BAYESIAN MODEL SELECTION FOR DISCRETE GRAPHICAL MODELS

LYNDSAY ROACH

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

JUNE 2023

Abstract

Graphical models allow for easy interpretation and representation of complex distributions. There is an expanding interest in model selection problems for high-dimensional graphical models, particularly when the number of variables increases with the sample size. A popular model selection tool is the Bayes factor, which compares the posterior probabilities of two competing models. Consider data given in the form of a contingency table where N objects are classified according to q random variables, where the conditional independence structure of these random variables are represented by a discrete graphical model G. We assume the cell counts follow a multinomial distribution with a hyper Dirichlet prior distribution imposed on the cell probability parameters. Then we can write the Bayes factor as a product of gamma functions indexed by the cliques and separators of G.

In this thesis, we study the behaviour of the Bayes factor when the dimension of a true discrete graphical model is fixed and when the dimension increases to infinity with the sample size. We prove that the Bayes factor is strong model selection consistent for both decomposable and non-decomposable discrete graphical models. When the true graph is non-decomposable, we prove that the Bayes factor selects a minimal triangulation of the true

graph. We support our theoretical results with various simulations.

In addition, we introduce a variation of the genetic algorithm, called the graphical local genetic algorithm, which can be implemented on large data sets. We use a local search operator and a normalizing constant proportionate to the posterior probability of the candidate models to determine optimal submodels, then reconstruct the full graph from the resulting subgraphs. We demonstrate the graphical local genetic algorithm's capabilities on both simulated data sets with known true graphs and on a real-world data set.

Acknowledgements

I would like to thank my supervisors, Professor Xin Gao and Professor Hélène Massam. Professor Xin Gao for sharing her extensive knowledge and for her constant words of encouragement. Professor Hélène Massam for having been an exceptional role model for myself and the statistics community at large.

I am grateful to the members of the Department of Mathematics and Statistics for a memorable experience during my time at York University. Thank you to Professor Nanwei Wang and Professor Kevin McGregor for taking the time to be on my thesis committee and for giving invaluable feedback.

Thank you to my family and friends for their endless support throughout my education.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Discrete graphical models which exhibit specific properties are a subset of the class of hierarchical log-linear models. Hierarchical log-linear models are used to analyse data given in the form of a contingency table with N objects classified according to a set of q criteria. Consider a vector of random variables $X = (X_v, v \in V)$ indexed by the set $V = \{1, 2, \ldots, q\}$ such that each X_v takes values in the finite set I_v. The resulting counts for each classification can be given in the form of a contingency table corresponding to

$$I = \bigtimes_{v \in V} I_v,$$

where I is the set of cells $i = (i_v, v \in V)$. The number of observations for cell i is denoted $n(i)$ and the probability of an object being observed in cell i is denoted $p(i)$. If $D \subset V$, the D-marginal table is the set of D-marginal cells $i_D = (i_v, v \in D)$. Given the marginal cell $i_D \in I_D$, we write the D-marginal cell count as $n_D(i_D) = \sum_{i' \in I; i_D = i'_D} n(i')$. For $N = \sum_{i \in I} n(i)$, we

assume the cell counts $(n(i), i \in I)$ follow a multinomial distribution and the cell probabilities are modelled by a hierarchical log-linear model.

Let Δ be a nonempty collection of subsets of V such that $\bigcup_{D\in\Delta} D = V$, and if $D \in \Delta$, $D_1 \subset D$, and $D_1 \neq \varnothing$, then $D_1 \in \Delta$. The collection of subsets Δ is called the generating class of the model. We arbitrarily select an element in each I_v and denote it by 0. Then we can impose the *baseline* constraints, meaning for $D \in \Delta$, if $i_v = 0$ for some $v \in D$, then $\theta_D(i_D) = 0$. Thus, we have the unique representation

$$\log p(i) = \theta_\varnothing + \sum_{D\in\Delta, i_v\neq 0, \forall v\in D} \theta_D(i_D),$$

where $\theta_\varnothing$ is a constant not depending on i. We give a full description of hierarchical log-linear models in Section 2.2.2.

A discrete graphical model for $X = (X_v, v \in V)$ is a representation of the conditional independencies between the random variables X_v using an undirected graph $G = (V, E)$ with vertex set V and edge set $E \subseteq V \times V$. A discrete graphical model is said to be decomposable or Markov with respect to G, if X_a is independent of X_b given $X_{V\setminus\{a,b\}}$, whenever (a, b) is not an edge in E. If the random vector X is Markov with respect to G, we can write the distribution of X as a decomposition of smaller components as follows,

$$p(x) \propto \prod_{C\in\mathcal{C}} \phi_C(x),$$

where $\mathcal{C}$ is a particular set of subgraphs of G called cliques. Being able to express models

according to an appropriate decomposable graph G makes it a convenient class of model to work with because it allows for useful closed-form expressions and for efficient computing. In addition, this property results in the decomposable chain rule, meaning one can construct either an increasing or decreasing sequence of decomposable graphs differing by one edge (Lauritzen, (1996)). The decomposable chain rule is a practical tool for comparing candidate models when selecting which variables to include. Consequently, model selection methods often restrict their search to decomposable models. More background definitions from graph theory are provided in Section 2.2.1.

The discrete graphical models which are decomposable correspond to the class of hierarchical log-linear models. They have applications spanning many disciplines and are often used for various machine learning applications, such as disease diagnostics and image recognition. As the technology for collecting and storing data improves, there is an increasing demand for exploring model selection problems in a high-dimensional setting, that is, when the number of variables increases with the sample size. An area of particular interest is determining a reliable model selection criterion. The different criteria for selecting a model consider aspects such as the fit of the model to the data and complexity of the model.

In the frequentist setting, it is common to use a penalized likelihood type of criterion. For example, Wainwright et al. (2007) and Ravikumar et al. (2010) propose a method based on ℓ_1-regularized logistic regression, where they show consistency for estimating the neighbourhood of every node in the graph simultaneously of an associated binary Ising model. In the case of continuous graphical models, Raskutti et al. (2009) give the sufficient

conditions for model selection consistency of ℓ_1-regularized Gaussian maximum likelihood. Meinshausen and Bühlmann (2006) demonstrate that the neighbourhood estimate with the lasso, introduced by Tibshirani (1996), converges to the true neighbourhood and that this method is an appropriate alternative to standard covariance selection for sparse high-dimensional graphs.

In the Bayesian setting, the Bayes factor is commonly used for model selection. It compares the posterior distribution of the data under two different models and indicates the support for one model over the other, regardless of whether either model is correct. For comparing two models, say G_1 and G_2, the pairwise Bayes factor is the ratio

$$BF_{G_1,G_2} = \frac{f(G_1|x)}{f(G_2|x)},$$

where $f(G|x)$ is the posterior probability of a graph G given data x. Fitch et al. (2014) focus on Gaussian graphical models with the hyper-inverse Wishart prior, which is the Diaconis-Ylvisaker conjugate prior for the Gaussian distribution. They study the behaviour of the Bayes factor for a fixed number of variables and they prove that when the true graph is non-decomposable, model selection procedures will favour a minimal triangulation of the true graph. They prove that the logarithm (log) of the Bayes factor between two minimal triangulations with the same number of edges is stochastically bounded. A minimal triangulation of a non-decomposable graph is a decomposable graph obtained from added the minimum number of edges to the non-decomposable graph to make it decomposable. Note that minimal triangulations are not necessarily unique. This result from Fitch et al.

(2014) is valuable because it allows one to exploit the decomposable property of a minimal triangulation. Therefore, when the true graph is non-decomposable it can be approximated by its minimal triangulation(s).

In Niu et al. (2021), they prove results analogous to those in Fitch et al. (2014), but for an increasing number of variables and they prove that strong model selection consistency holds in the high-dimensional setting both when the true graph is decomposable and non-decomposable. In order to do so, they reduce the Bayes factor to local moves comparing two graphs differing by only one edge. They make the distinction between an overfitting model and an underfitting model. When a candidate model is an overfitting model, it means that it contains all the edges of the true model plus at least one false edge and when a candidate model is an underfitting model, it means that it is missing at least one edge from the true model. Niu et al. (2021) conclude that in the underfitting case, the Bayes factor converges at a faster rate that in the overfitting case. This means that in the continuous setting, the Bayes factor gives a stronger penalty to a missing true edge over an additional false edge.

The main topic of this thesis is proving strong model selection consistency for the Bayes factor when the true discrete graphical model is non-decomposable and q increases to infinity. For discrete decomposable models, the Bayes factor can be written in a closed-form as a product of gamma functions. We define the Bayes factor and the assosiated Diaconis-Ylvisaker prior for graphical log-linear models in Section 2.3. We assume the cell probability parameters follow a Dirichlet prior distribution, thus for two decomposable graphs G_1 and G_2, we can write the corresponding Bayes factor as a product of gamma functions. Our theoretical

results are developed from our asymptotic approximation of the normalizing constant which is proportionate to the posterior probability, using known properties of the gamma function. We study the behaviour of the Bayes factor between competing decomposable models and show that a true non-decomposable model can be approximated by a minimal triangulation. Using these intermediate results, we are able to prove the desired strong model selection consistency results.

Graphical model selection often consists of forward or backward elimination procedures on decomposable graphs due to the decomposable chain rule. However, for q variables there are 2^q possible models, thus these methods become computationally intensive for high-dimensional data. Petijean et al. (2013) put forward an approach that they call *Chordalysis*, which is a forward selection method where they use data mining techniques to store and reuse the computed marginal likelihood ratios. Their method is effective; however, the efficiency of their algorithm relies on the decomposable property of the candidate graphs and cannot return a non-decomposable graph. Dobra and Mohammadi (2018) implement a Birth-Death Markov Chain Monte Carlo (BDMCMC) algorithm, which they speed up by computing all of the possible edges using parallel computing. They avoid being restricted to decomposable graph by using a marginal posterior probability based on the marginal pseudo-likelihood.

