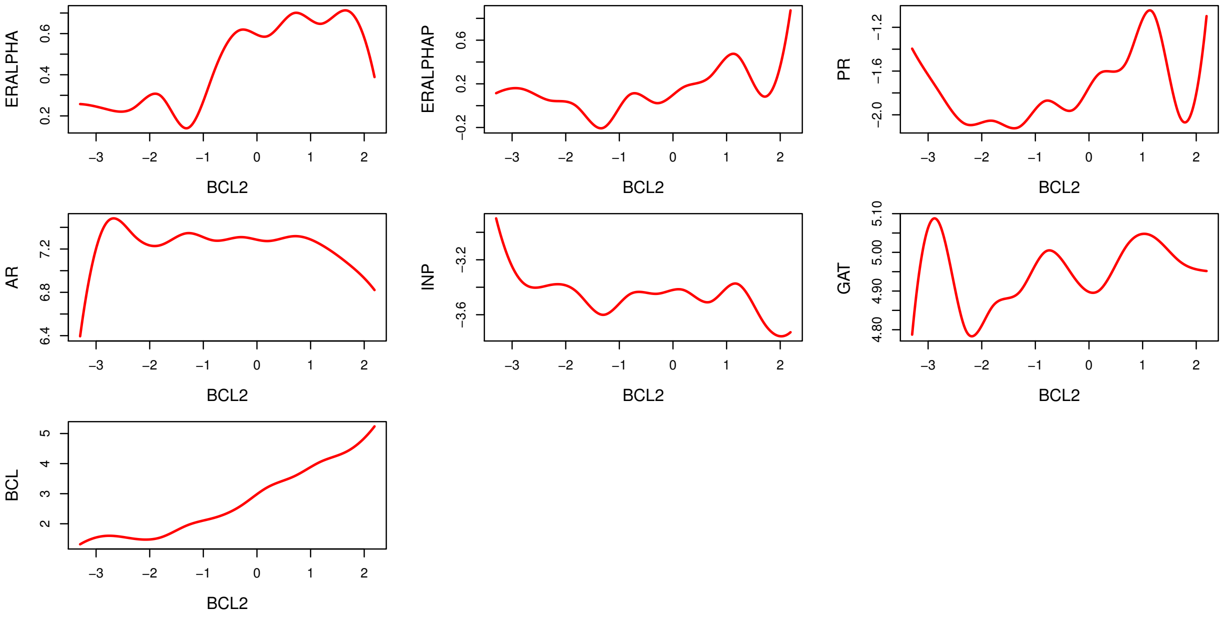


Figure C.9: Posterior mean of the nonlinear functions for proteins in hormone receptor&signaling pathway, mRNA selected is BCL2.

APPENDIX D

SOME RESULTS ON SAMPLE CORRELATION AND SAMPLE PARTIAL CORRELATION COEFFICIENTS [*]

D.1 Tail Behavior of Sample Partial Correlation

Theorem D.1.1. (When the population correlation is zero [74]). *Assume we have n i.i.d. samples from a multivariate Gaussian distribution. If the population correlation between X_i and X_j is zero, i.e. $\rho_{ij} = 0$, the density of the corresponding sample correlation coefficient $\hat{\rho}_{ij}$ as defined in Definition 3.2.2 is*

$$f_n(r \mid \rho_{ij} = 0) = \frac{\Gamma\big\{\frac{1}{2}(n-1)\big\}}{\Gamma\big\{\frac{1}{2}(n-2)\big\}\sqrt{\pi}}(1-r^2)^{\frac{1}{2}(n-4)}.$$

Theorem D.1.2. (When the population correlation is nonzero [75]). *The sample correlation coefficient in a sample of n from a bivariate normal distribution with population correlation coefficient ρ is distributed with density*

$$f_n(r \mid \rho) = \frac{n-2}{\sqrt{2\pi}}\frac{\Gamma(n-1)}{\Gamma(n-\frac{1}{2})}(1-\rho^2)^{\frac{1}{2}(n-1)}(1-r^2)^{\frac{1}{2}(n-4)}(1-\rho r)^{-n+\frac{3}{2}}F\Big(\frac{1}{2},\frac{1}{2};n-\frac{1}{2};\frac{1+\rho r}{2}\Big),$$

where $n > 2$, $-1 \leq r \leq 1$ and $F(\cdot,\cdot;\cdot;\cdot)$ is the hypergeometric function. When $\rho = 0$, the density becomes the same as in Theorem D.1.1.

Proposition D.1.1. (Mill's ratio). *Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the pdf and cdf of the standard normal distribution, respectively and $\widetilde{\Phi}(x) = 1 - \Phi(x)$. Then, we have $\phi(x)\big(\frac{1}{x} - \frac{1}{x^3}\big) \leq \widetilde{\Phi}(x) \leq \frac{\phi(x)}{x}$, for all $x > 0$.*

Proposition D.1.2. (Watson's inequality [76, 77]).

$$\sqrt{x+\frac{1}{4}} < \frac{\Gamma(x+1)}{\Gamma(x+\frac{1}{2})} \leq \sqrt{x+\frac{1}{\pi}} < \sqrt{x+\frac{1}{2}}, \quad \textit{for all } x \geq 0.$$

[*] Reprinted with permission from arXiv.org, "Bayesian Graph Selection Consistency Under Model Misspecification", arXiv preprint arXiv:1901.04134, 2019, by Niu, Yabo and Pati, Debdeep and Mallick, Bani K. In accordance arXiv copyright no modifications have been made except formatting.

Theorem D.1.3. (Tail behavior of sample correlation coefficient). *Let $\hat{\rho}_{ij}$ be the sample correlation coefficient between X_i and X_j with n samples from a p-dimensional normal distribution and the corresponding population correlation coefficient is ρ_{ij}, where $0 \leq |\rho_{ij}| < 1$. Then*

$$P\big(\,|\hat{\rho}_{ij} - \rho_{ij}| > \epsilon\big) < \frac{21}{(1-|\rho_{ij}|)^2}\frac{\exp(-n\epsilon^2/4)}{\epsilon\sqrt{n}}, \qquad \textit{for any } 0 < \epsilon < 1 - |\rho_{ij}|\,, n > 2.$$

Proof. First, let $r = \hat{\rho}_{ij}$ and $\rho = \rho_{ij}$, then by Theorem D.1.2, $f_n(x \mid \rho)$ is the pdf of r. Define

$$P_n(r_0, \rho) = P(r > r_0) = \int_{r_0}^{1} f_n(x \mid \rho)dx, \quad -1 \leq r_0 \leq 1.$$

By [75], we have

$$\begin{aligned} P_n(r_0,\rho) &= \frac{(n-2)\Gamma(n-1)}{\sqrt{2\pi}\Gamma(n-\frac{1}{2})}\left[M_0 + \frac{2M_0 - M_1}{4(2n-1)} + \frac{9(4M_0 - 4M_1 + M_2)}{32(2n-1)(2n+1)} + \dots\right] \\ &= \frac{(n-2)\Gamma(n-1)}{\sqrt{2\pi}\Gamma(n-\frac{1}{2})}(M_0 + R), \end{aligned}$$

where

$$\begin{aligned} M_k &= \int_{r_0}^{1}(1-\rho^2)^{\frac{1}{2}(n-1)}(1-x^2)^{\frac{1}{2}(n-4)}(1-\rho x)^{-n+k+\frac{3}{2}}dx, \qquad k = 0,1,2,\dots, \\ R &= \frac{2M_0 - M_1}{4(2n-1)} + \frac{9(4M_0 - 4M_1 + M_2)}{32(2n-1)(2n+1)} + \dots, \end{aligned}$$

and we know that the first term M_0 and the rest of the terms have the following inequality [75],

$$2(2n-1)\frac{1-|\rho|}{3-|\rho|} \leq \frac{M_0}{R} \leq 4(2n-1)\frac{1-|\rho|}{3-|\rho|}.$$

Let $\delta_\rho = \frac{1-|\rho|}{3-|\rho|}$. Since $0 \leq |\rho| < 1$, then $0 < \delta_\rho \leq \frac{1}{3}$. We can bound the residual term R by a fraction of M_0,

$$R \leq \frac{M_0}{2\delta_\rho(2n-1)} < \frac{M_0}{6\delta_\rho},$$

Therefore,

$$P_n(r_0,\rho) < \frac{(n-2)\Gamma(n-1)}{\sqrt{2\pi}\Gamma(n-\frac{1}{2})}\Big(1+\frac{1}{6\delta_\rho}\Big)M_0.$$

Next, we further simplify the bound of $P_n(r_0,\rho)$. By Proposition D.1.2, we have

$$\frac{\Gamma(n-1)}{\Gamma(n-\frac{1}{2})} < \frac{1}{\sqrt{n-\frac{5}{4}}} < \frac{1}{\sqrt{n-2}}.$$

Thus,

$$P_n(r_0,\rho) < \sqrt{\frac{n-2}{2\pi}}\Big(1+\frac{1}{6\delta_\rho}\Big)M_0 < \frac{1}{\sqrt{\pi}}\Big(1+\frac{1}{6\delta_\rho}\Big)\sqrt{n}M_0.$$

Let $r_0 = \rho+\epsilon > \rho$, where $0 < \epsilon \leq 1-\rho$. Next, we calculate the upper bound of $\sqrt{n}M_0$ for $0 \leq \rho < 1$ and $-1 < \rho < 0$ separately. But first, when $-1 < \rho < 1$ and $\rho < \rho+\epsilon \leq x \leq 1$, then $1-\rho x > 0$. Observe that,

$$\sqrt{n}M_0 = \sqrt{n}\int_{\rho+\epsilon}^{1}\Big(\frac{1-x^2}{1-\rho x}\Big)^{\frac{1}{2}(n-4)}\Big(\frac{1-\rho^2}{1-\rho x}\Big)^{\frac{1}{2}(n-1)}\frac{1}{1-\rho x}dx.$$

(I) When $0 \leq \rho < 1$. Since $\rho < \rho+\epsilon \leq x \leq 1$ and $\rho \geq 0$, we have $(1-\rho^2)^{-1} < (1-\rho x)^{-1} \leq (1-\rho)^{-1}$. Then

$$\sqrt{n}M_0 \leq \frac{\sqrt{n}}{1-\rho}\int_{\rho+\epsilon}^{1}\Big(1-\frac{x^2-\rho x}{1-\rho x}\Big)^{\frac{1}{2}(n-4)}\Big(1+\frac{\rho x-\rho^2}{1-\rho x}\Big)^{\frac{1}{2}(n-1)}dx.$$

Since $0 < \frac{x^2-\rho x}{1-\rho x} \leq 1$ and $0 < \frac{\rho x-\rho^2}{1-\rho x} \leq \rho$, we have

$$\begin{aligned}\sqrt{n}M_0 &\leq \frac{\sqrt{n}}{1-\rho}\int_{\rho+\epsilon}^{1}\exp\Big(-\frac{n}{2}\frac{x^2-\rho x}{1-\rho x}+2\frac{x^2-\rho x}{1-\rho x}\Big)\exp\Big(\frac{n}{2}\frac{\rho x-\rho^2}{1-\rho x}-\frac{1}{2}\frac{\rho x-\rho^2}{1-\rho x}\Big)dx\\ &\leq \frac{e^2\sqrt{n}}{1-\rho}\int_{\rho+\epsilon}^{1}\exp\Big(-\frac{n}{2}\frac{x^2-\rho x}{1-\rho x}\Big)\exp\Big(\frac{n}{2}\frac{\rho x-\rho^2}{1-\rho x}\Big)dx\\ &\leq \frac{e^2\sqrt{n}}{1-\rho}\int_{\rho+\epsilon}^{1}\exp\Big\{-\frac{n(x-\rho)^2}{2(1-\rho^2)}\Big\}dx.\end{aligned}$$

Thus, by Proposition D.1.1,

$$\begin{aligned}\sqrt{n}M_0 &\le e^2\sqrt{2\pi}\sqrt{\frac{1+\rho}{1-\rho}}\widetilde{\Phi}\Big(\frac{\epsilon\sqrt{n}}{\sqrt{1-\rho^2}}\Big)\\ &\le e^2\sqrt{2\pi}(1+\rho)\frac{\phi\Big(\frac{\epsilon\sqrt{n}}{\sqrt{1-\rho^2}}\Big)}{\epsilon\sqrt{n}} \le \frac{\exp(2+\rho/2)}{1-\rho}\cdot\frac{\exp(-n\epsilon^2/4)}{\epsilon\sqrt{n}}.\end{aligned}$$

(II) When $-1<\rho<0$. Since $\rho<\rho+\epsilon\le x\le 1$ and $\rho<0$, we have $(1-\rho)^{-1}\le(1-\rho x)^{-1}<(1-\rho^2)^{-1}$. Then

$$\sqrt{n}M_0 \le \frac{\sqrt{n}}{1-\rho^2}\int_{\rho+\epsilon}^1\Big(1-\frac{x^2-\rho x}{1-\rho x}\Big)^{\frac{1}{2}(n-4)}\Big(1+\frac{\rho x-\rho^2}{1-\rho x}\Big)^{\frac{1}{2}(n-1)}dx := \overline{M}.$$

(II.1) When $\rho+\epsilon<0$,

$$\begin{aligned}\overline{M} &= \frac{\sqrt{n}}{1-\rho^2}\bigg\{\int_{\rho+\epsilon}^0+\int_0^1\Big(1+\frac{\rho x-x^2}{1-\rho x}\Big)^{\frac{1}{2}(n-4)}\Big(1-\frac{\rho^2-\rho x}{1-\rho x}\Big)^{\frac{1}{2}(n-1)}dx\bigg\}\\ &:= A+B.\end{aligned}$$

Since $0\le\frac{\rho x-x^2}{1-\rho x}\le\Big(\frac{1-\sqrt{1-\rho^2}}{\rho}\Big)^2$ and $0<\frac{\rho^2-\rho x}{1-\rho x}\le\rho^2$ when $\rho<x\le 0$,

$$\begin{aligned}A &\le \frac{\sqrt{n}}{1-\rho^2}\int_{\rho+\epsilon}^0\exp\Big(\frac{n}{2}\frac{\rho x-x^2}{1-\rho x}-2\frac{\rho x-x^2}{1-\rho x}\Big)\exp\Big(-\frac{n}{2}\frac{\rho^2-\rho x}{1-\rho x}+\frac{1}{2}\frac{\rho^2-\rho x}{1-\rho x}\Big)dx\\ &\le \frac{e^{\frac{\rho^2}{2}}\sqrt{n}}{1-\rho^2}\int_{\rho+\epsilon}^0\exp\Big(\frac{n}{2}\frac{\rho x-x^2}{1-\rho x}\Big)\exp\Big(-\frac{n}{2}\frac{\rho^2-\rho x}{1-\rho x}\Big)dx\\ &\le \frac{e^{2-\frac{\rho}{2}}\sqrt{n}}{1-\rho^2}\int_{\rho+\epsilon}^0\exp\Big\{-\frac{n(x-\rho)^2}{2(1-\rho)}\Big\}dx,\ \text{since } 0<\frac{\rho^2}{2}<-\frac{\rho}{2}.\end{aligned}$$

Since $0\le\frac{x^2-\rho x}{1-\rho x}\le 1$ and $0<\rho^2\le\frac{\rho^2-\rho x}{1-\rho x}\le-\rho$ when $\rho<0\le x\le 1$,

$$\begin{aligned}B &\le \frac{\sqrt{n}}{1-\rho^2}\int_0^1\exp\Big(-\frac{n}{2}\frac{x^2-\rho x}{1-\rho x}+2\frac{x^2-\rho x}{1-\rho x}\Big)\exp\Big(-\frac{n}{2}\frac{\rho^2-\rho x}{1-\rho x}+\frac{1}{2}\frac{\rho^2-\rho x}{1-\rho x}\Big)dx\\ &\le \frac{e^{2-\frac{\rho}{2}}\sqrt{n}}{1-\rho^2}\int_0^1\exp\Big(-\frac{n}{2}\frac{x^2-\rho x}{1-\rho x}\Big)\exp\Big(-\frac{n}{2}\frac{\rho^2-\rho x}{1-\rho x}\Big)dx\end{aligned}$$

$$\le \frac{e^{2-\frac{\rho}{2}}\sqrt{n}}{1-\rho^2}\int_0^1 \exp\Big\{-\frac{n(x-\rho)^2}{2(1-\rho)}\Big\}dx,$$

Hence, when $-1<\rho<0$ and $\rho+\epsilon<0$, by Proposition D.1.1 we have

$$\begin{aligned}\sqrt{n}M_0 &\le e^{2-\frac{\rho}{2}}\sqrt{2\pi}\frac{\sqrt{1-\rho}}{1-\rho^2}\widetilde{\Phi}\Big(\frac{\epsilon\sqrt{n}}{\sqrt{1-\rho}}\Big)\\ &\le \frac{e^{2-\frac{\rho}{2}}\sqrt{2\pi}}{1+\rho}\frac{\phi\big(\frac{\sqrt{n}\epsilon}{\sqrt{1-\rho}}\big)}{\sqrt{n}\epsilon} \le \frac{\exp(2-\rho/2)}{1+\rho}\cdot\frac{\exp(-n\epsilon^2/4)}{\epsilon\sqrt{n}}.\end{aligned}$$

(II.2) When $\rho+\epsilon\ge 0$, similar to B, we still have

$$\sqrt{n}M_0 \le \frac{e^{2-\frac{\rho}{2}}\sqrt{n}}{1-\rho^2}\int_{\rho+\epsilon}^1 \exp\Big\{-\frac{n(x-\rho)^2}{2(1-\rho)}\Big\}dx \le \frac{\exp(2-\rho/2)}{1+\rho}\cdot\frac{\exp(-n\epsilon^2/4)}{\epsilon\sqrt{n}}.$$

So when $-1<\rho<1$ and $\rho<\rho+\epsilon<1$,

$$\begin{aligned}P(r>\rho+\epsilon) &< \frac{1}{\sqrt{\pi}}\Big(1+\frac{1}{6\delta_\rho}\Big)\frac{\exp(2+|\rho|/2)}{1-|\rho|}\cdot\frac{\exp(-n\epsilon^2/4)}{\epsilon\sqrt{n}}\\ &< \frac{7}{1-|\rho|}\Big(1+\frac{1}{6\delta_\rho}\Big)\frac{\exp(-n\epsilon^2/4)}{\epsilon\sqrt{n}}\\ &< \frac{10.5}{(1-|\rho|)^2}\frac{\exp(-n\epsilon^2/4)}{\epsilon\sqrt{n}}, \text{ for any } 0<\epsilon<1-\rho.\end{aligned}$$

For $P_n(r_0,\rho)$, we only need to consider when $r_0>\rho$, i.e. $r_0=\rho+\epsilon$. For the case which $r_0<\rho$, i.e. $-1<r_0=\rho-\epsilon<\rho$, we have the following equality,

$$\begin{aligned}P(r<\rho-\epsilon) &= 1-P(r>\rho-\epsilon)\\ &= 1-\int_{\rho-\epsilon}^1 f_n(-x\mid -\rho)dx\\ &= 1-\int_{-1}^{\epsilon-\rho} f_n(x\mid -\rho)dx\\ &= P(r>-\rho+\epsilon)\\ &< \frac{10.5}{(1-|\rho|)^2}\frac{\exp(-n\epsilon^2/4)}{\epsilon\sqrt{n}}, \qquad \text{for any } 0<\epsilon<1+\rho.\end{aligned}$$

Therefore,

$$P(|r-\rho|>\epsilon)<\frac{21}{(1-|\rho|)^2}\frac{\exp(-n\epsilon^2/4)}{\epsilon\sqrt{n}}, \quad \text{for any } 0<\epsilon<1-|\rho|\,.$$

□

Theorem D.1.4. (The CDF of sample partial correlation coefficient [74]). *If the cdf of sample correlation coefficient $\hat{\rho}_{ij}$ based on n samples from a normal distribution with population correlation coefficient ρ_{ij} is denoted by $F(r \mid n, \rho_{ij})$, then the cdf of the sample partial correlation coefficient $\hat{\rho}_{ij|s+1,\ldots,p}$, where $i, j < s+1$, based on n samples from a p-dimensional normal distribution with population partial correlation coefficient $\rho_{ij|s+1,\ldots,p}$ is $F(r \mid n-p+s, \rho_{ij|s+1,\ldots,p})$.*

The next corollary is an immediate result from Theorem D.1.3 and D.1.4.

Corollary D.1.1. (Tail behavior of sample partial correlation coefficient). *Let $\hat{\rho}_{ij|S}$ be the sample partial correlation coefficient between X_i and X_j, where $i, j \notin S$, holding X_S fixed based on n samples from a p-dimensional normal distribution and the corresponding population partial correlation coefficient is $\rho_{ij|S}$, where $0 \leq \left|\rho_{ij|S}\right| < 1$ and $|S| = d_S < p$. Then*

$$P\big(\left|\hat{\rho}_{ij|S}-\rho_{ij|S}\right|>\epsilon\big)<\frac{21}{(1-\left|\rho_{ij|S}\right|)^2}\frac{\exp\big\{-(n-d_S)\epsilon^2/4\big\}}{\epsilon\sqrt{n-d_S}}, \quad 0<\epsilon<1-\left|\rho_{ij|S}\right|.$$

D.2 Finding High Probability Region

Before introducing the next three lemmas, we first define some notations which are used by them and will be carried on using in the following proofs. Let $R_{ij|S} = \big\{\left|\hat{\rho}_{ij|S}-\rho_{ij|S}\right| \leq \epsilon\big\}$. If $(i,j) \notin E_t$, denote the set of all subsets (of V) which separate node i and j as $\Pi_{ij} = \big\{S \subseteq V\backslash\{i,j\} : \rho_{ij|S} = 0, (i,j) \notin E_t\big\}$, $1 \leq i < j \leq p$. Define

$$\Delta'_\epsilon = \Big\{\cap_{(i,j)\in E_t} R_{ij|V\backslash\{i,j\}}\Big\}\bigcap\Big\{\cap_{\substack{(i,j)\notin E_t,\\ \forall S\in\Pi_{ij}}} R_{ij|S}\Big\}, \text{ when } p<\infty,$$

$$\Delta'_\epsilon(n) = \Big\{\cap_{(i,j)\in E_t} R_{ij|V\backslash\{i,j\}}\Big\}\bigcap\Big\{\cap_{\substack{(i,j)\notin E_t,\\ \forall S\in\Pi_{ij}}} R_{ij|S}\Big\}, \text{ when } p \text{ grows with } n,$$

$$\Delta''_\epsilon(n) = \Big\{ \cap_{(i,j)\in E_t} R_{ij|V\setminus\{i,j\}} \Big\} \bigcap \Big\{ \cap_{(i,j)\notin E_t} \big(\cap_{S\in\Pi_{ij}} R_{ij|S} \big) \Big\}, \text{ when } p \text{ grows with } n,$$

where $\cap_{(i,j)\notin E_t,\, \forall S\in\Pi_{ij}}$ means intersection of $R_{ij|S}$ over all pairs of $(i,j) \notin E_t$ and for each pair any set of $S \in \Pi_{ij}$ can be used. The n in the bracket means the number of intersections depends on n. (When p grows with n, the number of edges in the true graph depends on n also.)

Lemma D.2.1. (Sample partial correlation simultaneous bounds for pairwise Bayes factor in finite graphs). *When the graph dimension p is finite, assume $\rho_U \neq 1$. Let $\epsilon_1(n) = \sqrt{\frac{\log(n-p)}{\tau(n-p)}}$. If $\tau > 0$, then $\mathbb{P}\big(\Delta'_{\epsilon_1}\big) \to 1$ as $n \to \infty$.*

Proof. For finite p, $\rho_U \neq 1$ is a positive constant which does not depend on n. By Corollary D.1.1, we have

$$\begin{aligned}
\mathbb{P}\big(\Delta'_{\epsilon_1}\big) &\geq 1 - \mathbb{P}\Big\{ \cup_{(i,j)\in E_t} R^C_{ij|V\setminus\{i,j\}} \Big\} - \mathbb{P}\Big\{ \cup_{\substack{(i,j)\notin E_t, \\ \forall S\in\Pi_{ij}}} R^C_{ij|S} \Big\} \\
&\geq 1 - 21\Big\{ \frac{|E_t|}{(1-\rho_U)^2} + p^2 - |E_t| \Big\}(n-p)^{-\frac{1}{4\tau}} \Big\{ \frac{1}{\tau}\log(n-p) \Big\}^{-\frac{1}{2}} \\
&\to 1, \text{ as } n \to \infty.
\end{aligned}$$

□

Lemma D.2.2. (Sample partial correlation simultaneous bounds for posterior ratio in high-dimensional graphs). *Under Assumption 3.4.1, i.e. the graph dimension $p = O(n^\alpha)$ grows with sample size n, where $0 < \alpha < 1$. Let $\epsilon_2(n) = (n-p)^{-\beta}$. If $0 < \beta < \frac{1}{2}$, under Assumption 3.4.5, then $\mathbb{P}\big\{\Delta'_{\epsilon_2}(n)\big\} \to 1$ as $n \to \infty$.*

Proof. By Corollary D.1.1, we have

$$\begin{aligned}
\mathbb{P}\big\{\Delta'_{\epsilon_2}(n)\big\} &\geq 1 - \mathbb{P}\Big\{ \cup_{(i,j)\in E_t} R^C_{ij|V\setminus\{i,j\}} \Big\} - \mathbb{P}\Big\{ \cup_{\substack{(i,j)\notin E_t, \\ \forall S\in\Pi_{ij}}} R^C_{ij|S} \Big\} \\
&\geq 1 - 21\Big\{ \frac{|E_t|}{(1-\rho_U)^2} + p^2 - |E_t| \Big\}(n-p)^{\beta-\frac{1}{2}} \exp\Big\{ -\frac{1}{4}(n-p)^{1-2\beta} \Big\} \\
&\to 1, \text{ as } n \to \infty.
\end{aligned}$$

□

Proposition D.2.1. (Lower and upper bound of binomial coefficient).

$$\Big(\frac{n}{k}\Big)^k \le \binom{n}{k} \le \Big(\frac{en}{k}\Big)^k,$$

where $k \le n$ and k, n are positive integers.

Lemma D.2.3. (Sample partial correlation simultaneous bounds for strong selection consistency in high-dimensional graphs). *Under Assumption 3.4.1, i.e. the graph dimension $p = O(n^\alpha)$ grows with sample size n, where $0 < \alpha < 1$. Let $\epsilon_3(n) = (n-p)^{-\beta}$, where $0 < \beta < \frac{1}{2}$. If $\alpha + 2\beta < 1$, under Assumption 3.4.5, then $\mathbb{P}\big\{\Delta''_{\epsilon_3}(n)\big\} \to 1$ as $n \to \infty$.*

Proof. By Corollary D.1.1, we have

$$\begin{aligned}
\mathbb{P}\big\{\Delta''_{\epsilon_3}(n)\big\} &\ge 1 - \mathbb{P}\Big\{\cup_{(i,j)\in E_t} R^C_{ij|V\setminus\{i,j\}}\Big\} - \mathbb{P}\Big\{\cup_{(i,j)\notin E_t}\big(\cup_{S\in\Pi_{ij}} R^C_{ij|S}\big)\Big\} \\
&\ge 1 - \sum_{(i,j)\in E_t} \mathbb{P}\big(R^C_{ij|V\setminus\{i,j\}}\big) - \sum_{(i,j)\notin E_t}\sum_{|S|=0}^{p-2}\binom{p-2}{|S|}\mathbb{P}\big(R^C_{ij|S}\big) \\
&\ge 1 - |E_t|\,\mathbb{P}\big(R^C_{ij|V\setminus\{i,j\}}\big) - \sum_{(i,j)\notin E_t}\sum_{|S|=0}^{p-2}(2e)^{p/2}\mathbb{P}\big(R^C_{ij|S}\big) \\
&\ge 1 - 21\Big\{\frac{|E_t|}{(1-\rho_U)^2} + p^3e^p\Big\}(n-p)^{\beta-\frac{1}{2}}\exp\Big\{-\frac{1}{4}(n-p)^{1-2\beta}\Big\} \\
&\to 1, \text{ as } n \to \infty.
\end{aligned}$$

□

Proposition D.2.2. (Sharp bounds for Beta CDF [78]). *Assume $Z \sim Beta(a,b)$, then*

$$\begin{aligned}
P(Z \le z) &< \frac{z^a(1-z)^b}{B(a,b)\{a-(a+b)z\}}, \quad z < \frac{a}{a+b}, \\
P(Z > z) &< \frac{z^a(1-z)^b}{B(a,b)\{(a+b)z-a\}}, \quad z > \frac{a}{a+b},
\end{aligned}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Theorem D.2.1. (Exact convergence rate of sample correlation coefficient when population correlation coefficient is zero). *Let $\hat{\rho}_{ij}$ be the sample correlation coefficient between X_i and X_j with n samples from a p-dimensional normal distribution. Assume its corresponding population correlation coefficient ρ_{ij} is zero. For any $0 < \epsilon < 1/2$, there exist two finite constant $0 < M_1(\epsilon) < 1/4$ and $M_2(\epsilon) > 3$, such that*

$$\mathbb{P}\Big(\hat{\rho}_{ij}^2 < \frac{M_1}{n}\Big) < \epsilon, \quad \mathbb{P}\Big(\hat{\rho}_{ij}^2 > \frac{M_2}{n}\Big) < \epsilon, \quad \textit{for any } n > 3.$$