The second topic of this thesis is proposing model selection algorithm for high-dimensional discrete graphical models. We introduce a variation of the genetic algorithm with a crossover-hill-climbing operator (Lozano et al., (2004)) for high-dimensional log-linear graphical models, called the graphical local genetic algorithm. We use the log of the normalizing constant

proportionate to posterior probability to measure the appropriateness of the candidate models. If the candidate model is decomposable, we compute the log of the normalizing constant directly, and if the candidate model is non-decomposable, we compute the log of the normalizing constant of its minimal triangulation. This allows the algorithm more flexibility than forward or backward elimination procedures. We use simulation results to show the flexibility of our algorithm, and we use our algorithm to analyse the real-world *Movies Dataset*.

The remainder of this thesis is organized as follows. In Chapter 2, we give the preliminary terminology from graph theory and hierarchical log-linear models and we prove strong model selection consistency for the Bayes factor when the true discrete graphical model is non-decomposable and q increases to infinity. In Chapter 3, we introduce a model selection algorithm for high-dimensional data and provide our experimental results. We give concluding remarks and suggest future work in Chapter 4.

Chapter 2

Bayesian Model Selection Consistency for Discrete Graphical Models

The Bayes factor is a popular method of model selection that compares the posterior probabilities of two competing models. Consider data given in the form of a contingency table where N objects are classified according to q random variables and the conditional independence structure of these random variables are represented by a discrete graphical model. We assume the cell counts follow a multinomial distribution with a hyper Dirichlet prior distribution imposed on the cell probability parameters. We examine the behaviour of the Bayes factor when the dimension of the model is fixed and when the dimension increases to infinity with the sample size. Our main result is proving strong model selection consistency for increasing dimension both when the true graph is decomposable and when the true graph is non-decomposable. When the true graph is non-decomposable, we prove that the Bayes

factor selects a minimal triangulation of the true graph.

2.1 Introduction

Graphical models allow for easy interpretation and representation of complex distributions. There is an expanding interest in model selection problems for high-dimensional graphical models, particularly when the number of variables increases with the sample size. In the following, we focus on discrete graphical models for data given in the form of a contingency table with N objects classified according to a set of q criteria. Consider a vector of random variables $X = (X_v, v \in V)$ indexed by the set $V = \{1, 2, \ldots, q\}$ such that each X_v takes values in a q-dimensional contingency table. Then the conditional independencies between the random variables X_v can be read off an undirected graph $G = (V, E)$ with vertex set V and edge set $E \subseteq V \times V$. The discrete graphical model for X is said to be decomposable or Markov with respect to G, if X_a is independent of X_b given $X_{V\setminus\{a,b\}}$, whenever (a, b) is not an edge in E. We assume the cell counts of the contingency table follow a multinomial distribution and the cell probabilities are modelled by a hierarchical log-linear model. The class of discrete graphical models which are Markov with respect to an undirected graph G is a subclass of the class of hierarchical log-linear models.

In this chapter, we examine the conditions required for Bayesian model selection consistency when the true model is non-decomposable. We concentrate on the behaviour of the Bayes factor between competing decomposable models and show that a non-decomposable model can be approximated by a suitable decomposable model. Although the class of decomposable graphs

can be considered limited, model selection problems are often restricted to decomposable graphs due to the convenient computational properties and the scalability of algorithms. The posterior probabilities become straightforward to calculate and it is possible to construct an increasing sequence of decomposable graphs with q vertices, differing by one edge, which is called the decomposable graph chain rule (Lauritzen, 1996).

In the Bayesian setting, the Bayes factor is commonly used for model selection. It compares the posterior distribution of the data under two different models and indicates the support for one model over the other, regardless of whether either model is correct. For decomposable graphs, the Bayes factor can be computed explicitly; however, this calls for a tractable family of prior distributions. Dawid and Lauritzen (1993) developed the hyper Markov laws for decomposable graphs which extend the Markov properties from the random variables to the probability distribution over the set of probability measures. For our purposes, the hyper Dirichlet on the cell probabilities parameter is of particular interest because it is the conjugate prior to a multinomial distribution. In our proofs, we primarily use the parametrization with respect to the cell probabilities; however, in some instances, the log-linear parametrization is more convenient. Since the log-linear model is a natural exponential family, the prior distribution on the log-linear parameters is the Diaconis-Ylvisaker (DY) conjugate prior from Diaconis and Ylvisaker (1979). Massam et al. (2009) derived the DY conjugate prior on the log-linear parameter for graphical models and prove that it is identical to the hyper Dirichlet through a one-to-one change of variables.

Since we use the hyper Dirichlet conjugate prior, the closed-form of the posterior prob-

ability for a decomposable graph G given data x is a product of gamma functions. We present a convenient approximation for the logarithm of gamma functions, then we use our approximation to write the logarithm of an expression proportionate to the posterior probability. When both models have equal probability, the log of the Bayes factor is the difference between the log of the posterior probabilities. We use the difference of log of the posterior probabilities to investigate the asymptotic behaviour of the Bayes factor between two decomposable graphical models. Initially, we prove that the Bayes factor favours the model containing all the true edges over an underfitting model, and when both models contain all the true edges, the one with fewer excess false edges is favoured. For a fixed dimension, we can show this by simply using the Bayes factor, but for increasing dimension, we require a prior model distribution to apply a stronger penalty on the extra edges. Thus, for increasing dimension, we study the behaviour of the posterior odds ratio which is the product of the Bayes factor and the ratio of model priors. When the true graph is non-decomposable, we use a minimal triangulation of the true non-decomposable model as its proxy and we prove that the Bayes factor favours a minimum triangulation over other competing models. We show that the Bayes factor between two possible minimal triangulations with the same finite number of edges is stochastically bounded. Lastly, we simulate the behaviour of the Bayes factor to justify our theoretical results.

To the best of our knowledge, this is the only article that addresses Bayesian model selection consistency for discrete graphical models when the dimension of the model increases with the sample size. There has been previous work done regarding undirected Gaussian

graphical models from the Bayesian perspective. Fitch et al. (2014) focus on Gaussian graphical models with the hyper-inverse Wishart prior. They study the asymptotic behaviour of the marginal likelihood ratio for a fixed number of variables and they prove that when the true graph is non-decomposable, model selection procedures will favour a minimal triangulation of the true graph. Also, they prove that the log of the marginal likelihood ratio between two minimal triangulations with the same number of edges is stochastically bounded. In Niu et al. (2021), they prove analogous results to Fitch et al. (2014) where the number of variables $q = O(n^{\alpha})$ with $\alpha < 1/3$ and they prove that strong model selection consistency holds in the high-dimensional setting in both the well-specified case and the misspecified case. To do so, they reduce the Bayes factor to local moves comparing two graphs differing by only one edge and convert the Bayes factor into a function of the sample partial correlation. Then they develop sharp concentration and tail bounds for the sample partial correlation.

In the well-specified case, the hyper-inverse Wishart prior is a particular case of the so-called DAG-Wishart prior. Cao et al. (2019) prove strong model selection consistency for Gaussian directed acyclic graphical (DAG) models using the DAG-Wishart prior with multiple shape parameters when the dimension increases at a sub-exponential rate with sample size. They use the modified Cholesky decomposition $\Omega = LD^{-1}L^T$ of the inverse covariance matrix $\Omega = \Sigma^{-1}$ to parametrize their model, where L is a lower triangle matrix and D is a diagonal matrix. Then they impose a sparsity pattern on the Gaussian DAG model by putting constraints on the off-diagonal entries of L. Other examples of papers on Gaussian DAG model selection consistency from the Bayesian point of view are Cao et

al. (2020) and Lee et al. (2019). In the frequentist setting, Wainwright et al. (2007) and Ravikumar et al. (2010) focus on undirected discrete graphical models. They estimate local structures of the model using ℓ_1-regularized logistic regression on each variable given the remaining variables. They prove neighbourhood selection consistency for every vertex in the graph simultaneously with the condition that the sample size grows more quickly than $(6d^6 + 2d^5)\log d$, where d is the maximum number of adjacent vertices to the vertex under consideration. Related ℓ_1-regularization methods are often used in the literature on graphical model selection because they lead to efficient algorithms. Equivalent approaches are implemented in the study of model selection consistency for Gaussian graphical models from a frequentist perspective in Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Raskutti et al. (2009), and Gao et al. (2012), among others.

The remainder of this chapter is organized as follows. In Section 2.2, we outline the prerequisite terminology from graph theory and hierarchical log-linear models. In Section 2.3, we define the Bayes factor for decomposable models. Section 2.4 gives the theoretical results for the pairwise Bayes factor between decomposable graphs and model selection consistency when the true graph is decomposable. Section 2.5 extends the theoretical results from Section 2.4 to the case where the true graph is assumed to be non-decomposable. Then we provide simulation results in Section 3.3.1 and our conclusion in Section 2.7.

2.2 Preliminaries

In the following subsections, we cover the necessary background definitions and concepts. Section 2.2.1 contains basic notions from graph theory, which can be found in full detail in Lauritzen (1996). Sections 2.2.2 and 2.2.3 review the parametrization of the hierarchical log-linear model and the multinomial distribution expressed as a natural exponential family, respectively. These last two subsections outline our model set-up as described in Letac and Massam (2012).

2.2.1 Undirected graphs

An *undirected* graph is a pair $G = (V, E)$, where $V = \{1, \ldots, q\}$ is a non-empty set of *vertices* and $E \subseteq V \times V$ is a set of unordered pair of vertices, called *edges*. A graph is *complete* if every pair of vertices has an edge. The number of edges in a complete graph is $|E| = \binom{q}{2} = q(q-1)/2$, where $|\cdot|$ denotes the cardinality of a set. A *subgraph* is a subset of vertices and edges from $G = (V, E)$. If $A \subseteq V$, then $G_A = (A, E_A)$ is called an *induced subgraph of* G, where $E_A = E \cap (A \times A)$ is obtained by including the edges of G with endpoints in A. If a subset of G induces a complete subgraph, we call this subgraph a *clique*.

A *path* of length n from u to v is a sequence of n distinct vertices, $u = u_0, \ldots, u_n = v$, such that $(u_{i-1}, u_i) \in E$ for $i = 1, \ldots, n$. An *n-cycle* is a path of length n with $u = v$. An edge is called a *chord* if it connects two non-adjacent vertices in a cycle. An undirected graph is called *triangulated* or *chordal* if every cycle of length $n \geqslant 4$ has a chord. A subset $S \subseteq V$ is called an (u, v)-*separator* if all paths from u to v intersect S. We say S *separates* A from

B if it is an (u, v)-separator for every $u \in A$ and $v \in B$.