Proof. By Theorem D.1.1, we know $\hat{\rho}_{ij}^2 \sim \text{Beta}\Big(\frac{1}{2}, \frac{n-2}{2}\Big)$. For any given ϵ, where $0 < \epsilon < \frac{1}{2}$, let $M_1 = \big(\frac{\epsilon}{\epsilon+1}\big)^2 < \frac{1}{4}$ and $M_2 = 6\log\big(\frac{5}{\epsilon}\big) > 3$. Thus, $\frac{M_1}{n} < \frac{1/2}{1/2+(n-2)/2}$ and $\frac{M_2}{n} > \frac{1/2}{1/2+(n-2)/2}$. By Proposition D.2.2,

$$\begin{aligned}
\mathbb{P}\Big(\hat{\rho}_{ij}^2 < \frac{M_1}{n}\Big) &< \frac{\big(\frac{M_1}{n}\big)^{\frac{1}{2}}\big(1-\frac{M_1}{n}\big)^{\frac{n-2}{2}}}{B\big(\frac{1}{2}, \frac{n-2}{2}\big)\big(\frac{1}{2} - \frac{n-1}{2}\frac{M_1}{n}\big)} \\
&< \frac{\Gamma\big(\frac{n-1}{2}\big)}{\Gamma\big(\frac{n-2}{2}\big)\sqrt{\pi}}\sqrt{\frac{M_1}{n}}\exp\Big(-\frac{M_1}{2}\frac{n-2}{n}\Big)\Big(\frac{1}{2} - \frac{M_1}{2}\frac{n-1}{n}\Big)^{-1} \\
&< \sqrt{\frac{n-2}{2n}}\sqrt{\frac{M_1}{\pi}}\Big(\frac{1}{2} - \frac{M_1}{2}\Big)^{-1} \\
&< \frac{\sqrt{M_1}}{1-\sqrt{M_1}} = \epsilon,
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}\Big(\hat{\rho}_{ij}^2 > \frac{M_2}{n}\Big) &< \frac{\big(\frac{M_2}{n}\big)^{\frac{1}{2}}\big(1-\frac{M_2}{n}\big)^{\frac{n-2}{2}}}{B\big(\frac{1}{2}, \frac{n-2}{2}\big)\big(\frac{n-1}{2}\frac{M_2}{n} - \frac{1}{2}\big)} \\
&< \frac{\Gamma\big(\frac{n-1}{2}\big)}{\Gamma\big(\frac{n-2}{2}\big)\sqrt{\pi}}\sqrt{\frac{M_2}{n}}\exp\Big(-\frac{M_2}{2}\frac{n-2}{n}\Big)\Big(\frac{M_2}{2}\frac{n-1}{n} - \frac{1}{2}\Big)^{-1} \\
&< \sqrt{\frac{M_2}{2\pi}}\exp\Big(-\frac{M_2}{6}\Big)\Big(\frac{M_2}{2}\frac{1}{2} - \frac{1}{2}\Big)^{-1} \\
&< 5\exp\Big(-\frac{M_2}{6}\Big) = \epsilon.
\end{aligned}$$

□

The next corollary is an immediate result from Theorem D.1.4 and D.2.1.

Corollary D.2.1. (Exact convergence rate of sample partial correlation coefficient when population partial correlation coefficient is zero). *Let $\hat{\rho}_{ij|S}$ be the sample partial correlation coefficient between X_i and X_j, where $i, j \notin S$, holding X_S fixed based on n samples from a p-dimensional normal distribution. Assume its corresponding population partial correlation coefficient $\rho_{ij|S}$ is zero. For any $0 < \epsilon < 1/2$, there exist two finite constant $0 < M_1(\epsilon) < 1/4$ and $M_2(\epsilon) > 3$, such that*

$$\mathbb{P}\Big(\hat{\rho}^2_{ij|S} < \frac{M_1}{n - d_S}\Big) < \epsilon, \quad \mathbb{P}\Big(\hat{\rho}^2_{ij|S} > \frac{M_2}{n - d_S}\Big) < \epsilon, \qquad \textit{for any } n > d_S + 3,\ d_S = |S|.$$

Lemma D.2.4. (Sample partial correlation simultaneous sharp bounds when population partial correlations are zero). *When the graph dimension p is finite, for any $0 < \epsilon < 1/2$, there exist two finite constant $0 < M_1(\epsilon) < 1/4$ and $M_2(\epsilon) > 3$, define*

$$R^0_{ij|S} = \Big\{\frac{M_1}{n} < \hat{\rho}^2_{ij|S} < \frac{M_2}{n - p}\Big\}, \quad \Delta^0_\epsilon = \cap_{\substack{(i,j)\notin E_t, \\ \forall S \in \Pi_{ij}}} R^0_{ij|S},$$

such that $\mathbb{P}\big(\Delta^0_\epsilon\big) > 1 - \epsilon$, when $n > p + 3$.

Proof. For any $0 < \epsilon < 1/2$, let

$$M_1 = \Big(\frac{\epsilon/p^2}{\epsilon/p^2 + 2}\Big)^2, \quad M_2 = 6 \log\Big(\frac{10p^2}{\epsilon}\Big).$$

By Theorem D.2.1 and Corollary D.2.1,

$$\mathbb{P}\Big(\hat{\rho}_{ij|S} < \frac{M_1}{n}\Big) < \frac{\epsilon}{2p^2}, \quad \mathbb{P}\Big(\hat{\rho}_{ij|S} > \frac{M_2}{n - p}\Big) < \frac{\epsilon}{2p^2},$$

for all $\hat{\rho}_{ij|S}$ such that $(i,j) \notin E_t$ and $S \in \Pi_{ij}$. Therefore,

$$\begin{aligned}\mathbb{P}\big(\Delta_\epsilon^0\big) &\geq 1 - \sum_{\substack{(i,j)\notin E_t,\\ \forall S\in\Pi_{ij}}} \mathbb{P}\Big(\hat{\rho}_{ij|S}^2 < \frac{M_1}{n}\Big) - \sum_{\substack{(i,j)\notin E_t,\\ \forall S\in\Pi_{ij}}} \mathbb{P}\Big(\hat{\rho}_{ij|S}^2 > \frac{M_2}{n-p}\Big)\\ &> 1 - p^2 \cdot \frac{\epsilon}{2p^2} - p^2 \cdot \frac{\epsilon}{2p^2} = 1 - \epsilon.\end{aligned}$$

□

Corollary D.2.2. *When the graph dimension* p *grows with* n*, for any* $0 < \epsilon < 1/2$ *and any positive integer* δ*, there exist two finite constant* $0 < M_1(\epsilon) < 1/4$ *and* $M_2(\epsilon) > 3$*, define*

$$R_{ij|S}^0 = \bigg\{\frac{M_1}{n} < \hat{\rho}_{ij|S}^2 < \frac{M_2}{n-p}\bigg\}, \quad \Delta_\epsilon^{0+} = \cap_{(i,j,S)\in\overline{E}_t} R_{ij|S}^0,$$

where

$$\overline{E}_t = \big\{(i,j,S) : (i,j) \notin E_t, S \in \Pi_{ij}, |\overline{E}_t| = \delta < \infty\big\},$$

we have $\mathbb{P}\big(\Delta_\epsilon^{0+}\big) > 1 - \epsilon$*, when* $n > p + 3$*.*

Proof. Let

$$M_1 = \Big(\frac{\epsilon/\delta}{\epsilon/\delta + 2}\Big)^2, \quad M_2 = 6\log\Big(\frac{10\delta}{\epsilon}\Big).$$

The rest of the proof proceeds the same as Lemma D.2.4. □

APPENDIX E

ENUMERATIONS OF ADDITION AND DELETION *

E.1 Enumerating Bayes Factors in the Deletion Case

Theorem E.1.1. (Condition of proper deletion while maintaining decomposability [79, 1, 15, 80]). *Removing an edge (x, y) from a decomposable graph G will result in a decomposable graph if and only if node x and y are contained in exactly one clique.*

For the rest of this appendix, we use lower-case letter x, y alone or with subscripts to represent nodes in the graph. We use the term "deletion" *only* in the case of deleting true edges. And true edges are the edges in the true graph G_t. Let $G_{+(x,y)\in E_t}$ and $G_{-(x,y)\in E_t}$ be any decomposable graph with and without the true edge (x, y), respectively. The remaining edges (excepting the true edge (x, y)) stays the same. (Notice $G_{+(x,y)\in E_t}$ does not need to be the true graph, except just containing the true edge (x, y).) Thus $G_{-(x,y)\in E_t}$ can be seen as the result of deleting the true edge (x, y) from $G_{+(x,y)\in E_t}$. From Theorem E.1.1, we know node x and y are contained in exactly one clique of $G_{+(x,y)\in E_t}$. The following Lemma E.1.1 provides upper and lower bound for Bayes factor in favor of deleting a true edge.

Lemma E.1.1. (Bayes factor of deleting one single true edge). *Denote C to be the only clique in $G_{+(x,y)\in E_t}$ that contains node x and y. Let $S = C\backslash\{x, y\}$. Then,*

$$\left(1+\frac{1}{g}\right)\sqrt{\frac{b+d_S-\frac{1}{2}}{b+n+d_S}}\big(1-\hat{\rho}^2_{xy|S}\big)^{\frac{n}{2}} < \mathrm{BF}\big(G_{-(x,y)\in E_t}; G_{+(x,y)\in E_t}\big)$$
$$< \left(1+\frac{1}{g}\right)\sqrt{\frac{b+d_S}{b+n+d_S-\frac{1}{2}}}\big(1-\hat{\rho}^2_{xy|S}\big)^{\frac{n}{2}},$$

where $d_S = |S| < p$. When $S = \emptyset$, $d_S = 0$ and the sample partial correlation coefficient $\hat{\rho}_{xy|S}$

*Reprinted with permission from arXiv.org, "Bayesian Graph Selection Consistency Under Model Misspecification", arXiv preprint arXiv:1901.04134, 2019, by Niu, Yabo and Pati, Debdeep and Mallick, Bani K. In accordance arXiv copyright no modifications have been made except formatting.

becomes the sample correlation coefficient $\hat{\rho}_{xy}$.

Proof. To proof this lemma, we enumerate all scenarios and calculate the Bayes factor above for every case. Similar enumeration also appears in [81].

CASE 1: Node x and y are contained in one clique C of $G_{+(x,y)\in E_t}$ which only has node x and y. In other words, removing edge (x,y) will result in adding an empty separator to the junction tree and also disconnecting clique C_1 and C_2, where C_1 is the clique before C and C_2 is the clique after C. They remain unchanged after deleting edge (x,y). This is the special scenario of CASE 2 where $S = \emptyset$. Figure E.1 illustrates the result of deleting edge (x,y) from $G_{+(x,y)\in E_t}$. Only the parts which are relative to the deletion are shown, the rest of the junction tree is omitted and will remain unchanged after the deletion. We use ellipses to denote cliques and squares to denote separators in the junction tree.

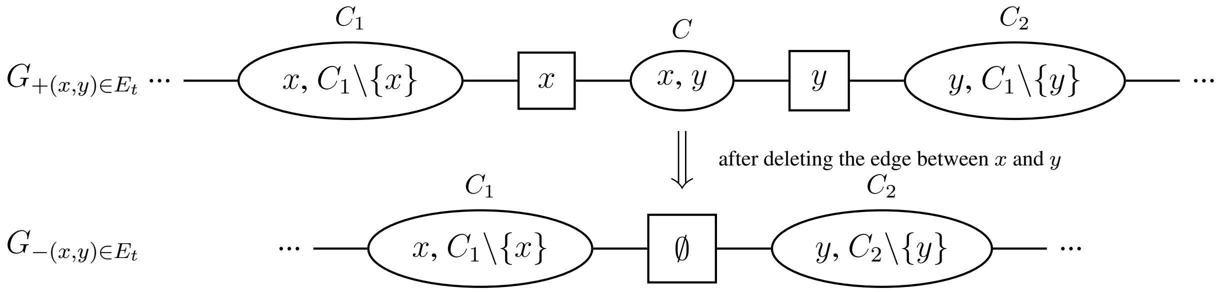


Figure E.1: Node x and y are in only one clique of $G_{+(x,y)\in E_t}$ that only contains themselves. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

$$\begin{aligned}&\mathrm{BF}\big(G_{-(x,y)\in E_t}; G_{+(x,y)\in E_t}\big)\\ &= \frac{f(\mathrm{Y} \mid G_{-(x,y)\in E_t})}{f(\mathrm{Y} \mid G_{+(x,y)\in E_t})} = \frac{1}{\frac{w(\{x,y\})}{w(\{x\})\cdot w(\{y\})}} = \frac{w(\{x\})\cdot w(\{y\})}{w(\{x,y\})}\end{aligned}$$

$$
\begin{aligned}
&= \left(1+\frac{1}{g}\right)\frac{\Gamma_2(\frac{b+1}{2})\Gamma^2(\frac{b+n}{2})}{\Gamma^2(\frac{b}{2})\Gamma_2(\frac{b+n+1}{2})}\left(\frac{\left|\mathrm{Y}_{xy}^T\mathrm{Y}_{xy}\right|}{\left|\mathrm{Y}_x^T\mathrm{Y}_x\right|\cdot\left|\mathrm{Y}_y^T\mathrm{Y}_y\right|}\right)^{\frac{n}{2}}\\
&= \left(1+\frac{1}{g}\right)\frac{\Gamma(\frac{b+1}{2})\Gamma(\frac{b+n}{2})}{\Gamma(\frac{b}{2})\Gamma(\frac{b+n+1}{2})}\left(\frac{\mathrm{Y}_x^T\mathrm{Y}_x\cdot\mathrm{X}_y^T\mathrm{Y}_y-(\mathrm{Y}_x^T\mathrm{Y}_y)^2}{\mathrm{Y}_x^T\mathrm{Y}_x\cdot\mathrm{X}_y^T\mathrm{Y}_y}\right)^{\frac{n}{2}}\\
&= \left(1+\frac{1}{g}\right)\frac{\Gamma(\frac{b+1}{2})\Gamma(\frac{b+n}{2})}{\Gamma(\frac{b}{2})\Gamma(\frac{b+n+1}{2})}\left(1-\hat{\rho}_{xy}^2\right)^{\frac{n}{2}}.
\end{aligned}
$$

By Proposition D.1.2,

$$
\sqrt{\frac{b-1}{2}+\frac{1}{4}}<\frac{\Gamma\left(\frac{b+1}{2}\right)}{\Gamma\left(\frac{b}{2}\right)}<\sqrt{\frac{b}{2}},\qquad \frac{1}{\sqrt{\frac{b+n}{2}}}<\frac{\Gamma\left(\frac{b+n}{2}\right)}{\Gamma\left(\frac{b+n+1}{2}\right)}<\frac{1}{\sqrt{\frac{b+n-1}{2}+\frac{1}{4}}}.
$$

Thus,

$$
\left(1+\frac{1}{g}\right)\sqrt{\frac{b-\frac{1}{2}}{b+n}}\left(1-\hat{\rho}_{xy}^2\right)^{\frac{n}{2}}<\mathrm{BF}\left(G_{-(x,y)\in E_t};G_{+(x,y)\in E_t}\right)<\left(1+\frac{1}{g}\right)\sqrt{\frac{b}{b+n-\frac{1}{2}}}\left(1-\hat{\rho}_{xy}^2\right)^{\frac{n}{2}}.
$$

CASE 2: Node x and y are contained in only one clique C of $G_{+(x,y)\in E_t}$ which consists of node x, y and a non-empty set S.

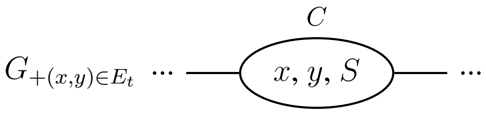


Figure E.2: When S is a non-empty set in $G_{+(x,y)\in E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

CASE 2.1: Both $\{x, S\}$ and $\{y, S\}$ are not separators in $G_{+(x,y)\in E_t}$. The cliques containing $\{x, S\}$ and $\{y, S\}$ are exactly $\{x, S\}$ and $\{y, S\}$ after the deletion in $G_{-(x,y)\in E_t}$, respectively [81]. Figure E.3 illustrates this scenario.

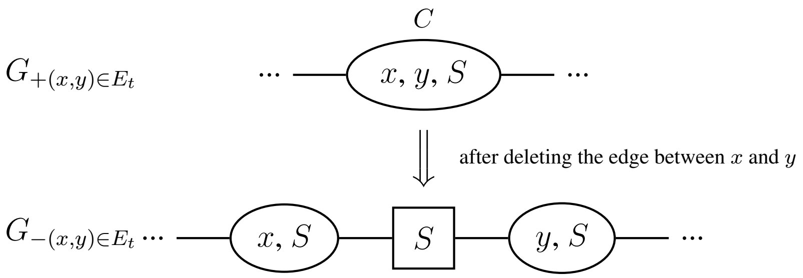


Figure E.3: Both $\{x, S\}$ and $\{y, S\}$ are not in other cliques of $G_{+(x,y)\in E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

Let

$$\begin{aligned}
\hat{\Sigma}_{SS} &= \mathrm{Y}_S^T \mathrm{Y}_S, \\
\hat{H}_S &= \mathrm{Y}_S(\mathrm{Y}_S^T \mathrm{Y}_S)^{-1}\mathrm{Y}_S^T, \\
\hat{\Sigma}_{xx|S} &= \mathrm{Y}_x^T \mathrm{Y}_x - \mathrm{Y}_x^T \hat{H}_S \mathrm{Y}_x, \\
\hat{\Sigma}_{yy|S} &= \mathrm{Y}_y^T \mathrm{Y}_y - \mathrm{Y}_y^T \hat{H}_S \mathrm{Y}_y, \\
\hat{\Sigma}_{xy|S} &= \mathrm{Y}_x^T \mathrm{Y}_y - \mathrm{Y}_x^T \hat{H}_S \mathrm{Y}_y.
\end{aligned}$$

Then we have

$$\begin{aligned}
\left|\mathrm{Y}_{xyS}^T \mathrm{Y}_{xyS}\right| &= \begin{vmatrix} \mathrm{Y}_x^T \mathrm{Y}_x & \mathrm{Y}_x^T \mathrm{Y}_y & \mathrm{Y}_x^T \mathrm{Y}_S \\ \mathrm{Y}_y^T \mathrm{Y}_x & \mathrm{Y}_y^T \mathrm{Y}_y & \mathrm{Y}_y^T \mathrm{Y}_S \\ \mathrm{Y}_S^T \mathrm{Y}_x & \mathrm{Y}_S^T \mathrm{Y}_y & \mathrm{Y}_S^T \mathrm{Y}_S \end{vmatrix} = \left|\mathrm{Y}_S^T \mathrm{Y}_S\right| \cdot \begin{vmatrix} \mathrm{Y}_x^T \mathrm{Y}_x - \mathrm{Y}_x^T \hat{H}_S \mathrm{Y}_x & \mathrm{Y}_x^T \mathrm{Y}_y - \mathrm{Y}_x^T \hat{H}_S \mathrm{Y}_y \\ \mathrm{Y}_y^T \mathrm{Y}_x - \mathrm{Y}_y^T \hat{H}_S \mathrm{Y}_x & \mathrm{Y}_y^T \mathrm{Y}_y - \mathrm{Y}_y^T \hat{H}_S \mathrm{Y}_y \end{vmatrix} \\
&= \left|\hat{\Sigma}_{SS}\right| \cdot \left(\hat{\Sigma}_{xx|S}\hat{\Sigma}_{yy|S} - \hat{\Sigma}_{xy|S}^2\right), \\
\left|\mathrm{Y}_{xS}^T \mathrm{Y}_{xS}\right| &= \begin{vmatrix} \mathrm{Y}_x^T \mathrm{Y}_x & \mathrm{Y}_x^T \mathrm{Y}_S \\ \mathrm{Y}_S^T \mathrm{Y}_x & \mathrm{Y}_S^T \mathrm{Y}_S \end{vmatrix} = \left|\mathrm{Y}_S^T \mathrm{Y}_S\right| \cdot \left|\mathrm{Y}_x^T \mathrm{Y}_x - \mathrm{Y}_x^T \hat{H}_S \mathrm{Y}_x\right| = \left|\hat{\Sigma}_{SS}\right| \cdot \hat{\Sigma}_{xx|S}, \\
\left|\mathrm{Y}_{yS}^T \mathrm{Y}_{yS}\right| &= \begin{vmatrix} \mathrm{Y}_y^T \mathrm{Y}_y & \mathrm{Y}_y^T \mathrm{Y}_S \\ \mathrm{Y}_S^T \mathrm{Y}_y & \mathrm{Y}_S^T \mathrm{Y}_S \end{vmatrix} = \left|\mathrm{Y}_S^T \mathrm{Y}_S\right| \cdot \left|\mathrm{Y}_y^T \mathrm{Y}_y - \mathrm{Y}_y^T \hat{H}_S \mathrm{Y}_y\right| = \left|\hat{\Sigma}_{SS}\right| \cdot \hat{\Sigma}_{yy|S}.
\end{aligned}$$

$$
\begin{aligned}
&\mathrm{BF}\big(G_{-(x,y)\in E_t}; G_{+(x,y)\in E_t}\big)\\
&= \frac{f(\mathrm{Y} \mid G_{-(x,y)\in E_t})}{f(\mathrm{Y} \mid G_{+(x,y)\in E_t})} = \frac{\frac{w(\{x,S\})\cdot w(\{y,S\})}{w(S)}}{w(\{x,y,S\})} = \frac{w(\{x,S\})\cdot w(\{y,S\})}{w(S)\cdot w(\{x,y,S\})}\\
&= \left(1+\frac{1}{g}\right)\frac{\Gamma_{d_S}\big(\frac{b+d_S-1}{2}\big)\Gamma_{d_S+2}\big(\frac{b+d_S+1}{2}\big)\Gamma^2_{d_S+1}\big(\frac{b+n+d_S}{2}\big)}{\Gamma^2_{d_S+1}\big(\frac{b+d_S}{2}\big)\Gamma_{d_S}\big(\frac{b+n+d_S-1}{2}\big)\Gamma_{d_S+2}\big(\frac{b+n+d_S+1}{2}\big)}\left(\frac{\big|\mathrm{Y}_S^T\mathrm{Y}_S\big|\cdot\big|\mathrm{Y}_{xyS}^T\mathrm{Y}_{xyS}\big|}{\big|\mathrm{Y}_{xS}^T\mathrm{Y}_{xS}\big|\cdot\big|\mathrm{Y}_{yS}^T\mathrm{Y}_{yS}\big|}\right)^{\frac{n}{2}}\\
&= \left(1+\frac{1}{g}\right)\frac{\Gamma\big(\frac{b+d_S+1}{2}\big)\Gamma\big(\frac{b+n+d_S}{2}\big)}{\Gamma\big(\frac{b+d_S}{2}\big)\Gamma\big(\frac{b+n+d_S+1}{2}\big)}\left(\frac{\hat{\Sigma}_{xx|S}\hat{\Sigma}_{yy|S}-\hat{\Sigma}^2_{xy|S}}{\hat{\Sigma}_{xx|S}\hat{\Sigma}_{yy|S}}\right)^{\frac{n}{2}}\\
&= \left(1+\frac{1}{g}\right)\frac{\Gamma\big(\frac{b+d_S+1}{2}\big)\Gamma\big(\frac{b+n+d_S}{2}\big)}{\Gamma\big(\frac{b+d_S}{2}\big)\Gamma\big(\frac{b+n+d_S+1}{2}\big)}\big(1-\hat{\rho}^2_{xy|S}\big)^{\frac{n}{2}}.
\end{aligned}
$$

By Proposition D.1.2,

$$
\sqrt{\frac{b+d_S-1}{2}+\frac{1}{4}} < \frac{\Gamma\big(\frac{b+d_S+1}{2}\big)}{\Gamma\big(\frac{b+d_S}{2}\big)} < \sqrt{\frac{b+d_S}{2}}, \qquad \frac{1}{\sqrt{\frac{b+n+d_S}{2}}} < \frac{\Gamma\big(\frac{b+n+d_S}{2}\big)}{\Gamma\big(\frac{b+n+d_S+1}{2}\big)} < \frac{1}{\sqrt{\frac{b+n+d_S-1}{2}+\frac{1}{4}}}.
$$

Thus,

$$
\begin{aligned}
\left(1+\frac{1}{g}\right)\sqrt{\frac{b+d_S-\frac{1}{2}}{b+n+d_S}}\big(1-\hat{\rho}^2_{xy|S}\big)^{\frac{n}{2}} &< \mathrm{BF}\big(G_{-(x,y)\in E_t}; G_{+(x,y)\in E_t}\big)\\
&< \left(1+\frac{1}{g}\right)\sqrt{\frac{b+d_S}{b+n+d_S-\frac{1}{2}}}\big(1-\hat{\rho}^2_{xy|S}\big)^{\frac{n}{2}}.
\end{aligned}
$$

CASE 2.2: Only one of $\{x,S\}$ and $\{y,S\}$ is a separator in $G_{+(x,y)\in E_t}$. The cliques containing $\{x,S\}$ or $\{y,S\}$ are a superset of $\{x,S\}$ or $\{y,S\}$ after the deletion in $G_{-(x,y)\in E_t}$, respectively [81]. Figure E.4 shows when $\{x,S\}$ is in other cliques (only one of those supersets is shown here which is $\{x,S,P\}$ and $P\neq\emptyset$, others are omitted for simplicity), thus $\{x,S\}$ is a separator in $G_{+(x,y)\in E_t}$. Figure E.5 shows when $\{y,S\}$ is in other cliques (which is $\{y,S,Q\}$ and $Q\neq\emptyset$), thus $\{y,S\}$ is a separator in $G_{+(x,y)\in E_t}$.

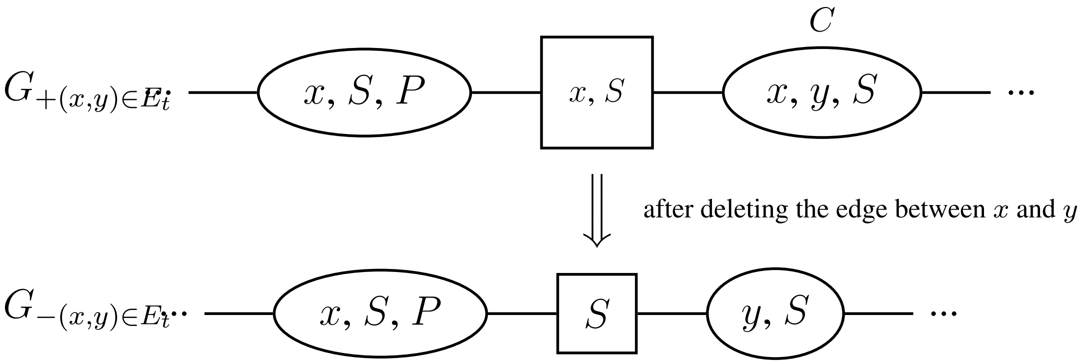


Figure E.4: Only x and S are in a superset $\{x, S, P\}$ of $G_{+(x,y)\in E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

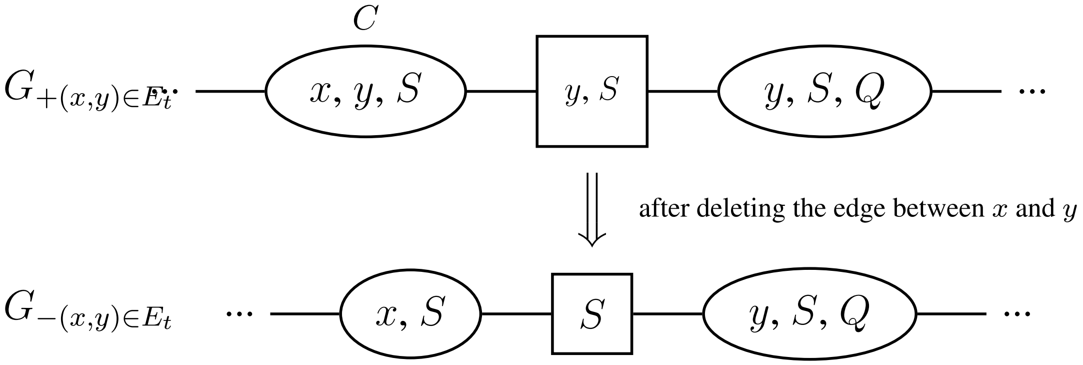


Figure E.5: Only y and S are in a superset $\{y, S, Q\}$ of $G_{+(x,y)\in E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

$$\text{BF}\big(G_{-(x,y)\in E_t}; G_{+(x,y)\in E_t}\big) = \frac{w(\{x, S\}) \cdot w(\{y, S\})}{w(S) \cdot w(\{x, y, S\})}.$$

This is the same as **CASE 2.1**.