A triple (A, B, S) of disjoint subsets of V such that $V = A \cup B \cup S$ is called a *decomposition* if S separates A from B and S is a complete subset of V. An undirected graph is said to be *decomposable* if it is complete, or if there exists a decomposition (A, B, S) into decomposable subgraphs $G_{A\cup S}$ and $G_{B\cup S}$. Equivalently, a graph is decomposable if and only if it is triangulated. A collection of random variables $(X_v)_{v\in V}$ with associated graph G are said to be Markov relative to G if for any decomposition (A, B, S),

$$X_A \perp\!\!\!\perp X_B | X_S.$$

An important property of decomposable graphs is that their cliques form a perfect ordering. Let $C_1, \ldots, C_k$ be a sequence of cliques and $S_2, \ldots, S_k$ be a sequence of separators of an undirected graph G. The ordering $(C_1, S_2, C_2, S_3, \ldots, C_k)$ is said to be *perfect* if for all $i > 1$ there is a $j < i$ such that $S_i \subseteq C_j$ and the sets S_i are complete for all i, where $S_j = H_{j-1} \cap C_j$ and $H_{j-1} = \cup_{j=1}^{i-1} C_j$. Let $R_j = C_j \backslash H_{j-1}$, then for every j, (H_{j-1}, R_j, S_j) is a decomposition of G.

If the random vector $X = (X_v, v \in V)$ is Markov with respect to G, then by the Hammersley-Clifford theorem, we can write factorize the distribution of X as follows,

$$p(x) \propto \prod_{C\in\mathcal{C}} \phi_C(x), \tag{2.1}$$

where $\mathcal{C}$ is the set of cliques in G. Furthermore, the distribution of X can be written as a

product of *factors* indexed according to the conditional dependencies encoded in the cliques and separators of G.

Example. Figure 2.1 is a decomposable graph with cliques $C_1 = \{ab\}$, $C_2 = \{bc\}$ and separator $S_2 = \{b\}$. The ordering (C_1, S_2, C_2) is perfect since $S_2 \subseteq C_1$ and S_2 is complete. Let X_a, X_b and X_c be random variables which are Markov with respect to this decomposable graph G, then the conditional independence

$$X_a \perp\!\!\!\perp X_c | X_b$$

is encoded in G.

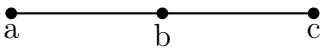


Figure 2.1: An example of a decomposable graph G with two cliques.

The joint distribution of $X = (X_a, X_b, X_c)$ can be written as a factorization indexed by the cliques and separators of G, that is,

$$p(x_a, x_b, x_c) = p(x_b)p(x_a|x_b)p(x_c|x_b) = \frac{p(x_b)p(x_a, x_b)p(x_b, x_c)}{p(x_b)p(x_b)} = \frac{p(x_a, x_b)p(x_b, x_c)}{p(x_b)}.$$

A graph G is said to be *non-decomposable* if it contains at least one chordless n-cycle of length $n \geqslant 4$. A graph $G^{\triangle} = (V, E \cup F)$ is called a *triangulation* of $G = (V, E)$ if $G^{\triangle}$ is chordal. The edges in set F are called *fill-in edges* and it is required that $E \cap F = \varnothing$. A triangulation is said to be *minimal* if $(V, E \cup F')$ is non-chordal for every $F' \subsetneq F$. Strictly speaking, a triangulation $G^{\triangle}$ is minimal if and only if the removal of any single fill-in edge from it results in a non-chordal graph (Rose et al., 1976). For more information on minimal

triangulations, see Heggernes (2006).

Example. Graph (a) is the smallest non-decomposable graph and it has three possible triangulations. Graph (d) is the complete graph on four vertices and it is a triangulation of (a); however, it is not minimal. If we remove any one edge, we obtain another decomposable graph. Graphs (b) and (c) are minimal since removing the edge (b, c) from (b), or the edge (a, d) from (c) will result in a non-decomposable graph.

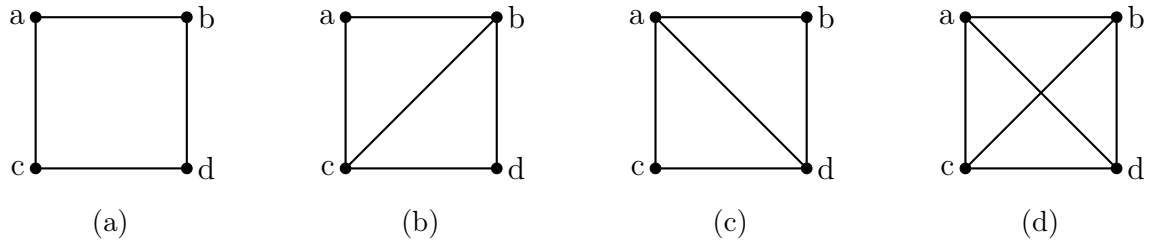


Figure 2.2: The smallest non-decomposable graph and its triangulations.

2.2.2 Hierarchical log-linear models

Let $V = \{1, 2, \ldots, q\}$ be a set of indices corresponding to q criteria. Let $X = (X_v, v \in V)$ be a vector of discrete random variables such that X_v takes values in the finite set I_v with $|I_v|$ levels. Consider N objects classified according to these q criteria, then the resulting counts can be presented in a contingency table corresponding to

$$I = \bigtimes_{v \in V} I_v,$$

where I is the set of cells $i = (i_v, v \in V)$ and $i_v \in I_v$. The number of observations for cell i is denoted $n(i)$ and the probability of an object being observed in cell i is denoted $p(i)$. If

$D \subset V$, the D-marginal table

$$I_D = \bigtimes_{v \in D} I_v$$

is the set of D-marginal cells $i_D = (i_v, v \in D)$. Given the marginal cell $i_D \in I_D$, we write the D-marginal cell count as $n_D(i_D) = \sum_{i' \in I; i_D = i'_D} n(i')$. For $N = \sum_{i \in I} n(i)$, we assume the cell counts $(n(i), i \in I)$ follow a multinomial distribution with probability density function

$$P(n(i), i \in I) = \frac{N!}{\prod_{i \in I} n(i)!} \prod_{i \in I} p(i)^{n(i)}. \tag{2.2}$$

Let Δ be a nonempty collection of subsets of V such that if $D \in \Delta$, $D_1 \subset D$, and $D_1 \neq \varnothing$, then $D_1 \in \Delta$. We assume $\bigcup_{D \in \Delta} D = V$. The collection of subsets Δ is called a simplicial complex or the generating class of the model (Letac and Massam, 2012). We denote the space of real functions $i \mapsto x(i)$ defined on I as $\mathbb{R}^I$, then let Ω_Δ be the linear subspace of $\mathbb{R}^I$ such that $x \in \Omega_\Delta$ if and only if $x = \sum_{D \in \Delta} \theta_D$, where $\theta_D \in \mathbb{R}^I$ for $D \in \Delta$ are functions depending only on i_D. This linear subspace can be written as

$$\Omega_\Delta = \left\{ x \in \mathbb{R}^I : \exists \theta_D \in \mathbb{R}^I, D \in \Delta \text{ such that } \theta_D(i) = \theta_D(i_D) \text{ and } x = \sum_{D \in \Delta} \theta_D \right\}.$$

The set of positive cell probabilities $p = (p(i))_{i \in I}$ on I such that $\log p \in \Omega_\Delta$ is the hierarchical model generated by Δ, also referred to as a multiplicative model in Darroch and Speed (1983). To guarantee a unique representation of $\log p$, we must impose a constraint on the

parameters $\theta_D(i_D)$. We arbitrarily select an element in each I_v to be the baseline level and denote it by 0. By abuse of notation, we denote also by 0 the cell in I with all its levels equal to 0. We choose the log-linear parametrization $\log p(i)$ and constrain the parameters by imposing that for $D \in \Delta$, if $i_v = 0$ for some $v \in D$, then $\theta_D(i_D) = 0$, which gives the unique representation

$$\log p(i) = \theta_\varnothing + \sum_{D\in\Delta, i_v\neq 0, \forall v\in D} \theta_D(i_D),$$

where $\theta_\varnothing$ is a constant not depending on i. Next, we adopt a more concise notation. We define

$$S(i) = \{v \in V, i_v \neq 0\}$$

to be the support of cell i, and

$$J = \{j \in I, S(j) \in \Delta\}$$

to be the subset of I corresponding to the set of free parameters. For a given $D \in \Delta$ and $\theta_D(i_D)$ such that $i_\gamma \neq 0$, $\forall \gamma \in D$, there is only one $j \in J$ such that $S(j) = D$ and $j_D = j_{S(j)} = i_D$, and conversely. Therefore, we can write

$$\theta_D(i_D) = \theta_j \text{ for the unique } j \in J \text{ with } S(j) = D, i_D = j_D. \tag{2.3}$$

To further simplify the notation we write $j \lhd i$ when $S(j) \subseteq S(i)$ and $j_{S(j)} = i_{S(j)}$. We say that j *is to the left of* i. Now we can express (2.3) in terms of the *free* parameters $\{\theta_j, j \in J\}$, which becomes

$$\log p(i) = \theta_\varnothing + \sum_{j \lhd i} \theta_j, \tag{2.4}$$

where $\theta_\varnothing$ is a unique number such that $\sum_{i \in I} p(i) = 1$. As shown in Letac and Massam (2012), by Möbius inversion theorem applied to (2.4), we are able to express the unique representation of the log-linear parameter as

$$\theta_j = \sum_{j' \in J : j' \lhd j} (-1)^{|S(j)| - |S(j')|} \log \frac{p(j')}{p(0)}. \tag{2.5}$$

Example. Let $V = \{a, b, c\}$, $\Delta = \{a, b, c, ab, bc\}$ and $I_a = I_b = I_c = \{0, 1\}$. Then the random variables X_a, X_b, and X_c can be modelled by the graph $G = (V, E)$ represented in Figure 2.1. The set of indices of the free parameters is

$$J = \{(100), (010), (001), (110), (011)\}.$$

For $i = 101$ the set of j in J such that $j \lhd i$ is $\{100, 001\}$. For $i = 111$ the set of j in J such that $j \lhd i$ is $\{100, 010, 001, 110, 011\}$ and so on. For these two cells, using the unique representation (2.4) we can write,

$$\log p(101) = \theta_0 + \theta_{100} + \theta_{001}$$

$$\log p(111) = \theta_0 + \theta_{100} + \theta_{010} + \theta_{001} + \theta_{110} + \theta_{011}.$$

For $j = 100$ the set of j' in J such that $j' \lhd j$ is $\{100\}$. For $j = 110$ the set of j' in J such that $j' \lhd j$ is $\{100, 010, 110\}$. For these two cells, using the unique parametrization (2.5) we can write,

$$\theta_{100} = \log \frac{p(100)}{p(0)}$$
$$\theta_{110} = \log \frac{p(110)p(0)}{p(100)p(010)}.$$

2.2.3 The multinomial distribution as a natural exponential family

As previously stated, we assume the cell counts for a q-dimensional contingency table with sample size N follow a multinomial distribution. In this subsection, we show that the distribution of cell counts can be written as a natural exponential family (NEF). We take the multinomial density function and employ the representation (2.4). Then the multinomial density (2.2) has the form

$$\begin{aligned}
\prod_{i\in I} p(i)^{n(i)} &= \exp\left\{\sum_{i\in I} n(i)\log p(i)\right\} \\
&= \exp\left\{\sum_{i\in I\setminus\{0\}} n(i)\log\frac{p(i)}{p(0)} + N\theta_0\right\} \\
&= \exp\left\{\sum_{i\in I\setminus\{0\}} n(i)\left(\sum_{j\in J, j\lhd i}\theta_j\right) + N\theta_0\right\} \\
&= \exp\left\{\sum_{j\in J}\theta_j\left(\sum_{i: j\lhd i} n(i)\right) + N\theta_0\right\} \\
&= \exp\left\{\sum_{j\in J}\theta_j n_{S(j)}(j_{S(j)}) + N\theta_0\right\},
\end{aligned}$$

where $n_{S(j)}(j_{S(j)})$ is the $S(j)$-marginal count. When $D = S(j)$, for $j \in J$ we write

$$t(j) = n_D(i_D). \tag{2.6}$$

For $i \in I$, we introduce vectors

$$f_i = \sum_{j\in J, j\lhd i} e_j,$$

where $(e_j)_{j\in J}$ is the canonical basis of $\mathbb{R}^J$. Let $\theta = (\theta_j, j \in J)$ be the vector of the free parameters. From the baseline constraints imposed when defining (2.4), we know $\theta_\varnothing = \log p(0)$; thus for $i \in I$, we can write

$$
\begin{aligned}
\log \frac{p(i)}{p(0)} &= \sum_{j \in J, j \lhd i} \theta_j \\
\log \frac{p(i)}{p(0)} &= \langle \theta, f_i \rangle \\
\frac{p(i)}{p(0)} &= e^{\langle \theta, f_i \rangle} \\
\sum_{i \in I} \frac{p(i)}{p(0)} &= \sum_{i \in I} e^{\langle \theta, f_i \rangle} \\
\implies p(0) &= \left(\sum_{i \in I} e^{\langle \theta, f_i \rangle} \right)^{-1},
\end{aligned}
$$

where $\langle u, v \rangle$ denotes the inner product between two vectors $u, v \in \mathbb{R}^d$. Let F be the $|I| \times |J|$ *design* matrix, where the i^{th} row is f_i^T and the superscript denotes the transpose of a vector or matrix. It is stated in Proposition 2.1 in Letac and Massam (2012) that

$$
\log \frac{p(i)}{p(0)} = (F\theta)_i.
$$

Thus, if $n = (n(i), i \in I)$ is the vector of cell counts, we can write the multinomial density (2.2) as

$$\exp\left\{\sum_{i\in I} n(i)(F\theta)_i - N\log\sum_{i\in I} e^{\langle\theta,f_i\rangle}\right\}$$
$$= \exp\left\{\langle n, F\theta\rangle - N\log\sum_{i\in I} e^{\langle\theta,f_i\rangle}\right\}$$
$$= \exp\left\{\langle F^T n, \theta\rangle - N\log\sum_{i\in I} e^{\langle\theta,f_i\rangle}\right\}.$$

Then let $t = F^T n$ denote the vector of marginal counts as defined in (2.6). Therefore, we can write the likelihood function for a multinomial distribution as

$$L(\theta) = \exp\left\{\langle\theta, t\rangle - N\log\sum_{i\in I} e^{\langle\theta,f_i\rangle}\right\}, \tag{2.7}$$

and the log-likelihood function as

$$\ell(\theta) = \langle\theta, t\rangle - N\log\sum_{i\in I} e^{\langle\theta,f_i\rangle}. \tag{2.8}$$

We denote the vector of sufficient statistics as $t = (t(j), j \in J)$, the canonical log-linear parameter as $\theta = (\theta_j, j \in J)$ and the cumulant generating function as

$$k(\theta) = \log\sum_{i\in I} e^{\langle\theta,f_i\rangle}.$$

Example. Consider the binary random variables X_a, X_b and X_c, modelled by the decomposable graph represented in Figure 2.1. Thus, we have a 3-dimensional contingency table with the set of cells

$$I = \{(000), (100), (010), (001), (110), (011), (101), (111)\}$$

and the vector of cell counts

$$n = (n(000), n(100), n(010), n(001), n(110), n(011), n(101), n(111))^T.$$

The vectors f_i, $i \in I$ corresponding to our model are

$$\begin{aligned}
f_{000} &= \Big(0,0,0,0,0\Big)^T, \\
f_{100} &= \Big(1,0,0,0,0\Big)^T, \\
f_{010} &= \Big(0,1,0,0,0\Big)^T, \\
f_{001} &= \Big(0,0,1,0,0\Big)^T, \\
f_{110} &= \Big(1,1,0,1,0\Big)^T, \\
f_{011} &= \Big(0,1,1,0,1\Big)^T, \\
f_{101} &= \Big(1,0,1,0,0\Big)^T, \\
f_{111} &= \Big(1,1,1,1,1\Big)^T.
\end{aligned}$$

Thus, we have design matrix

$$F = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

and

$$
\begin{aligned}
F^T n &= \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} n(000) \\ n(100) \\ n(010) \\ n(001) \\ n(110) \\ n(011) \\ n(101) \\ n(111) \end{pmatrix} \\
&= \begin{pmatrix} n(100) + n(110) + n(101) + n(111) \\ n(010) + n(110) + n(011) + n(111) \\ n(001) + n(011) + n(101) + n(111) \\ n(110) + n(111) \\ n(011) + n(111) \end{pmatrix} \\
&= \begin{pmatrix} t(100) \\ t(010) \\ t(001) \\ t(110) \\ t(011) \end{pmatrix}.
\end{aligned}
$$

We can write the log-likelihood function using the formulation (2.8) with sufficient statistic

$$t = (t(100), t(010), t(001), t(110), t(011))^T$$

$$= (n_a(1), n_b(1), n_c(1), n_{ab}(1,1), n_{bc}(1,1))^T,$$

canonical parameter $\theta = (\theta_{100}, \theta_{010}, \theta_{001}, \theta_{110}, \theta_{011})^T$ and cumulant generating function

$$k(\theta) = \log(1 + e^{\theta_{100}} + e^{\theta_{010}} + e^{\theta_{001}} + e^{\theta_{100}+\theta_{010}+\theta_{110}} + e^{\theta_{010}+\theta_{010}+\theta_{011}}$$

$$+ e^{\theta_{100}+\theta_{001}} + e^{\theta_{100}+\theta_{010}+\theta_{001}+\theta_{110}+\theta_{011}}).$$

Let $M(\theta) = \sum_{i\in I} e^{\langle\theta, f_i\rangle}$. For $j, m, l \lhd i$, we define the following marginal probabilities:

$$P_j(\theta) = \frac{\partial k}{\partial \theta_j} = \frac{\sum_{i\in I} e^{\langle\theta, f_i\rangle} f_{i,j}}{M(\theta)}, \tag{2.9}$$

$$P_{jm}(\theta) = \frac{\left(\sum_{i\in I} e^{\langle\theta, f_i\rangle} f_{i,j} \cdot f_{i,m}\right)}{M(\theta)}, \tag{2.10}$$

and

$$P_{jml}(\theta) = \frac{\left(\sum_{i\in I} e^{\langle\theta, f_i\rangle} f_{i,j} \cdot f_{i,m} \cdot f_{i,l}\right)}{M(\theta)}. \tag{2.11}$$

When we take the first derivative of (2.9) with respect to θ_m, we obtain the following expression which is a function of marginal probabilities, that is

$$\frac{\partial P_j(\theta)}{\partial \theta_m} = \frac{\left(\sum_{i\in I} e^{\langle \theta, f_i\rangle} f_{i,j} \cdot f_{i,m}\right) M(\theta) - \left(\sum_{i\in I} e^{\langle \theta, f_i\rangle} f_{i,j}\right) \cdot \left(\sum_{i\in I} e^{\langle \theta, f_i\rangle} f_{i,m}\right)}{[M(\theta)]^2}$$

$$= P_{jm}(\theta) - P_j(\theta) \cdot P_m(\theta). \tag{2.12}$$

We will use (2.12) to express the first, second and third derivatives of the log-likehood functions in terms of marginal probabilities. Taking the first derivative of (2.8), we obtain the j^{th} entry of the score vector

$$\frac{\partial \ell}{\partial \theta_j} = t(j) - N P_j(\theta). \tag{2.13}$$

The $(j,m)^{th}$ entry of the second derivative of (2.8) is

$$\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_m} = -N \left[P_{jm}(\theta) - P_j(\theta) \cdot P_m(\theta)\right], \tag{2.14}$$

and the $(j,m,l)^{th}$ entry of the third derivative of (2.8) is

$$\begin{aligned}\frac{\partial^{(3)}\ell}{\partial\theta_j\partial\theta_m\partial\theta_l} &= -N\left[\frac{\partial P_{jm}(\theta)}{\partial\theta_l} - \frac{\partial P_j(\theta)}{\partial\theta_l}P_m(\theta) - \frac{\partial P_m(\theta)}{\partial\theta_l}P_j(\theta)\right] \\ &= -N[P_{jml}(\theta) - P_{jm}(\theta)P_l(\theta) - P_{jl}(\theta)P_m(\theta) + P_j(\theta)P_l(\theta)P_m(\theta) \\ &\quad - P_{ml}(\theta)P_j(\theta) + P_m(\theta)P_l(\theta)P_j(\theta)].\end{aligned} \tag{2.15}$$