CASE 2.3: Both $\{x, S\}$ and $\{y, S\}$ are separators in $G_{+(x,y)\in E_t}$. The cliques containing both $\{x, S\}$ and $\{y, S\}$ are supersets of them after the deletion in $G_{-(x,y)\in E_t}$ [81]. Figure E.6 shows

$\{x, S\}$ in superset $\{x, S, P\}$ and $\{y, S\}$ in superset $\{y, S, Q\}$, where $P, Q \neq \emptyset$ and $P \cap Q = \emptyset$, thus $\{x, S\}$ and $\{y, S\}$ are separators in $G_{+(x,y)\in E_t}$.

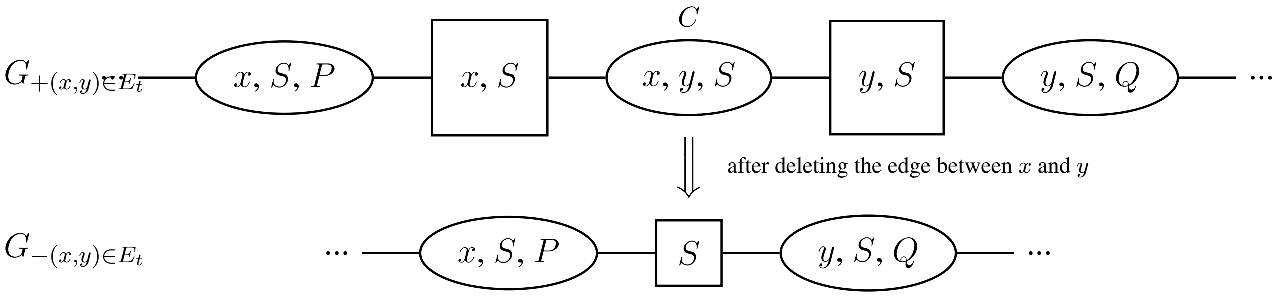


Figure E.6: $\{x, S\}$ and $\{y, S\}$ are in superset $\{x, S, P\}$ and $\{y, S, Q\}$ of $G_{+(x,y)\in E_t}$, respectively. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

$$\text{BF}\big(G_{-(x,y)\in E_t}; G_{+(x,y)\in E_t}\big) = \frac{w(\{x, S\}) \cdot w(\{y, S\})}{w(S) \cdot w(\{x, y, S\})}.$$

This is also the same as **CASE 2.1**. □

E.2 Enumerating Bayes factors in the addition case

Theorem E.2.1. (Condition of proper addition while maintaining decomposability [79, 15, 80]). *Adding an edge (x, y) to a decomposable graph G will result in a decomposable graph if and only if x and y are unconnected and contained in cliques that are adjacent in some junction tree of G.*

Notice we use the term "addition" *only* in the case of adding false edges, i.e., edges which are not in the true graph G_t. Let $G_{+(x,y)\notin E_t}$ and $G_{-(x,y)\notin E_t}$ be any decomposable graph with and without the false edge (x, y), respectively. And except the false edge (x, y), the rest of them are the same. ($G_{-(x,y)\notin E_t}$ does not need to be the true graph, except not having the false edge (x, y).) Therefore, $G_{+(x,y)\notin E_t}$ can be seen as the result of adding the false edge (x, y) to $G_{-(x,y)\notin E_t}$. By Theorem E.2.1, we know node x and y are contained in cliques that are adjacent in at least one junction tree of $G_{-(x,y)\notin E_t}$. Thus we have the following lemma.

Lemma E.2.1. (Bayes factor of adding one single false edge). *Let C_1 and C_2 be the cliques which contain x and y, respectively. Assume C_1 and C_2 are two adjacent nodes in at least one junction tree of $G_{-(x,y)\notin E_t}$. Let $S = C_1 \cap C_2$. Then,*

$$\left(\frac{g}{g+1}\right)\sqrt{\frac{b+n+d_S-\frac{1}{2}}{b+d_S}}\big(1-\hat{\rho}^2_{xy|S}\big)^{-\frac{n}{2}} < \text{BF}\big(G_{+(x,y)\notin E_t}; G_{-(x,y)\notin E_t}\big)$$

$$< \left(\frac{g}{g+1}\right)\sqrt{\frac{b+n+d_S}{b+d_S-\frac{1}{2}}}\big(1-\hat{\rho}^2_{xy|S}\big)^{-\frac{n}{2}},$$

where $d_S = |S| < p$. When $S = \emptyset$, $d_S = 0$ and the sample partial correlation coefficient $\hat{\rho}_{xy|S}$ becomes the sample correlation coefficient $\hat{\rho}_{xy}$.

Proof. Similar to the deletion case, we enumerate all scenarios and calculate the corresponding Bayes factors. The addition case can be partially seen as the reversion of the deletion case, only the edge added here is *not* a true edge. Same enumeration can be found in the appendix of [15].

CASE 1: Clique C_1 and C_2 are disconnected in $G_{-(x,y)\notin E_t}$, i.e. node x and y are not adjacent and not connected. (The graph can be seen as two separate subgraphs.) In other words, adding edge (x, y) will result in creating a new clique to the current junction tree of $G_{-(x,y)\notin E_t}$, and also connecting clique C_1 and C_2. They remain unchanged after adding edge (x, y). This is the special scenario of CASE 2 where $S = \emptyset$. Figure E.7 illustrates the result of adding a false edge (x, y) to $G_{-(x,y)\notin E_t}$. Here $P = C_1\backslash\{x\}$ and $Q = C_2\backslash\{y\}$, thus $P \cap Q = \emptyset$ and $P, Q \neq \emptyset$.

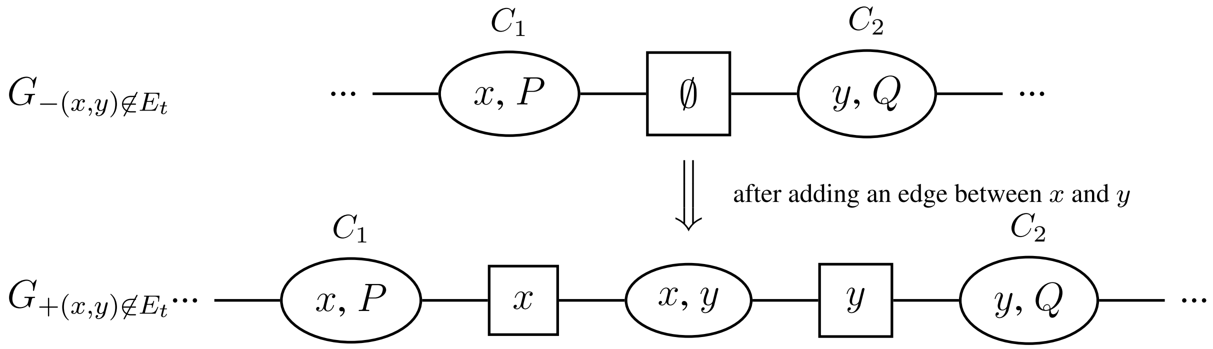


Figure E.7: Clique C_1 and C_2 are disconnected in $G_{-(x,y)\notin E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

$$
\begin{aligned}
&\mathrm{BF}\big(G_{+(x,y)\notin E_t}; G_{-(x,y)\notin E_t}\big) \\
&= \frac{f(\mathrm{Y} \mid G_{+(x,y)\notin E_t})}{f(\mathrm{Y} \mid G_{-(x,y)\notin E_t})} = \frac{w(\{x, y\})}{w(\{x\}) \cdot w(\{y\})} \\
&= \left(\frac{g}{g+1}\right) \frac{\Gamma^2(\frac{b}{2})\Gamma_2(\frac{b+n+1}{2})}{\Gamma_2(\frac{b+1}{2})\Gamma^2(\frac{b+n}{2})} \left(\frac{\left|\mathrm{Y}_{xy}^T \mathrm{Y}_{xy}\right|}{\left|\mathrm{Y}_x^T \mathrm{Y}_x\right| \cdot \left|\mathrm{Y}_y^T \mathrm{Y}_y\right|}\right)^{-\frac{n}{2}} \\
&= \left(\frac{g}{g+1}\right) \frac{\Gamma(\frac{b}{2})\Gamma(\frac{b+n+1}{2})}{\Gamma(\frac{b+1}{2})\Gamma(\frac{b+n}{2})} \left(\frac{\mathrm{Y}_x^T \mathrm{Y}_x \cdot \mathrm{X}_y^T \mathrm{Y}_y - (\mathrm{Y}_x^T \mathrm{Y}_y)^2}{\mathrm{Y}_x^T \mathrm{Y}_x \cdot \mathrm{X}_y^T \mathrm{Y}_y}\right)^{-\frac{n}{2}} \\
&= \left(\frac{g}{g+1}\right) \frac{\Gamma(\frac{b}{2})\Gamma(\frac{b+n+1}{2})}{\Gamma(\frac{b+1}{2})\Gamma(\frac{b+n}{2})} \big(1 - \hat{\rho}_{xy}^2\big)^{-\frac{n}{2}}.
\end{aligned}
$$

By Proposition D.1.2,

$$
\frac{1}{\sqrt{\frac{b}{2}}} < \frac{\Gamma\big(\frac{b}{2}\big)}{\Gamma\big(\frac{b+1}{2}\big)} < \frac{1}{\sqrt{\frac{b-1}{2} + \frac{1}{4}}}, \quad \sqrt{\frac{b+n-1}{2} + \frac{1}{4}} < \frac{\Gamma\big(\frac{b+n+1}{2}\big)}{\Gamma\big(\frac{b+n}{2}\big)} < \sqrt{\frac{b+n}{2}}.
$$

Thus,

$$
\left(\frac{g}{g+1}\right)\sqrt{\frac{b+n-\frac{1}{2}}{b}}\big(1-\hat{\rho}_{xy}^2\big)^{-\frac{n}{2}} < \mathrm{BF}\big(G_{+(x,y)\notin E_t}; G_{-(x,y)\notin E_t}\big) < \left(\frac{g}{g+1}\right)\sqrt{\frac{b+n}{b-\frac{1}{2}}}\big(1-\hat{\rho}_{xy}^2\big)^{-\frac{n}{2}}.
$$

CASE 2: Clique C_1 and C_2 are connected by a non-empty separator S in $G_{-(x,y)\notin E_t}$ and $P \cap Q = \emptyset$.

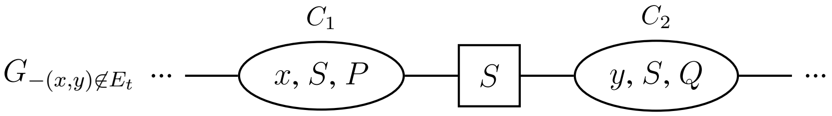


Figure E.8: When S is a non-empty separator in $G_{-(x,y)\notin E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

CASE 2.1: When P, Q are both empty sets, i.e. clique C_1 contains only $\{x, S\}$ and clique C_2 contains only $\{y, S\}$ in $G_{-(x,y)\notin E_t}$. In this case, adding an edge between x and y will consolidate C_1 and C_2 to create a single clique which consists of x, y and S. Figure E.9 shows this scenario.

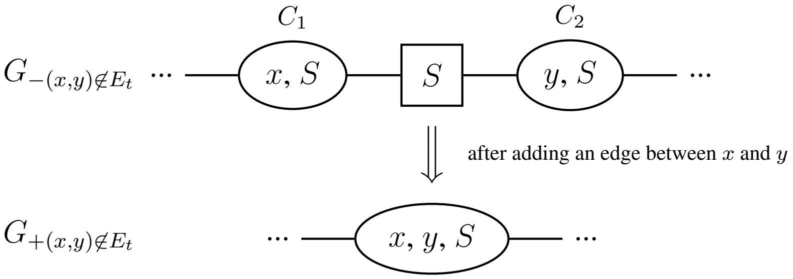


Figure E.9: When $P, Q = \emptyset$, i.e. $C_1 = \{x, S\}$ and $C_2 = \{y, S\}$ in $G_{-(x,y)\notin E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

$$
\begin{aligned}
&\mathrm{BF}\big(G_{+(x,y)\notin E_t}; G_{-(x,y)\notin E_t}\big)\\
&= \frac{f(\mathrm{Y} \mid G_{+(x,y)\notin E_t})}{f(\mathrm{Y} \mid G_{-(x,y)\notin E_t})} = \frac{w(\{x,y,S\})}{\frac{w(\{x,S\})\cdot w(\{y,S\})}{w(S)}} = \frac{w(\{x,y,S\})\cdot w(S)}{w(\{x,S\})\cdot w(\{y,S\})}\\
&= \left(\frac{g}{g+1}\right) \frac{\Gamma^2_{d_S+1}\big(\frac{b+d_S}{2}\big)\Gamma_{d_S}\big(\frac{b+n+d_S-1}{2}\big)\Gamma_{d_S+2}\big(\frac{b+n+d_S+1}{2}\big)}{\Gamma_{d_S}\big(\frac{b+d_S-1}{2}\big)\Gamma_{d_S+2}\big(\frac{b+d_S+1}{2}\big)\Gamma^2_{d_S+1}\big(\frac{b+n+d_S}{2}\big)} \left(\frac{\big|\mathrm{Y}_S^T\mathrm{Y}_S\big|\cdot\big|\mathrm{Y}_{xyS}^T\mathrm{Y}_{xyS}\big|}{\big|\mathrm{Y}_{xS}^T\mathrm{Y}_{xS}\big|\cdot\big|\mathrm{Y}_{yS}^T\mathrm{Y}_{yS}\big|}\right)^{-\frac{n}{2}}
\end{aligned}
$$

$$= \left(\frac{g}{g+1}\right) \frac{\Gamma\left(\frac{b+d_S}{2}\right)\Gamma\left(\frac{b+n+d_S+1}{2}\right)}{\Gamma\left(\frac{b+d_S+1}{2}\right)\Gamma\left(\frac{b+n+d_S}{2}\right)} \left(\frac{\hat{\Sigma}_{xx|S}\hat{\Sigma}_{yy|S} - \hat{\Sigma}_{xy|S}^2}{\hat{\Sigma}_{xx|S}\hat{\Sigma}_{yy|S}}\right)^{-\frac{n}{2}}$$

$$= \left(\frac{g}{g+1}\right) \frac{\Gamma\left(\frac{b+d_S}{2}\right)\Gamma\left(\frac{b+n+d_S+1}{2}\right)}{\Gamma\left(\frac{b+d_S+1}{2}\right)\Gamma\left(\frac{b+n+d_S}{2}\right)} \left(1 - \hat{\rho}_{xy|S}^2\right)^{-\frac{n}{2}}.$$

By Proposition D.1.2,

$$\frac{1}{\sqrt{\frac{b+d_S}{2}}} < \frac{\Gamma\left(\frac{b+d_S}{2}\right)}{\Gamma\left(\frac{b+d_S+1}{2}\right)} < \frac{1}{\sqrt{\frac{b+d_S-1}{2} + \frac{1}{4}}}$$

and

$$\sqrt{\frac{b+n+d_S-1}{2} + \frac{1}{4}} < \frac{\Gamma\left(\frac{b+n+d_S+1}{2}\right)}{\Gamma\left(\frac{b+n+d_S}{2}\right)} < \sqrt{\frac{b+n+d_S}{2}}.$$

Thus,

$$\left(\frac{g}{g+1}\right)\sqrt{\frac{b+n+d_S-\frac{1}{2}}{b+d_S}}\left(1 - \hat{\rho}_{xy|S}^2\right)^{-\frac{n}{2}} < \mathrm{BF}\left(G_{+(x,y)\notin E_t}; G_{-(x,y)\notin E_t}\right)$$
$$< \left(\frac{g}{g+1}\right)\sqrt{\frac{b+n+d_S}{b+d_S-\frac{1}{2}}}\left(1 - \hat{\rho}_{xy|S}^2\right)^{-\frac{n}{2}}.$$

CASE 2.2: One of P, Q is an empty set, i.e. clique C_1 contains only $\{x, S\}$ or clique C_2 contains only $\{y, S\}$ in $G_{-(x,y)\notin E_t}$. In this case, adding an edge between node x and y will not create a new clique, but extending the original separator S by node x or y. Figure E.10 shows when $P \neq \emptyset$ and $Q = \emptyset$, where $C_1 = \{x, S, P\}$, $C_2 = \{y, S\}$ in $G_{-(x,y)\notin E_t}$. Figure E.11 shows when $Q \neq \emptyset$ and $P = \emptyset$, where $C_1 = \{x, S\}$, $C_2 = \{y, S, Q\}$ in $G_{-(x,y)\notin E_t}$.