2.3 The Bayes factor for decomposable models

Consider a sample of q-dimensional random vectors taking values in a q-dimensional contingency table, as described in Section 2.2.2, where the cell counts follow a multinomial distribution with density (2.2). It is shown in Lauritzen (1996) that if a probability distribution is Markov with respect to a decomposable graph G, then it can be written as a product of *factors* over the cliques and the separators of G. Let $\mathcal{C}$ be the set of cliques, $\mathcal{S}$ be the set of separators, and $\nu(S)$ be the multiplicity of separator $S \in \mathcal{S}$. Let $p^C(i_C)$ and $p^S(i_S)$ denote the C-marginal and the S-marginal cell probabilities, respectively, where C denotes a clique in $\mathcal{C}$. Let p be the vector of the C-marginal and the S-marginal cell probabilities. Then from the distribution (2.2), we can write the log-likelihood function as

$$\ell(p) = \sum_{C\in\mathcal{C}}\sum_{i_C\in I_C} n_C(i_C)\log p^C(i_C) - \sum_{S\in\mathcal{S}}\nu(S)\sum_{i_S\in I_S} n_S(i_S)\log p^S(i_S), \tag{2.16}$$

with maximum likelihood estimate (MLE)

$$\widehat{p}(i) = \frac{\prod_{C\in\mathcal{C}} n_C(i_C)}{N\prod_{S\in\mathcal{S}} n_S(i_S)^{\nu(S)}}. \tag{2.17}$$

The number of parameters in a decomposable model with log-likelihood (2.16) is the number of parameters corresponding to the cliques minus the number of parameters corresponding to the separators, to avoid redundancy. The number of free parameters is one less than the number of parameters since the sum of the cell probabilities is equal to 1. We denote the number of free parameters in a decomposable model as

$$k = -1 + \sum_{C \in \mathcal{C}} |I_v| - \sum_{S \in \mathcal{S}} \nu(S) \cdot |I_v|. \tag{2.18}$$

Note that the number of levels $|I_v|$ can vary across different $v \in V$. For convenient notation, we assume $|I_v| = 2$, meaning $|I_C| = 2^{|C|}$ and $|I_S| =^{|S|}$.

2.3.1 The hyper Dirichlet prior

The density of the multinomial random vector X given a graph G can be written as the NEF

$$f(x|\theta, G) = \exp\{\langle \theta, t \rangle - Nk(\theta)\}. \tag{2.19}$$

From Diaconis and Ylvisaker (1979), we know the conjugate prior on θ can be written as

$$\pi_{s,\alpha}(\theta|G) = \frac{1}{I_G(s,\alpha)} \exp\{\langle \theta, s \rangle - \alpha k(\theta)\}, \tag{2.20}$$

with normalizing constant $I_G(s,\alpha)$ and hyperparameters $s = (s(j), j \in J)$ and α. The vector s consists of *fictive counts* from a fictive contingency table, and the set of indices J correspond to the same subset of I as for the sufficient statistic t. The real number α is the total fictive

counts and the choice of α will change the shape of the prior distribution. The sufficient statistic and the hyperparamters both depend on G', and in the high-dimensional case, α will depend on N; however, we suppress G' and N in the notation for readability. In Lemma 3.1 of Massam et al. (2009) show that $I_G(s,\alpha) < +\infty$ holds if and only if $\alpha > 0$ and there exists a contingency table with cells $i \in I$ such that

$$s(j) = \alpha \sum_{i_{S(j)}=j_{S(j)}} p(i), \quad \text{for} \quad j \in J,$$

where each $p(i) > 0$. See Massam et al. (2009) for a discussion on how to obtain (s,α). The posterior probability of G given x is

$$f(G|x) = \frac{\int \exp\{\langle \theta_G, t+s\rangle - (N+\alpha)k(\theta)\}d\theta_G}{\sum_{G'\in\mathcal{G}} \int \exp\{\langle \theta_{G'}, t+s\rangle - (N+\alpha)k(\theta)\}d\theta_{G'}} = \frac{I_G(t+s, N+\alpha)}{\sum_{G'\in\mathcal{G}} I_{G'}(t+s, N+\alpha)}. \quad (2.21)$$

From (2.21), when each model is assigned an equal probability, we see that the pairwise Bayes factor comparing model G_1 to model G_2 is a ratio of the normalizing constants

$$BF_{G_1,G_2} = \frac{f(G_1|x)}{f(G_2|x)} = \frac{I_{G_1}(t+s, N+\alpha)}{I_{G_2}(t+s, N+\alpha)}. \quad (2.22)$$

Let $s(i)$ denote the fictive cell count for cell $i \in I$ and let $s_D(i_D)$ denote any D-marginal fictive cell count – not necessarily corresponding to the set of indices J. Massam et al. (2009)

proved that the prior $\pi_{s,\alpha}(\theta|G)$ exhibits the strong hyper Markov property for graphical models as defined by Dawid and Lauritzen (1993), and is identical to the hyper Dirichlet with normalizing constant

$$I_G(s,\alpha) = \frac{\prod_{C\in\mathcal{C}} \prod_{i_C\in I_C} \Gamma(s_C(i_C))}{\Gamma(\alpha) \prod_{S\in\mathcal{S}} \left[\prod_{i_S\in I_S} \Gamma(s_S(i_S))\right]^{\nu(S)}}. \tag{2.23}$$

Their Proposition 4.1 demonstrates how to obtain the marginal fictive cell counts from a linear combination of the components of s and α. Therefore, the Bayes factor (2.22) becomes the product of gamma functions, where their arguments are the sum of true cell counts and fictive cell counts over the cliques and the separators. When the two models differ by only one edge, the Bayes factor becomes a localized comparison.

Example. Consider the binary random variables X_a, X_b and X_c and the two decomposable models $G_1 = \{ab, c\}$ and $G_2 = \{ab, bc\}$. Let $n^*(i) = n(i) + s(i)$ be the sum of the true and the fictive counts for cell i, let $n^*_D(i_D) = n_D(i_D) + s_D(i_D)$ be the sum of the true and the fictive D-marginal cell counts, and let $N^* = N + \alpha$ be the sum of the total true and the total fictive cell counts. For convenience, we denote the components of (2.23) as

$$\widetilde{\Gamma}^C = \prod_{i_C\in I_C} \Gamma(n^*_C(i_C))$$

and

$$\widetilde{\Gamma}^S = \prod_{i_S\in I_S} \Gamma(n^*_S(i_S)).$$

Then the Bayes factor comparing G_1 and G_2 is

$$\begin{aligned} BF_{G_1,G_2} &= \frac{\widetilde{\Gamma}^{ab}\widetilde{\Gamma}^{c}\widetilde{\Gamma}^{b}\Gamma(N^*)\Gamma(N^*)}{\Gamma(N^*)\Gamma(N^*)\Gamma(N^*)\widetilde{\Gamma}^{ab}\widetilde{\Gamma}^{bc}} \\ &= \frac{\widetilde{\Gamma}^{c}\widetilde{\Gamma}^{b}}{\Gamma(N^*)\widetilde{\Gamma}^{bc}}, \end{aligned}$$

where

$$\widetilde{\Gamma}^{c} = \Gamma(n^*(000) + n^*(100) + n^*(010) + n^*(110))\Gamma(n^*(001) + n^*(101) + n^*(011) + n^*(111)),$$

$$\widetilde{\Gamma}^{b} = \Gamma(n^*(000) + n^*(100) + n^*(001) + n^*(101))\Gamma(n^*(010) + n^*(110) + n^*(011) + n^*(111)),$$

$$\widetilde{\Gamma}^{bc} = \Gamma(n^*(000) + n^*(100))\Gamma(n^*(010) + n^*(110))\Gamma(n^*(001) + n^*(101))\Gamma(n^*(011) + n^*(111)).$$

Since G_1 and G_2 differ by only one edge, the Bayes factor simplifies to the ratio comparing no edge from vertices b to c versus the inclusion of the edge (b, c).

2.3.2 The model prior

When the dimension of the model is fixed the Bayes factor is sufficient to guarantee the true model is selected; however, when the dimension of the model is increasing we require a prior distribution on the model to encourage sparsity. We follow the approach used in Niu et al. (2021). To limit the number of false edges, we assume given a decomposable graph G that the probability of a model parameter being included in G follows a Bernoulli distribution.

Let $Q_{\max} = 2^5(q-1)$ be the upper bound for the total number of parameters in the model,

where we assume 5 is the most complex interaction between variables. This assumption is for convenience; results hold for any finite constant. For $l = 1, \ldots, Q_{\max}$, let $e_l = 1$ if the l^{th} parameter is in G, and 0 otherwise. Then the prior distribution of a model G given probability ρ is

$$\pi(G|\rho) \propto \left[\prod_{l=1}^{Q_{\max}} \rho^{e_l} (1-\rho)^{1-e_l} \right] \cdot \mathbb{1}_{\mathcal{D}_q}(G), \tag{2.24}$$

where $\mathcal{D}_q$ is the set of all decomposable graphs with q vertices. We state our assumption about ρ in Section 2.4.

Therefore, when the dimension is increasing we will examine the behaviour of the posterior odds ratio, defined as

$$PR_{G_1,G_2} = \frac{f(G_1|x)}{f(G_2|x)} = \frac{f(x|G_1)\pi(G_1)}{f(x|G_2)\pi(G_2)}, \tag{2.25}$$

where $f(x|G) = \int \exp\{\langle \theta_G, t+s \rangle - (N+\alpha)k(\theta)\} d\theta_G$.