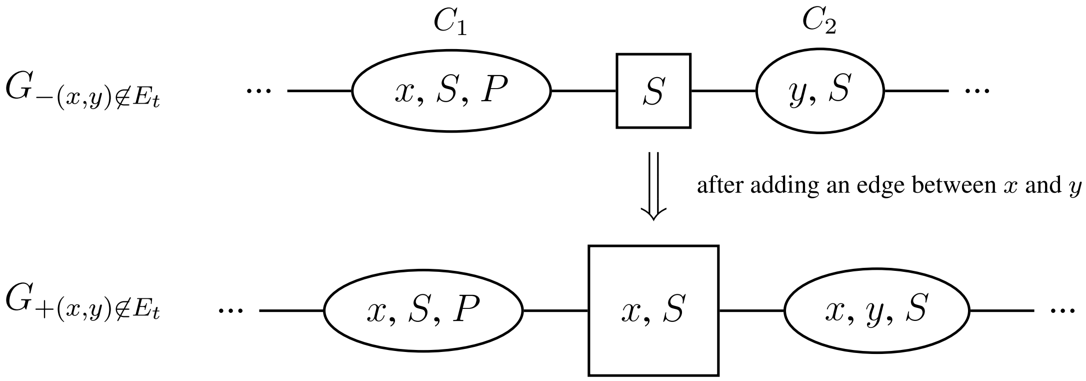


Figure E.10: $P \neq \emptyset$ and $Q = \emptyset$, where $C_1 = \{x, S, P\}$, $C_2 = \{y, S\}$ in $G_{-(x,y)\notin E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

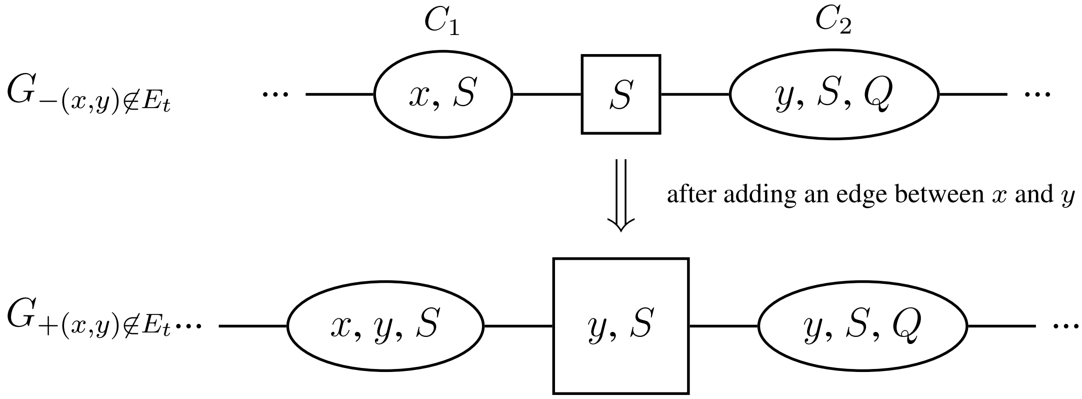


Figure E.11: $Q \neq \emptyset$ and $P = \emptyset$, where $C_1 = \{x, S\}$, $C_2 = \{y, S, Q\}$ in $G_{-(x,y)\notin E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

$$\text{BF}\big(G_{+(x,y)\notin E_t}; G_{-(x,y)\notin E_t}\big) = \frac{w(\{x, y, S\}) \cdot w(S)}{w(\{x, S\}) \cdot w(\{y, S\})}.$$

This is the same as **CASE 2.1**.

CASE 2.3: When P, Q are both non-empty sets and $P \cap Q = \emptyset$, i.e. $C_1 = \{x, S, P\}$, $C_2 = \{y, S, Q\}$ in $G_{-(x,y)\notin E_t}$. In this case, adding an edge between x and y will create a new clique

$\{x, y, S\}$ and two new separators $\{x, S\}$ and $\{y, S\}$. Figure E.12 illustrates this case.

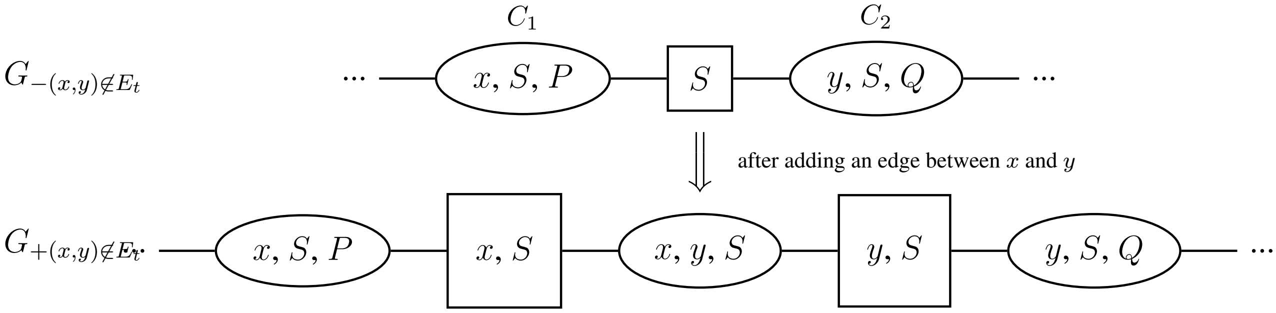


Figure E.12: $P, Q \neq \emptyset$ and $P \cap Q = \emptyset$, where $C_1 = \{x, S, P\}$, $C_2 = \{y, S, Q\}$ in $G_{-(x,y)\notin E_t}$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

$$\text{BF}\big(G_{+(x,y)\notin E_t}; G_{-(x,y)\notin E_t}\big) = \frac{w(\{x, y, S\}) \cdot w(S)}{w(\{x, S\}) \cdot w(\{y, S\})}.$$

This is also the same as **CASE 2.1**. □

APPENDIX F

PAIRWISE BAYES FACTOR CONSISTENCY AND POSTERIOR RATIO CONSISTENCY – ANY GRAPH G_A VERSUS THE TRUE GRAPH G_T [*]

F.1 Preparation

Lemma F.1.1. (Decomposable graph chain rule [1]). *Let $G = (V, E)$ be a decomposable graph and let $G' = (V, E')$ be a subgraph of G that also is decomposable with $|E \backslash E'| = k$. 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 $|E_t| > |E_a^1|$. By Lemma F.1.1, there exists a decreasing sequence of decomposable graphs from G_c to G_a that differ by exactly one edge, say $\left\{\overline{G}_i^{c\to a}\right\}_{i=0}^{|E_c|-|E_a|}$, where $G_c = \overline{G}_0^{c\to a} \supsetneqq \overline{G}_1^{c\to a} \supsetneqq \cdots \supsetneqq \overline{G}_{|E_c|-|E_a|-1}^{c\to a} \supsetneqq \overline{G}_{|E_c|-|E_a|}^{c\to a} = G_a$. There are $|E_c| - |E_a|$ steps for moving from G_c to G_a. Let $\left\{\rho_{\overline{x}_i\overline{y}_i|\overline{S}_i}\right\}_{i=1}^{|E_c|-|E_a|}$ be the corresponding population partial correlation (or correlation, when $\overline{S}_i = \emptyset$) sequence and $\left\{\text{BF}(\overline{G}_i^{c\to a}; \overline{G}_{i-1}^{c\to a})\right\}_{i=1}^{|E_c|-|E_a|}$ 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 $\text{BF}(\overline{G}_i^{c\to a}; \overline{G}_{i-1}^{c\to a})$ 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 removal of true edges that are deletion cases; $|E_c| - |E_a| - |E_t| + |E_a^1|$ steps are removal of false edges that can be seen as the reciprocal of addition cases.

Lemma F.1.2. (Origin of the exponential rate in the deletion case). *Assume $G_t \not\subset G_a$. In $\left\{\rho_{\overline{x}_i\overline{y}_i|\overline{S}_i}\right\}_{i=1}^{|E_c|-|E_a|}$, among all population partial correlations that are corresponding to the removal of true edges, at least one is non-zero and it is not a population correlation ($\overline{S}_i \neq \emptyset$).*

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 maintaining decomposability along the way. Let $(\overline{x}_*, \overline{y}_*) \in E_t \backslash E_a^1$.

[*] Reprinted with permission from arXiv.org, "Bayesian Graph Selection Consistency Under Model Misspecification", arXiv preprint arXiv:1901.04134, 2019, by Niu, Yabo and Pati, Debdeep and Mallick, Bani K. In accordance arXiv copyright no modifications have been made except formatting.

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. Let $\overline{S}_*$ be the corresponding separator. Thus we know $\overline{S}_* = V \backslash \{\overline{x}_*, \overline{y}_*\} \neq \emptyset$, since $(\overline{x}_*, \overline{y}_*)$ is removed from G_c. In fact, the removal of any edge from a complete graph still maintains decomposability, i.e. $\overline{G}_1^{c \to a}$ 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 $\rho_L \leq \left|\rho_{\overline{x}_*, \overline{y}_* | V \backslash \{\overline{x}_*, \overline{y}_*\}}\right| \leq \rho_U$. Therefore, we complete the proof of this lemma. □

Lemma F.1.3. (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. Assume S does not separate x from y in G. By the definition of separators, 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 F.1.1, there exists an increasing sequence of decomposable graphs from G_t to G_a that differ by exactly one edge, say $\left\{\widetilde{G}_i^{t \to a}\right\}_{i=0}^{|E_a|-|E_t|}$, 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 addition of false edges that are addition cases. Let $\left\{\rho_{\widetilde{x}_i \widetilde{y}_i | \widetilde{S}_i}\right\}_{i=1}^{|E_a|-|E_t|}$ be the corresponding population partial correlation (or correlation, when $\widetilde{S}_i = \emptyset$) sequence and $\left\{\text{BF}(\widetilde{G}_i^{t \to a}; \widetilde{G}_{i-1}^{t \to a})\right\}_{i=1}^{|E_a|-|E_t|}$ 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 $\text{BF}(\widetilde{G}_i^{t \to a}; \widetilde{G}_{i-1}^{t \to a})$ 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.

Lemma F.1.4. (Origin of the polynomial rate in the addition case). *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 $\left\{\rho_{\widetilde{x}_i \widetilde{y}_i | \widetilde{S}_i}\right\}_{i=1}^{|E_a|-|E_t|}$ are zero. (or correlation, when $\widetilde{S}_i = \emptyset$)*

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

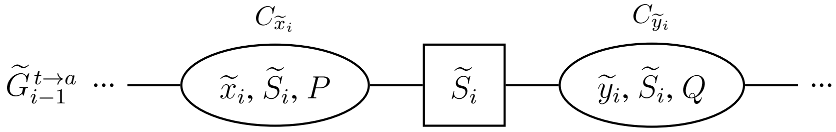

$C_{\widetilde{x}_i}$
$C_{\widetilde{y}_i}$
$\widetilde{G}^{t\to a}_{i-1}$ ···
$\widetilde{x}_i, \widetilde{S}_i, P$
$\widetilde{S}_i$
$\widetilde{y}_i, \widetilde{S}_i, Q$
···

Figure F.1: $\widetilde{G}^{t\to a}_{i-1}$ before adding edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ where $\widetilde{S}_i \neq \emptyset$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

First, when $\widetilde{S}_i \neq \emptyset$. Since edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ is added in the ith step, by Lemma E.2.1, $C_{\widetilde{x}_i}$ and $C_{\widetilde{y}_i}$ are adjacent in some junction tree of $\widetilde{G}^{t\to a}_{i-1}$ 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 $\widetilde{G}^{t\to a}_{i-1}$. Since $\left\{\widetilde{G}^{t\to a}_{i}\right\}_{i=0}^{|E_a|-|E_t|}$ is an increasing sequence by edge, by Lemma F.1.3, we know $\widetilde{S}_i$ also separates $\widetilde{x}_i$ from $\widetilde{y}_i$ in $\widetilde{G}^{t\to a}_{0} = G_t$. By the global Markov property, $\rho_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i} = 0$.

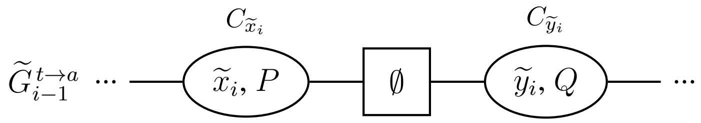


Figure F.2: $\widetilde{G}^{t\to a}_{i-1}$ before adding edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ where $\widetilde{S}_i = \emptyset$. Reprinted with permission from arXiv preprint, arXiv:1901.04134.