2.3.3 Asymptotic approximation of the posterior probability

Our results rely on the logarithm of the Bayes factor, which becomes the log of the difference of normalizing constants of the form $I_G(t+s, N+\alpha)$. Thus, we use Lemma 2.3.1 and Lemma 2.3.2 to write an asymptotic approximation for the log of this normalizing constant, which is proportionate to the posterior probability for a decomposable model (2.21). For our proofs in Sections 2.3.3, 2.4 and 2.5, we use background lemmas which state the necessary large

deviation results. In Appendix 4, we provide the proofs for Lemmas 4.1.1-4.2.1.

Here we give a convenient expression for $\log\Gamma(x)$ with $x \geqslant 1$, for $x \in \mathbb{R}$, using the inequality for the gamma function from Theorem 1.6 in Batir (2008).

Lemma 2.3.1. *For all positive real numbers $x \geqslant 1$,*

$$\log\Gamma(x) = x\log x - x - 1/2\log x + c_1,$$

where $c_1 \in (\frac{1}{2}\log 2\pi, \frac{1}{2}\log 3\pi)$.

Proof of Lemma 2.3.1. Theorem 1.6 in Batir (2008) states that for all positive real numbers $x \geqslant 1$, we have

$$x^x e^{-x}\sqrt{2\pi(x+a)} < \Gamma(x+1) < x^x e^{-x}\sqrt{2\pi(x+b)},$$

with a, b constant. They specify that $a = 1/6$ and $b = \frac{e^2}{2\pi} - 1$ are the best possible constants. For our purposes, these constants are negligible.

Since $\Gamma(x+1) = x\Gamma(x)$, then

$$x^x e^{-x}\sqrt{2\pi(x+a)} < x\Gamma(x) < x^x e^{-x}\sqrt{2\pi(x+b)}$$

$$x^{x-1} e^{-x}\sqrt{2\pi(x+a)} < \Gamma(x) < x^{x-1} e^{-x}\sqrt{2\pi(x+b)}$$

$$x^{x-1} e^{-x}\sqrt{2\pi x} < \Gamma(x) < x^{x-1} e^{-x}\sqrt{3\pi x}$$

$$x^x e^{-x} x^{-1/2}\sqrt{2\pi} < \Gamma(x) < x^x e^{-x} x^{-1/2}\sqrt{3\pi}$$

$$\log\left(x^x e^{-x} x^{-1/2}\sqrt{2\pi}\right) < \log\Gamma(x) < \log\left(x^x e^{-x} x^{-1/2}\sqrt{3\pi}\right)$$

$$x\log x - x - 1/2\log x + 1/2\log 2\pi < \log\Gamma(x) < x\log x - x - 1/2\log x + 1/2\log 3\pi.$$

Therefore,

$$\log\Gamma(x) = x\log x - x - 1/2\log x + c_1,$$

where $c_1 \in (\frac{1}{2}\log 2\pi, \frac{1}{2}\log 3\pi)$. □

We use Lemma 2.3.1 to rewrite log of the normalizing constant, which has the form of (2.23) with arguments $t+s$ and $N+\alpha$, without gamma functions. Since Lemma 2.3.1 applies to $x \geqslant 1$; in practice, if the true count of any marginal cell is zero then we assign 1 to the corresponding marginal fictive cell count; otherwise, the fictive cell counts are constants greater than zero and the total number of fictive counts is assumed to be much less than the total sample size. This ensures that each component of the vector $t+s$ is greater than or equal to 1. In this scenario, even though the fictive counts are not evenly distributed across the cells, this does not affect the asymptotic result since the fictive counts are small

compared to the true counts. Our theoretical results require the MLE to exist, meaning there are no cells with zero counts. Therefore, in the relevant proof, we show that the MLE for each cell is non-zero with probability tending to 1.

Let $\widehat{p}^*(i) = \frac{n(i)+s(i)}{N+\alpha}$ be the frequency estimate of sum of true and fictive counts for cell $i \in I$ with the same Markov property as the MLE (2.17). We denote the sum of the true and the fictive D-marginal cell counts by $n_D^*(i_D)$ and the of the true and fictive total cell counts by N^*. Then we let $\widehat{p}^*$ be the vector of marginal cell frequencies indexed by the cliques $C \in \mathcal{C}$ and the separators $S \in \mathcal{S}$, namely, $\widehat{p}^{*C}(i_C) = \frac{n_C(i_C)+s_C(i_C)}{N+\alpha} = \frac{n_C^*(i_C)}{N^*}$ and $\widehat{p}^{*S}(i_S) = \frac{n_S(i_S)+s_S(i_S)}{N+\alpha} = \frac{n_S^*(i_S)}{N^*}$. To simplify our expression when the sample size $N \to \infty$, we evaluate the log-likelihood at $\widehat{p}^*$; that is,

$$\ell(\widehat{p}^*) = \sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} n_C^*(i_C) \log \widehat{p}^{*C}(i_C) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} n_S^*(i_S) \log \widehat{p}^{*S}(i_S). \tag{2.26}$$

Lemma 2.3.2. *Assume the true C-marginal and S-marginal cell probabilities $p_0^C(i_C)$ and $p_0^S(i_S)$ are bounded away from 0 and 1, and $|\log p_0^C(i_C)| < c_2$ and $|\log p_0^S(i_S)| < c_2$, for a positive constant c_2. Let $\widehat{p}^*$ denote the vector of frequency estimators for the vector of true marginal probabilities p_0. Then, if k is the number of parameters in the model and the sample size $N \to \infty$, we have*

1. *when q is fixed,*

$$\log I_G(t+s, N+\alpha) = \ell(\widehat{p}^*) - \frac{k}{2}\log(N+\alpha) + O(1),$$

2. when $q_N \to \infty$ as $N \to \infty$,

$$\log I_G(t+s, N+\alpha) = \ell(\widehat{p}^*) - \frac{k_N}{2}(\log(N+\alpha) + O(1)) + C_N,$$

with probability $1 - q_N O(Q_N{}^{-Q_N})$*, where* $C_N = 2^5(q_N - 1)(c_2 + \epsilon_N)$*,* $Q_N = q_N{}^2$*,* $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$*, and* C *is a positive universal constant.*

Proof of Lemma 2.3.2. Let n^* be the vector of marginal counts such that $n^*_D(i_D) = n_D(i_D) + s_D(i_D)$ is the sum of the true D-marginal cell counts and the fictive D-marginal cell counts, and let $N^* = N + \alpha$ be the sum of the total cell counts and the total fictive counts for a decomposable model G with m cliques and $(m-1)$ separators. We must consider two cases: when q is fixed and when q_N is increasing. First, we find an asymptotic expression for $\log I_G(n^*, N^*)$ when q is fixed.

Let $\mathcal{C}$ and $\mathcal{S}$ denote the set of cliques and the set of separators for a model G, respectively. By Lemma 2.3.1, taking the logarithm of $I_G(n^*, N^*)$ gives

$$\begin{aligned}
\log I_G(n^*, N^*) = &\sum_{C \in \mathcal{C}} \sum_{i_C \in I_C} \left[n^*_C(i_C) \log n^*_C(i_C) - n^*_C(i_C) - \frac{1}{2} \log n^*_C(i_C) \right] \\
&- \sum_{S \in \mathcal{S}} \nu(S) \sum_{i_S \in I_S} \left[n^*_S(i_S) \log n^*_S(i_S) - n^*_S(i_S) - \frac{1}{2} \log n^*_S(i_S) \right] \\
&- N^* \log N^* + N^* + \frac{1}{2} \log N^* + O(1),
\end{aligned}$$

where $O(1)$ is a constant equal to the finite number of parameters k times a constant c_1, such that $c_1 \in (\frac{1}{2}\log 2\pi, \frac{1}{2}\log 3\pi)$.

Since $N \to \infty$, we approximate the cell counts which are arguments of a $\log(\cdot)$ function with the marginal frequency estimates $\widehat{p}^{*C}(i_C) = \frac{n^*_C(i_C)}{N^*}$ and $\widehat{p}^{*S}(i_S) = \frac{n^*_S(i_S)}{N^*}$, and we denote the vector of marginal frequency estimates by $\widehat{p}^*$. Then we rearrange the terms to obtain the log-likelihood function evaluated at $\widehat{p}^*$ plus a penalty term, that is

$$
\begin{aligned}
\log I_G(n^*, N^*) &= \sum_{C\in\mathcal{C}} \sum_{i_C\in I_C} n_C^*(i_C) \log N^* \widehat{p}^{*C}(i_C) - \sum_{C\in\mathcal{C}} \sum_{i_C\in I_C} n_C^*(i_C) \\
&- \sum_{C\in\mathcal{C}} \sum_{i_C\in I_C} \frac{1}{2} \log N^* \widehat{p}^{*C}(i_C) - \sum_{s\in\mathcal{S}} \nu(S) \sum_{i_S\in I_S} n_S^*(i_S) \log N^* \widehat{p}^{*S}(i_S) \\
&+ \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S\in I_S} n_S^*(i_S) + \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S\in I_S} \frac{1}{2} \log N^* \widehat{p}^{*S}(i_S) - N^* \log N^* \\
&+ N^* + \frac{1}{2} \log N^* + O(1) \\
&= \sum_{C\in\mathcal{C}} \sum_{i_C\in I_C} n_C^*(i_C) \log \widehat{p}^{*C}(i_C) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S\in I_S} n_S^*(i_S) \log \widehat{p}^{*S}(i_S) \\
&+ mN^* \log N^* - (m-1)N^* \log N^* - mN^* - \sum_{C\in\mathcal{C}} \sum_{i_C\in I_C} \frac{1}{2} \log N^* \widehat{p}^{*C}(i_C) \\
&+ (m-1)N^* + \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S\in I_S} \frac{1}{2} \log N^* \widehat{p}^{*S}(i_S) - N^* \log N^* + N^* \\
&+ \frac{1}{2} \log N^* + O(1) \\
&= \ell(\widehat{p}^*) - \sum_{C\in\mathcal{C}} \sum_{i_C\in I_C} \frac{1}{2} \log N^* \widehat{p}^{*C}(i_C) + \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S\in I_S} \frac{1}{2} \log N^* \widehat{p}^{*S}(i_S) \\
&+ \frac{1}{2} \log N^* + O(1),
\end{aligned}
$$

where $\ell(\widehat{p}^*)$ is the log-likelihood function (2.26) evaluated at the marginal frequency estimator $\widehat{p}^*$. Since the dimension of the model is fixed,

$$
- \sum_{C\in\mathcal{C}} \sum_{i_C\in I_C} \frac{1}{2} \log \widehat{p}^{*C}(i_C) + \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S\in I_S} \frac{1}{2} \log \widehat{p}^{*S}(i_S) \tag{2.27}
$$

is approximately constant, thus it can be absorbed into $O(1)$. Therefore,

$$\begin{aligned}
\log I_G(n^*, N^*) &= \ell(\widehat{p}^*) - \sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \frac{1}{2} \log N^* + \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} \frac{1}{2} \log N^* + \frac{1}{2} \log N^* + O(1) \\
&= \ell(\widehat{p}^*) - \sum_{C\in\mathcal{C}} \frac{2^{|C|}}{2} \log N^* + \sum_{S\in\mathcal{S}} \nu(S) \frac{2^{|S|}}{2} \log N^* + \frac{1}{2} \log N^* + O(1) \\
&= \ell(\widehat{p}^*) + \left(\frac{1 - \sum_{C\in\mathcal{C}} 2^{|C|} + \sum_{S\in\mathcal{S}} \nu(S) 2^{|S|}}{2} \right) \log N^* + O(1). \\
&= \ell(\widehat{p}^*) - \frac{k}{2} \log N^* + O(1),
\end{aligned}$$

where k is the number of parameters of the model (2.18).