Next, when $\widetilde{S}_i = \emptyset$, 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 $\widetilde{G}^{t\to a}_{i-1}$, 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 $G_t \subsetneq \widetilde{G}^{t\to a}_{i-1}$, 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 order of deleting edges in Lemma F.1.2 (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 any order of adding edges (as long as decomposability is satisfied) according to Lemma F.1.4. Following the notations in Lemma F.1.2 and F.1.4, we have the decomposition of Bayes factor in favor of G_a as follows.

When $G_t \not\subset G_a$,

$$\begin{aligned}
\text{BF}(G_a; G_t) &= \frac{f(\text{Y} \mid G_a)}{f(\text{Y} \mid G_t)} = \frac{f(\text{Y} \mid G_a)}{f(\text{Y} \mid G_c)} \cdot \frac{f(\text{Y} \mid G_c)}{f(\text{Y} \mid G_t)} \\
&= \frac{p(\text{Y} \mid G_a)}{p(\text{Y} \mid \overline{G}^{c\to a}_{|E_c|-|E_a|-1})} \frac{p(\text{Y} \mid \overline{G}^{c\to a}_{|E_c|-|E_a|-1})}{p(\text{Y} \mid \overline{G}^{c\to a}_{|E_c|-|E_a|-2})} \cdots \frac{p(\text{Y} \mid \overline{G}^{c\to a}_{2})}{p(\text{Y} \mid \overline{G}^{c\to a}_{1})} \frac{p(\text{Y} \mid \overline{G}^{c\to a}_{1})}{p(\text{Y} \mid G_c)} \\
&\times \frac{p(\text{Y} \mid G_c)}{p(\text{Y} \mid \widetilde{G}^{t\to c}_{|E_c|-|E_t|-1})} \frac{p(\text{Y} \mid \widetilde{G}^{t\to c}_{|E_c|-|E_t|-1})}{p(\text{Y} \mid \widetilde{G}^{t\to c}_{|E_c|-|E_t|-2})} \cdots \frac{p(\text{Y} \mid \widetilde{G}^{t\to c}_{2})}{p(\text{Y} \mid \widetilde{G}^{t\to c}_{1})} \frac{p(\text{Y} \mid \widetilde{G}^{t\to c}_{1})}{p(\text{Y} \mid G_t)} \\
&= \prod_{i=1}^{|E_c|-|E_a|} \text{BF}(\overline{G}^{c\to a}_i; \overline{G}^{c\to a}_{i-1}) \cdot \prod_{i=1}^{|E_c|-|E_t|} \text{BF}(\widetilde{G}^{t\to c}_i; \widetilde{G}^{t\to c}_{i-1}) \\
&= \text{BF}_{c\to a} \cdot \text{BF}_{t\to c}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{PR}(G_a; G_t) = \frac{p(G_a \mid \text{Y})}{p(G_t \mid \text{Y})} &= \frac{f(\text{Y} \mid G_a)\pi(G_a)}{f(\text{Y} \mid G_t)\pi(G_t)} = \text{BF}(G_a; G_t)\frac{\pi(G_a)}{\pi(G_t)} \\
&= \text{BF}_{c\to a} \cdot \text{BF}_{t\to c} \cdot \left(\frac{q}{1-q}\right)^{|E_a|-|E_t|}.
\end{aligned}$$

$\text{BF}_{c\to a}$ contains $|E_c| - |E_a|$ terms, in which $|E_t| - |E_a^1|$ terms are deletion cases and $|E_c| - |E_a| - |E_t| + |E_a^1|$ terms are the reciprocal of addition cases. $\text{BF}_{t\to c}$ has $|E_c| - |E_t|$ terms that are all addition cases.

When $G_t \subsetneq G_a$,

$$\mathrm{BF}(G_a; G_t) = \prod_{i=1}^{|E_a|-|E_t|} \mathrm{BF}(\widetilde{G}_i^{t\to a}; \widetilde{G}_{i-1}^{t\to a}) = \mathrm{BF}_{t\to a},$$

$$\mathrm{PR}(G_a; G_t) = \mathrm{BF}_{t\to a} \cdot \left(\frac{q}{1-q}\right)^{|E_a|-|E_t|}.$$

F.2 Proof of Theorem 3.4.1

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

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

Given any decomposable graph $G_a \neq G_t$, when $G_t \not\subset G_a$, by Lemma F.1.2, 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 F.1.4. Let

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

Since $\rho_U \neq 1$, by the proof of Lemma D.2.1, we have

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

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 F.1.4. (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 D.2.1, we also have

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

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, 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 $|E_c| > |E_t|$. We begin by simplifying the upper bound of $\text{BF}_{t\to c}$. (for $G_t = G_c$, $\text{BF}_{t\to c} = 1$) By Lemma E.2.1 and F.1.4,

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

Next, we examine $\text{BF}_{c\to a}$. Based on Lemma F.1.2 and its proof, we divide it into two parts, i.e deletion cases and the reciprocal of addition cases. For 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 addition 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.

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

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,

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

When $G_t \subsetneq G_a$, by Lemma E.2.1 and F.1.4 we have

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

F.3 Proof of Theorem 3.4.2

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

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

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 F.1.2, 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 F.1.4. Let

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

Since $0 < \beta^* < \frac{1}{2}$ and Assumption 3.4.5, by Lemma D.2.2, when $n \to \infty$,

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

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 F.1.4. (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 3.4.5, by Lemma D.2.2, when $n \to \infty$,

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

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 3.4.1, 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 Lemma E.2.1 and F.1.4,

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

Similar to the proof of Theorem 3.4.1, we have

$$\begin{aligned}
\mathrm{BF}_{c\to a} = & \prod_{i=1}^{|E_c|-|E_a|} \mathrm{BF}(\overline{G}_i^{c\to a}; \overline{G}_{i-1}^{c\to a}) \\
< & \big\{2p(n+1)\big\}^{\frac{|E_c|-|E_a|}{2}} \big(1-\hat{\rho}^2_{\overline{x}_*\overline{y}_*|V\setminus\{\overline{x}_*,\overline{y}_*\}}\big)^{\frac{n}{2}}, \text{ when } p > b \\
< & \big\{2p(n+1)\big\}^{\frac{|E_c|-|E_a|}{2}} \exp\Big(-\frac{n\rho_L^2}{2} + n\epsilon_2 - \frac{n\epsilon_2^2}{2}\Big) \\
< & \big\{2p(n+1)\big\}^{\frac{|E_c|-|E_a|}{2}} \exp\Big\{-\frac{n\rho_L^2}{2} + \frac{n}{(n-p)^{\beta^*}} - \frac{1}{2}n^{1-2\beta^*}\Big\} \\
< & \exp\Big\{-\frac{n\rho_L^2}{2} + \frac{n}{(n-p)^{\beta^*}} - \frac{1}{2}n^{1-2\beta^*} + 3p^2\log n\Big\}.
\end{aligned}$$

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

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

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

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

By the construction of β^*, we have

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

and $1-2\lambda > \sigma+\gamma$. Therefore, $-n\rho_L^2/2$ is the leading term in the upper bound of PR$(G_a; G_t \mid G_t \not\subset G_a)$. Thus, 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 Lemma E.2.1 and F.1.4 we have

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

So

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

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

F.4 Proof of Theorem 3.4.3

From $\gamma > \alpha$, we have $\frac{1-\gamma}{2} < \frac{1-\alpha}{2}$; from $\gamma > 1-4\alpha$, we have $\frac{1-\gamma}{2} < 2\alpha$; from $\lambda < \frac{1}{2}(1-3\alpha)$, we have $\alpha+\lambda < \frac{1-\alpha}{2}$; from $\lambda < \alpha$, we have $\alpha+\lambda < 2\alpha$. For any $\beta^{\#}$ satisfies

$$\max\Big\{\alpha+\lambda, \frac{1-\gamma}{2}\Big\} < \beta^{\#} < \min\Big\{\frac{1-\alpha}{2}, 2\alpha\Big\},$$

let $\epsilon_{3,n} = (n-p)^{-\beta^{\#}}$. Then define

$$R'''_{ij|S} = \big\{|\hat{\rho}_{ij|S} - \rho_{ij|S}| < \epsilon_{3,n}\big\}.$$

Denote

$$\Delta''_{\epsilon_3}(n) = \Big\{ \cap_{(i,j)\in E_t} R'''_{ij|V\setminus\{i,j\}} \Big\} \bigcap \Big\{ \cap_{(i,j)\notin E_t} \big(\cap_{S\in\Pi_{ij}} R'''_{ij|S} \big) \Big\}.$$

Since $0 < \alpha < \frac{1}{3}$, thus $0 < \beta^{\#} < \frac{1-\alpha}{2} < \frac{1}{2}$. By Assumption 3.4.5 and Lemma D.2.3,

$$\mathbb{P}\big\{\Delta''_{\epsilon_3}(n)\big\} \to 1, \text{ as } n \to \infty.$$

For any decomposable graph G_a, there exists a set $\Delta_{a,\epsilon_3}(n)$ defined in Theorem 3.4.2, such that $\Delta''_{\epsilon_3}(n) \subset \Delta_{a,\epsilon_3}(n)$. For the following proof, we restrict it to the event $\Delta''_{\epsilon_3}(n)$. Thus, the upper bound of Bayes factors derived under $\Delta''_{\epsilon_3}(n)$ is a uniform upper bound for all decomposable graphs that are not G_t. Following the proof of Theorem 3.4.2, when $G_t \not\subset G_a$,

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

By the construction of $\beta^{\#}$, we have

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

and $1 - 2\lambda > \gamma + \sigma$. Therefore, $-n\rho_L^2/2$ is the leading term in the upper bound of $\text{PR}(G_a; G_t \mid G_t \not\subset G_a)$. For simplicity, only the leading term is used in the following calculation.

When $G_t \subsetneq G_a$,

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

Since $\beta^{\#} > \frac{1-\gamma}{2}$, then $\big(|E_a| - |E_t|\big)\log(2q)$ is the leading term above and $|E_a| - |E_t| > 0$. Thus, when n is sufficiently large, for any decomposable graph $G_a \neq G_t$, we have

$$\begin{aligned} &\text{PR}(G_a; G_t \mid G_t \not\subset G_a) < \exp\Big(- D_1 n \rho_L^2 \Big), \\ &\text{PR}(G_a; G_t \mid G_t \subsetneq G_a) < \exp\Big\{ - D_2 n^{\gamma} \big(|E_a| - |E_t|\big) \Big\}, \end{aligned}$$

where D_1 and D_2 are two positive finite constants.

$$\begin{aligned} \sum_{G_t \not\subset G_a} \text{PR}(G_a; G_t) &= \sum_{|E_a^1|=0}^{|E_t|-1} \binom{|E_t|}{|E_a^1|} \sum_{|E_a|-|E_a^1|=0}^{|E_c|-|E_t|} \binom{|E_c| - |E_t|}{|E_a| - |E_a^1|} \text{PR}(G_a; G_t \mid G_t \not\subseteq G_a) \\ &< \exp(p^2 \log 2) \exp(-D_1 n \rho_L^2) \to 0, \ \text{as } n \to \infty. \end{aligned}$$

$$\begin{aligned} \sum_{G_t \subsetneq G_a} \text{PR}(G_a; G_t) &= \sum_{|E_a|=|E_t|+1}^{|E_c|} \binom{|E_c| - |E_t|}{|E_a| - |E_t|} \text{PR}(G_a; G_t \mid G_t \subsetneq G_a) \\ &< \sum_{i=1}^{|E_c|-|E_t|} \binom{|E_c| - |E_t|}{i} \big(e^{-D_2 n^{\gamma}}\big)^i \\ &= (1 + e^{-D_2 n^{\gamma}})^{|E_c|-|E_t|} - 1 \\ &< \exp\big\{\big(|E_c| - |E_t|\big) e^{-D_2 n^{\gamma}}\big\} - 1 \to 0, \ \text{as } n \to \infty. \end{aligned}$$

(i) When $G_t = G_0$, where G_0 is the null graph with no edges.

$$\sum_{G_a \neq G_0} \text{PR}(G_a; G_0) = \sum_{G_0 \subsetneq G_a} \text{PR}(G_a; G_0) \to 0, \ \text{as } n \to \infty;$$

(ii) When $G_t \neq G_0$ and $G_t \neq G_c$,

$$\sum_{G_a \neq G_t} \text{PR}(G_a; G_t) = \sum_{G_t \not\subset G_a} \text{PR}(G_a; G_t) + \sum_{G_t \subsetneq G_a} \text{PR}(G_a; G_t) \to 0, \ \text{as } n \to \infty;$$

(iii) When $G_t = G_c$,

$$\sum_{G_a \neq G_c} \text{PR}(G_a; G_c) = \sum_{G_c \not\subseteq G_a} \text{PR}(G_a; G_c) \to 0, \text{ as } n \to \infty.$$

Therefore,

$$\pi(G_t \mid \text{Y}) = \frac{1}{1 + \sum_{G_a \neq G_t} \text{PR}(G_a; G_t)} \to 1, \text{ as } n \to \infty.$$

F.5 Proof of Corollary 3.4.2

According to the proof of Theorem 3.4.3, in the set $\Delta''_{\epsilon_3}(n)$, all Bayes factors in favor of G_a converge to zero uniformly. Thus, we have

$$\mathbb{P}\Big\{ \max_{G_a \neq G_t} \pi(G_a \mid \text{Y}) < \pi(G_t \mid \text{Y}) \Big\} \to 1, \quad \text{as } n \to \infty.$$

Therefore,

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

APPENDIX G

PROOFS UNDER MODEL MISSPECIFICATION [*]

G.1 Preparation

Let $G_m = (V, E_m)$ be any minimal triangulation of G_t, where $E_m = E_t \cup F$, $F \neq \emptyset$. 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 F.1.2, $\{\overline{x}_i, \overline{y}_i\}_{i=1}^{|E_c|-|E_a|}$ 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 F.1.2 still holds here, i.e. at least one population partial correlation in $\{\rho_{\overline{x}_i\overline{y}_i|\overline{S}_i}\}_{i=1}^{|E_c|-|E_a|}$ corresponding to the removal of a true edge is non-zero and it is not a correlation. The proof carries out the same as in Lemma F.1.2, just let the first step of moving from G_c to G_a be the deletion of one 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 the reciprocal of addition cases. There is no deletion case here since G_a has all the true edges in 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 F.1.4 still holds here.

Lemma G.1.1. *For any edge sequence $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{|E_a|-|E_m|}$ from G_m to G_a describe above, all pop-*

[*] Reprinted with permission from arXiv.org, "Bayesian Graph Selection Consistency Under Model Misspecification", arXiv preprint arXiv:1901.04134, 2019, by Niu, Yabo and Pati, Debdeep and Mallick, Bani K. In accordance arXiv copyright no modifications have been made except formatting.

ulation partial correlations in $\{\rho_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i}\}_{i=1}^{|E_a|-|E_m|}$ *are zero.* (*or correlation, when* $\widetilde{S}_i = \emptyset$)

Proof. This proof follows similarly to the proof of Lemma F.1.4. 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 $\widetilde{S}_i \neq \emptyset$. Since adding edge $(\widetilde{x}_i, \widetilde{y}_i) \notin E_t$ to graph $\widetilde{G}_{i-1}^{m\to a}$ maintains the decomposability of graph $\widetilde{G}_i^{m\to a}$. By Lemma E.2.1, $\widetilde{x}_i$ 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 F.1.3, $\widetilde{S}_i$ separates node $\widetilde{x}_i$ from $\widetilde{y}_i$ in G_t, $\rho_{\widetilde{x}_i\widetilde{y}_i|\widetilde{S}_i} = 0$.

When $\widetilde{S}_i = \emptyset$, $\widetilde{x}_i$ 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 G.1.1. *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 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.*

G.2 Proof of Theorem 3.5.1

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 3.4.1. 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

$$\begin{aligned}
\mathrm{BF}_{m\to c} &< \exp\left\{p^2 - \left(\frac{1}{2} - \frac{1}{\tau^*}\right)(|E_c| - |E_m|)\log n\right\} \to 0,\\
\mathrm{BF}_{c\to a} &< \exp\left\{ - \frac{n\rho_L^2}{2} + p^2\log n + \sqrt{n\log n} - \frac{1}{2\tau^*}\log(n-p) + 2p^2\log p\right\} \to 0.
\end{aligned}$$

Hence,

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

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

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

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

$$\mathrm{BF}_{m\to c} < 2^{p^2} n^{-\frac{|E_c|-|E_a|}{2}} \exp\Bigg[-\Bigg\{\frac{|E_a| - |E_m|}{2(|E_c| - |E_m|)} - \frac{1}{\tau^*}\Bigg\}(|E_c| - |E_m|)\log n\Bigg],$$
$$\mathrm{BF}_{c\to a} < (4p)^{p^2} n^{\frac{|E_c|-|E_a|}{2}}, \text{ when } n > 1.$$

Hence,

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

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

Part 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 G.1.1, $\rho_{m_1,i} = 0$, $i = 1, 2, \ldots, |E_c| - |E_{m_1}|$. By Lemma D.2.4, 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 D.2.4), we have $\mathbb{P}(\Delta_\epsilon^0) > 1 - \epsilon/2$, for $n > p + 3$. Let

$$R_{m_1,i} = \Big\{\frac{M_1}{n} < \hat{\rho}^2_{m_1,i} < \frac{M_2}{n-p}\Big\},$$

and denote

$$\Delta_{m_1} = \bigcap_{i=1}^{|E_c|-|E_{m_1}|} R_{m_1,i}.$$

Then

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

By Lemma E.2.1, when $n > b + p$, we have

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

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

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

Thus we have

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

Similarly,

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

Therefore, let $A_1 = \frac{1}{4}e^{-M_2}$ and $A_2 = 4e^{2M_2p^2}$,

$$\mathbb{P}\big\{A_1 < \text{BF}(G_{m_1}; G_{m_2}) < A_2\big\} > 1 - \epsilon.$$

Part 3. 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},

$$\text{BF}(G_{m_i}; G_a) \to \infty, \quad i = 1, 2, \ldots, l,$$

where $G_a \notin \mathcal{M}_t$. Therefore,

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

G.3 Proof of Theorem 3.5.2

Part 1. From $\gamma > 1 - 2\alpha$, we have $\frac{1-\gamma+2\alpha}{2} < 2\alpha$; from $\lambda < \frac{1}{2} - \alpha$, we have $\alpha + \lambda < \frac{1}{2}$; from $\lambda < \alpha$, we have $\alpha + \lambda < 2\alpha$; from $\gamma > 2\alpha$, we have $\frac{1-\gamma+2\alpha}{2} < \frac{1}{2}$. Let β^* satisfy

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

then follow the construction of $\Delta_{a,\epsilon_2}(n)$ in the proof of Theorem 3.4.2 using β^* specified above. After that, we restrict the following proof to the set $\Delta_{a,\epsilon_2}(n)$. For case (1), when $|E_a^1| < |E_m^1| = |E_t|$, by the construction of β^*, we have

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

and $1 - 2\lambda > \sigma + \gamma$. Thus,

$$\begin{aligned}
&\text{PR}(G_a; G_m \mid G_m \not\subset G_a, |E_a^1| < |E_m^1|) \\
&< \exp\Big\{ -\frac{n\rho_L^2}{2} + \frac{n}{(n-p)^{\beta^*}} - \frac{1}{2} n^{1-2\beta^*} + \frac{2np^2}{(n-p)^{2\beta^*}} + \big(|E_a| - |E_m|\big)\log(2q)\Big\} \to 0.
\end{aligned}$$

For case (2), when $G_m \subsetneq G_a$, i.e. $|E_a^1| = |E_m^1| = |E_t|$ and $|E_a| > |E_m|$, since $\beta^* > \frac{1-\gamma}{2}$, we have

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

For case (3), when $|E_a^1| = |E_m^1| = |E_t|$ and $G_m \not\subset G_a$, also $|E_a| > |E_m|$, since $\beta^* > \frac{1-\gamma+2\alpha}{2}$, we have

$$\begin{aligned}&\mathrm{PR}(G_a; G_m \mid G_m \not\subset G_a, |E_a^1| = |E_m^1|)\\ &\qquad < \{2p(n+1)\}^{\frac{|E_c|-|E_a|}{2}} \exp\left\{\frac{(|E_c| - |E_m|)n}{(n-p)^{2\beta^*}} + \big(|E_a| - |E_m|\big)\log(2q)\right\} \to 0.\end{aligned}$$

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

Part 2. Since the number of fill-in edges is finite, then the number of cycles length greater than 3 without a chord in G_t is finite and the length of the longest cycle without a chord is also finite. Thus instead of adding one chord for each of those cycles that are length greater than 3 in G_t, we can complete the subgraphs induced by those cycles with finite number of edges. Let G_{m_c} be the graph after completing all subgraphs induced by those cycles. Then G_{m_c} is decomposable and $|E_{m_c}| - |E_t|$ is finite. We also know $G_{m_1}, G_{m_2} \subsetneq G_{m_c}$. Let $\delta_c = |E_{m_c}| - |E_{m_1}| = |E_{m_c}| - |E_{m_2}|$.

Let $\{\hat{\rho}_{m_1,i}\}_{i=1}^{|E_{m_c}|-|E_{m_1}|}$ and $\{\rho_{m_1,i}\}_{i=1}^{|E_{m_c}|-|E_{m_1}|}$ be the sample and population partial correlation sequence corresponding to each step from G_{m_1} to G_{m_c}. By Lemma G.1.1, $\rho_{m_1,i} = 0$, $i = 1, 2, \ldots, |E_{m_c}| - |E_{m_1}|$. By Corollary D.2.2, 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 Corollary D.2.2), we have $P(\Delta_\epsilon^{0+}) > 1 - \epsilon/2$, for $n > p + 3$. Let

$$R'_{m_1,i} = \left\{\frac{M_1}{n} < \hat{\rho}^2_{m_1,i} < \frac{M_2}{n-p}\right\},$$

and denote

$$\Delta'_{m_1} = \bigcap_{i=1}^{|E_{m_c}|-|E_{m_1}|} R'_{m_1,i}.$$

Then

$$\mathbb{P}(\Delta'_{m_1}) \geq \mathbb{P}(\Delta_{\epsilon}^{0+}) \geq 1 - \epsilon/2.$$

By Lemma E.2.1, when $n > b + p$, we have

$$\Big(\frac{1}{2n}\Big)^{\frac{\delta_c}{2}} \prod_{i=1}^{\delta_c} (1 - \hat{\rho}^2_{m_1,i})^{-\frac{n\delta_c}{2}} < \text{BF}(G_{m_c}; G_{m_1}) < \Big(\frac{2}{n}\Big)^{\frac{\delta_c}{2}} \prod_{i=1}^{\delta_c} (1 - \hat{\rho}^2_{m_1,i})^{-\frac{n\delta_c}{2}}.$$

Under the event Δ'_{m_1}, when $n > p + M_2$,

$$\Big(\frac{e^{M_1}}{2n}\Big)^{\frac{\delta_c}{2}} < \text{BF}(G_{m_c}; G_{m_1}) < \Big(\frac{2e^{2M_2}}{n}\Big)^{\frac{\delta_c}{2}}.$$

Thus we have

$$\mathbb{P}\Bigg\{\Big(\frac{e^{M_1}}{2n}\Big)^{\frac{\delta_c}{2}} < \text{BF}(G_{m_c}; G_{m_1}) < \Big(\frac{2e^{2M_2}}{n}\Big)^{\frac{\delta_c}{2}}\Bigg\} > 1 - \frac{\epsilon}{2}.$$

Similarly,

$$\mathbb{P}\Bigg\{\Big(\frac{2e^{2M_2}}{n}\Big)^{-\frac{\delta_c}{2}} < \text{BF}(G_{m_2}; G_{m_c}) < \Big(\frac{e^{M_1}}{2n}\Big)^{-\frac{\delta_c}{2}}\Bigg\} > 1 - \frac{\epsilon}{2}.$$

Therefore, let $A_1 = \frac{1}{4}e^{-M_2\delta_c}$ and $A_2 = 4e^{M_2\delta_c}$,

$$\mathbb{P}\big\{A_1 < \text{BF}(G_{m_1}; G_{m_2}) < A_2\big\} > 1 - \epsilon.$$

Part 3. From $\gamma > 1 - 2\alpha$, we have $\frac{1-\gamma+2\alpha}{2} < 2\alpha$; from $\lambda < \frac{1-3\alpha}{2}$, we have $\alpha + \lambda < \frac{1-\alpha}{2}$; from $\lambda < \alpha$, we have $\alpha + \lambda < 2\alpha$; from $\gamma > 3\alpha$, we have $\frac{1-\gamma+2\alpha}{2} < \frac{1-\alpha}{2}$. Let β^* satisfy

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

then follow the construction of $\Delta''_{\epsilon_3}(n)$ in the proof of Theorem 3.4.3 using β^* specified above. After that, we restrict the following proof to the set $\Delta''_{\epsilon_3}(n)$. Let $G_{m_1}, G_{m_2}, \ldots, G_{m_h}$ be all the minimal triangulations of G_t, where h is a positive integer that depends on n. By Part 1, we have

$$\begin{aligned}
\text{PR}(G_a; G_m \mid G_m \not\subset G_a, |E_a^1| < |E_m^1|) &< \exp\Big(- D_1 n \rho_L^2 \Big), \\
\text{PR}(G_a; G_m \mid G_m \subsetneq G_a) &< \exp\Big\{ - D_2 n^\gamma \big(|E_a| - |E_m| \big) \Big\}, \\
\text{PR}(G_a; G_m \mid G_m \not\subset G_a, |E_a^1| = |E_m^1|) &< \exp\Big\{ - D_3 n^\gamma \big(|E_a| - |E_m| \big) \Big\},
\end{aligned}$$

where D_1, D_2, D_3 are three positive finite constants. And

$$\sum_{\substack{G_a \notin \mathcal{M}_t, \\ G_{m_1} \not\subset G_a, \\ |E_a^1| < |E_{m_1}^1|}} \text{PR}(G_a; G_{m_1}) < \exp(p^2) \exp\big(- D_1 n \rho_L^2 \big) \to 0,$$

$$\begin{aligned}
\sum_{\substack{G_a \notin \mathcal{M}_t, \\ G_{m_1} \subsetneq G_a}} \text{PR}(G_a; G_{m_1}) &< \sum_{i=1}^{|E_c| - |E_{m_1}|} \binom{|E_c| - |E_{m_1}|}{i} \big(e^{-D_2 n^\gamma} \big)^i \\
&< \exp\big\{ (|E_c| - |E_{m_1}|) e^{-D_2 n^\gamma} \big\} - 1 \to 0,
\end{aligned}$$

$$\begin{aligned}
\sum_{\substack{G_a \notin \mathcal{M}_t, \\ G_{m_1} \not\subset G_a, \\ |E_a^1| = |E_{m_1}^1|}} \text{PR}(G_a; G_{m_1}) &< \sum_{i=1}^{|E_c| - |E_{m_1}|} \binom{|E_c| - |E_{m_1}^1|}{|E_{m_1}| - |E_{m_1}^1| + i} \big(e^{-D_3 n^\gamma} \big)^i \\
&< \exp(p^2) \exp\big(- D_3 n^\gamma \big) \to 0.
\end{aligned}$$

Thus

$$\begin{aligned}
&\sum_{G_a \notin \mathcal{M}_t} \frac{1}{\sum_{i=1}^h \text{PR}(G_{m_i}; G_a)} < \sum_{G_a \notin \mathcal{M}_t} \frac{1}{\text{PR}(G_{m_1}; G_a)} \\
&= \sum_{\substack{G_a \notin \mathcal{M}_t, \\ G_{m_1} \subsetneq G_a}} \text{PR}(G_a; G_{m_1}) + \sum_{\substack{G_a \notin \mathcal{M}_t, \\ G_{m_1} \not\subset G_a}} \text{PR}(G_a; G_{m_1}) + \sum_{\substack{G_a \notin \mathcal{M}_t, \\ G_{m_1} \not\subset G_a, \\ |E_a^1| = |E_{m_1}^1|}} \text{PR}(G_a; G_{m_1}) \to 0.
\end{aligned}$$

Therefore,

$$\sum_{G_m \in \mathcal{M}_t} \pi(G_m \mid \mathrm{Y}) = \frac{1}{1 + \sum_{G_a \notin \mathcal{M}_t} \frac{1}{\sum_{i=1}^{h} \mathrm{PR}(G_{m_i}; G_a)}} \to 1, \quad \text{as } n \to \infty.$$

G.4 Proof of Corollary 3.5.1

Under the event $\Delta''_{\epsilon_3}(n)$ in the proof of Theorem 3.5.2, given any $G_m \in \mathcal{M}_t$, all Bayes factors in favor of G_a converge to zero uniformly. Thus, we have

$$\mathbb{P}\Big\{ \max_{G_a \notin \mathcal{M}_t} \pi(G_a \mid \mathrm{Y}) < \min_{G_m \in \mathcal{M}_t} \pi(G_m \mid \mathrm{Y}) \Big\} \to 1, \quad \text{as } n \to \infty.$$

Therefore,

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