Next, when the dimension of the model is increasing as the sample size increases to infinity, we need to consider the multiplicity of the constant c_1 in the interval $(\frac{1}{2}\log 2\pi, \frac{1}{2}\log 3\pi)$ and we need to establish an upper bound for (2.27). By Lemma 2.3.1, we have

$$\begin{aligned}
\log I_G(n^*, N^*) = \ell(\widehat{p}^*) - \frac{k_N}{2}(\log N^* + O(1)) - \sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \frac{1}{2} \log \widehat{p}^{*C}(i_C) \\
+ \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} \frac{1}{2} \log \widehat{p}^{*S}(i_S).
\end{aligned}$$

We write $(\log N^* + O(1))$ to account for adding a constant c_1 for each log of a gamma function.

We assume the true C-marginal and S-marginal cell probabilities $p_0^C(i_C)$ and $p_0^S(i_S)$ are bounded away from 0 and 1, and $|\log p_0^C(i_C)| < c_2$ and $|\log p_0^S(i_S)| < c_2$ for a positive constant c_2. Moreover, we assume the most complex interaction is a 5-way interaction and we consider $q_N - 1$ to be the maximum number of cliques in a decomposable graph. By Lemma 4.1.11 with $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$,

$$\begin{aligned}\left| -\frac{1}{2}\left(\sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \log \widehat{p}^{*C}(i_C) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} \log \widehat{p}^{*S}(i_S) \right) \right| &< \left| \sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \log \widehat{p}^{*C}(i_C) \right| \\ &< \left| \sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} (\log p_0^C(i_C) + \epsilon_N) \right| \\ &< 2^5(q_N - 1)(c_2 + \epsilon_N).\end{aligned}$$

with probability $1 - q_N O({Q_N}^{-Q_N})$. Therefore,

$$\log I_G(n^*, N^*) = \ell(\widehat{p}^*) - \frac{k_N}{2}(\log N^* + O(1)) + C_N$$

with probability $1 - q_N O({Q_N}^{-Q_N})$, where $C_N = 2^5(q_N - 1)(c_2 + \epsilon_N)$.

□

Notice that our first expression in Lemma 2.3.2 is similar to the well-known Bayesian information criterion (BIC); namely, BIC $= -2\ell(\widehat{p}) + k \log N$. In this case we would have

$$\log I_G(s+t, N+\alpha) = -\frac{\text{BIC}}{2} + O(1) = \ell(\widehat{p}) - k \log N + O(1).$$

However, the BIC requires the Laplace approximation of the integral

$$\int f(x|\theta, G)\pi_{s,\alpha}(\theta|G)d\theta_G,$$

which is suitable for a fixed dimension, but it is harder to control the error term when the dimension is increasing. Therefore, instead of the Laplace approximation, we use our $\log \Gamma(x)$ approximation in Lemma 2.3.1 since our main goal is to focus on the high-dimensional scheme. Our approximation is preferable because it allows us to avoid high-dimensional integration and to control each error term.

2.4 Theoretical results when the true graph is decomposable

In this section, we present our pairwise Bayes factor consistency results and our strong consistency results when the true graph is decomposable. First, we state the graph notation we follow from Niu et al. (2021) and the necessary assumptions to support our results.

Denote $\mathcal{G}_q$ as the q-dimensional graph space and $\mathcal{D}_q$ as the q-dimensional decomposable graph space. We use $G_t = (V, E_t)$ to denote the true graph, and suppose $G_a = (V, E_a)$ is

any competing decomposable graph which is not the true graph G_t, then let $E_a \cap E_t$ be the set of true edges in E_a. Moreover, we write $G_a \subsetneq G_t$ to denote $E_a \subsetneq E_t$, $G_a \not\subset G_t$ to denote $E_a \not\subset E_t$, and $G_a \neq G_t$ to denote $E_a \neq E_t$.

When q_N is increasing with N, the concept of a 'true graph' is in fact a sequence of true graphs depending on N. However, this notion is challenging to capture in our theoretical results. Thus, we assume that at every value of q_N, there exists a true graph G_t.

Assumption 1. Assume that the true cell probabilities $p_0(i)$ for each cell $i \in I$ are bounded away from 0 and 1, and that $|\log p_0(i)|$ for $i \in I$ is bounded by a positive constant. Similarly, the true C-marginal and S-marginal cell probabilities $p_0^C(i_C)$ and $p_0^S(i_S)$ are bounded away from 0 and 1, and $|\log p_0^C(i_C)| < c_2$ and $|\log p_0^S(i_S)| < c_2$ for a positive constant c_2.

This assumption ensures that we can control our asymptotic results in the high-dimensional setting. The number of cells and the number of model parameters increase as the number of variables increases, so this assumption allows us to find upper bounds for summations that are indexed by the marginal cells.

Assumption 2. In the high-dimensional setting, the number of variables $q_N \to \infty$ as the sample size $N \to \infty$ and ${q_N}^4 \log q_N = o(N)$.

This assumption restricts the dimension of the model when the dimension is increasing with the sample size. To prove model selection consistency for a competing overfitting model, we require that the number of variables cannot increase faster than $N^{1/4}$. This assumption is more strict, but comparable to the assumptions in Fitch et al. (2014) and in Niu et al. (2021), where they let $p = O(n^{1/3})$ in the high-dimensional setting.

Assumption 3. Assume the most complex interaction between the variables is a 5-way interaction, meaning the order of any clique or separator is at most 5. We assume all variables are binary, thus we consider $2^5(q_N - 1)$ to be the upper bound for the number of parameters in a given model. Without loss of generality, we can change 5-way to any m-way interaction as long as m is bounded, and we can change the base-2 to the highest number of levels $|I_C|$ associated with any clique in G.

To control the complexity of the model when the parameters increases as $q_N \to \infty$, we assume that 5 is the highest order of interaction. We can obtain the upper bound for maximum number of parameters; that is, $-1 + \sum_{C \in \mathcal{C}} 2^{|C|} - \sum_{S \in \mathcal{S}} \nu(S) 2^{|S|} < \sum_{C \in \mathcal{C}} 2^{|C|} < 2^5(q_N - 1)$, where $q_N - 1$ is the maximum number of cliques in a decomposable graph. The most cliques a decomposable graph can have is one less than the number of vertices, meaning the graph represents only 2-way interactions. This assumption holds as long as as the order of largest clique is bounded.

Assumption 4. The ratio of the total fictive counts over the total true cell counts is bounded, such that $0 < \frac{\alpha}{N} < \epsilon_N$, where $\epsilon_N = (CN^{-1} q \log N)^{1/2}$ when q is fixed and $\epsilon_N = (CN^{-1} Q_N \log Q_N)^{1/2}$ when $q_N \to \infty$.

This assumption states that the total fictive counts are negligible compared to the total true cell counts.

Assumption 5. The smallest Kullback-Leibler divergence between the true model G_t and an underfitting model G_a is bounded; that is, $\mathrm{E}_t \log f_t(x) - \mathrm{E}_t \log f_a(x) > c_m$, where the positive constant c_m is a universal lower bound for the Kullback-Leibler divergence.

When the competing model is an underfitting model, we apply the Kullback-Leibler theorem to find the behaviour of the Bayes factor. Since the Kullback-Leibler divergence between models G_t and G_a depends on the model G_a, we require that the smallest c_m is bounded to prove strong model selection consistency.

Assumption 6. Under the true model, let $H(\theta) = \mathrm{E}_t\left(-\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^T}\right)$ denote the Fisher Information, and let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and the smallest eigenvalues of $H(\theta)$, respectively. We assume that the eigenvalues of the Fisher Information matrix under the true model are bounded, meaning there exists constants $M_1 > 0$ and $M_2 < +\infty$ such that $M_1 < \lambda_{\min}(H(\theta)) \leqslant \lambda_{\max}(H(\theta)) < M_2$.

This is a standard assumption and it ensures that the random variable $u^T U(\theta)$ is bounded, where u is a unit vector and $U(\theta)$ is a score vector.

Assumption 7. Let ρ denote the probability of including a parameter such that $\log \rho = -\gamma \log q_N$, for some constant $\gamma > 0$.

This assumption gives the condition for the model prior; namely, the probability ρ is inversely proportional to the number of variables q_N.

2.4.1 Pairwise Bayes factor consistency for decomposable graphs

We study the behaviour of the Bayes factor (2.22) between two decomposable models, say G_a and G_b, when q is fixed and when q_N is increasing with N. Two cases arise: when G_a is an underfitting model and G_b is an overfitting model, and when G_a and G_b are both overfitting models. We focus on decomposable graphs because they allow for the Bayes factor to be

written explicitly. This is still possible when the true graph is non-decomposable since it can be approximated by a minimal triangulation which, by definition, is a decomposable graph.

Here we give four lemmas for decomposable graphs in both the overfitting and the underfitting case, which lay the foundation to later prove strong model selection consistency. We begin by stating the lemmas for the pairwise comparison of decomposable graphs with a fixed dimension.

Lemma 2.4.1. *Let q be fixed. If G_a and G_b are both decomposable graphs, where G_a is an underfitting model with $|E_a \cap E_t| < |E_t|$, and G_b is an overfitting model, with $|E_b \cap E_t| = |E_t|$, then for $\epsilon_N = (18CN^{-1}q \log N)^{1/2}$ and positive constants c_m, C_1 and C_2,*

$$BF_{G_a,G_b} < \exp\left\{-c_m N^* + C_1 N^* \epsilon_N + C_2 \log N^* + O(1)\right\},$$

with probability greater than $1 - O(N^{-q})$ as $N \to \infty$.

Proof of Lemma 2.4.1. For a decomposable model G_a, let $\widehat{p}_a^*$ be the vector of frequency estimates for the sum of the true and fictive marginal counts, let p_{0a} be the vector of true marginal cell probabilities, and let k_a be the number of parameters in the model. Similarly, for a decomposable model G_b. We assume G_a is an underfitting model and G_b is an overfitting model. By Lemma 2.3.2, we can write the logarithm of the Bayes factor comparing G_a with G_b as

$$\log BF_{G_a,G_b} = \ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*) + \frac{(k_b - k_a)}{2} \log N^* + O(1). \tag{2.28}$$

For N observations, we define $\ell(\widehat{p}_a^*)$ and $\mathrm{E}_t \log f_a(x)$ as we did in the proof of Lemma 2.4.1, where $\log f_a$ represents the log density under graph G_a, and the expectation is under the true density $f_t(x)$. We also have the equivalent expressions for $\ell(\widehat{p}_b^*)$ and $\mathrm{E}_t \log f_b(x) = \mathrm{E}_t \log f_t(x)$, corresponding to the model G_b. Next, we write

$$\ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*) = \mathrm{E}_t \log f_a(x) - \mathrm{E}_t \log f_t(x) + \ell(\widehat{p}_a^*) - \mathrm{E}_t \log f_a(x) - \left[\ell(\widehat{p}_b^*) - \mathrm{E}_t \log f_t(x)\right].$$

Under Assumption 5, for all the underfitting models,

$$\mathrm{E}_t \log f_t(x) - \mathrm{E}_t \log f_a(x) > c_m N^*,$$

where c_m is the lower bound for the Kullback–Leibler divergence.

By Lemma 4.1.6 with $\epsilon_N = (18CN^{-1}q \log N)^{1/2}$,

$$|\ell(\widehat{p}_a^*) - \mathrm{E}_t \log f_a(x)| < 2^5(q-1)N^*\epsilon_N$$

and

$$|\ell(\widehat{p}_b^*) - \mathrm{E}_t \log f_t(x)| < 2^5(q-1)N^*\epsilon_N$$

with probability $1 - 2^6(q-1)O(N^{-q})$.

Since q is finite, we let $C_1 = 2^6(q-1)$ and $C_2 = \frac{(k_a - k_b)}{2}$ be constants. Therefore,

$$\mathrm{P}\left(BF_{G_a,G_b} < \exp\left\{-c_m N^* + C_1 N^* \epsilon_N + C_2 \log N^* + O(1)\right\}\right) > 1 - O(N^{-q}).$$

Notice that $N^*\epsilon_N = N\epsilon_N + \alpha\epsilon_N$ and $N\epsilon_N = (18CqN\log N)^{1/2}$. Since $\log N < N$ and $N < N^*$ implies that $(N\log N)^{1/2} < N^*$, then the leading term is $-c_m N^*$ and $BF_{G_a,G_b} \xrightarrow{\mathrm{P}} 0$ as $N \to \infty$.

□

Lemma 2.4.2. *Let q be fixed and let α be the total fictive counts. If G_a and G_b are both decomposable overfitting models, with $|E_a \cap E_t| = |E_b \cap E_t| = |E_t|$ and $k_a > k_b$, then for $\epsilon_N = (18CN^{-1}q\log N)^{1/2}$ and positive constants C_1, and C_2,*

$$BF_{G_a,G_b} < \exp\left\{-C_1 \log N^* + C_2 \log(\log N)\{1 + o(1)\} + O(1)\right\},$$

with probability greater than $1 - O((\log N)^{-\tilde{a}}) - O(N^{-q})$ as $N \to \infty$, where $\tilde{a} = (k_a - k_b)/6$.

Proof of Lemma 2.4.2. Let $\widehat{p}_t$ be the vector of marginal frequency estimators for the true model, and let the decomposable models G_a and G_b both be overfitting models. We can write the logarithm of the Bayes factor comparing G_a with G_b as

$$\log BF_{G_a,G_b} = \ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*) - \frac{(k_a - k_b)}{2}\log N^* + O(1), \tag{2.29}$$

where $k_a > k_b$. We define $\ell(\widehat{p}_a^*)$ and $\ell(\widehat{p}_b^*)$ as we did in the proof of Lemma 2.4.2. We can

write

$$|\ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*)| = |\ell(\widehat{p}_a^*) - \ell(\widehat{p}_a) - \{\ell(\widehat{p}_b^*) - \ell(\widehat{p}_b)\} + \ell(\widehat{p}_a) - \ell(\widehat{p}_b)|$$

$$\leqslant |\ell(\widehat{p}_a^*) - \ell(\widehat{p}_a)| + |\ell(\widehat{p}_b^*) - \ell(\widehat{p}_b)| + |\ell(\widehat{p}_a) - \ell(\widehat{p}_b)|$$

By Assumption 4,

$$|\ell(\widehat{p}_a^*) - \ell(\widehat{p}_a)| < 2^5(q-1)C'N^*\frac{\alpha}{N^*} = 2^5(q-1)C'\alpha$$

and

$$|\ell(\widehat{p}_b^*) - \ell(\widehat{p}_b)| < 2^5(q-1)C''N^*\frac{\alpha}{N^*} = 2^5(q-1)C''\alpha,$$

where C' and C'' are positive constants.

Next, since the log-linear parametrization (2.5) is a unique representation we can use $\ell(\widehat{\theta}_a) - \ell(\widehat{\theta}_b)$ in the place of $\ell(\widehat{p}_a) - \ell(\widehat{p}_b)$. Let $\xi = N^{-1/2}H_a^{-1/2}(\theta_0)U_a(\theta_0)$ and $A = H_a^{1/2}(\theta_0)\left(\{H_a(\theta_0)\}^{-1} - D_a\{H_b(\theta_0)\}^{-1}D_a\right)H_a(\theta_0)^{1/2}$ with $tr(B) = k_a - k_b$. By Lemma 4.2.1 and the proof of Lemma 2.4.4, we know that we can write

$$2\{\ell(\widehat{\theta}_a) - \ell(\widehat{\theta}_b)\} = \xi^T A\xi\{1 + o_p(1)\}$$

with probability $1 - O(N^{-q})$, where the largest eigenvalue of $H_a(\theta_0)$ satisfies $[\lambda_{\max}(H_a(\theta_0))]^{-1} \leqslant M_1^{-1}$ for a positive constant M_1 such that $M_1 < \lambda_{\min}(H_a(\theta_0))$ and $\xi \sim subGaussian(1/M_1^2)$. Following the proof of Lemma 2.4.2, using Lemma A.4 in Gao and Carroll (2017) and Corollary 4.2 in Spokoiny and Zhilova (2013), we choose $g = (4q \log N)^{1/2}$ and $K = (k_a - k_b)\log(\log N)$. Therefore, $(k_a - k_b)\log(\log N) > [2(k_a - k_b)]^{1/2}/3$ and $q \log N > (k_a - k_b)\log(\log N)$ for large N. If $\xi^* = M_1\xi$, then

$$\begin{aligned} \mathrm{P}\left(|\xi^{*T} A \xi^*| \geqslant (k_a - k_b)(1 + \log(\log N))\right) &\leqslant 10.4 \exp\{-(k_a - k_b)\log(\log N)/6\} \\ &= O((\log N)^{-\tilde{a}}), \end{aligned}$$

where $\tilde{a} = (k_a - k_b)/6$.

Since q is finite, we let $C_1 = \frac{(k_a - k_b)}{2}$, and $C_2 = \frac{(k_a - k_b)}{2M_1}$ be constants. Then from (2.29),

$$\begin{aligned} &\mathrm{P}\bigg(BF_{G_a,G_b} < \exp\bigg\{ 2^5(q-1)C' + 2^5(q-1)C''\alpha - \frac{(k_a - k_b)}{2}\log N^* \\ &+ \frac{(k_a - k_b)}{2M_1}(1 + \log(\log N))\{1 + o_p(1)\} + O(1) \bigg\} \bigg) \\ &= \mathrm{P}\bigg(BF_{G_a,G_b} < \exp\big\{ - C_1 \log N^* + C_2 \log(\log N)\{1 + o_p(1)\} + O(1) \big\} \bigg) \\ &> 1 - O((\log N)^{-\tilde{a}}) - O(N^{-q}). \end{aligned}$$

Since $\log(\log N) < \log N^*$, then the leading term is $-C_1 \log N^*$ and $BF_{G_a,G_b} \xrightarrow{\mathrm{P}} 0$ as $N \to \infty$.

□

Lemma 2.4.1 states that the Bayes factor will support an overfitting model, which contains all the true edges, over an underfitting model with at least one missing true edge. Lemma 2.4.2 states that when comparing two overfitting models, the Bayes factor supports the model with less superfluous edges. We see that in the first scenario, the Bayes factor will converge to zero at an exponential rate; however, in the second scenario, the Bayes factor converges to zero at a polynomial rate. This means that the Bayes factor with the hyper Dirichlet prior gives a stronger penalty to the removal of a true edge than it does to add a false edge. These results coincide with those of Theorem 4.1 in Niu et al. (2021), which uses the hyper-inverse Wishart prior in the Gaussian setting.

Next, we state the lemmas for the pairwise comparison of decomposable graphs with increasing dimension. When $q_N \to \infty$, we require a prior distribution of the model (2.24); therefore, Lemmas 2.4.3 and 2.4.4 examine the behaviour of the posterior odds ratio (2.25).

Lemma 2.4.3. *Let q_N be increasing with N. If G_a and G_b are both decomposable graphs, where G_a is an underfitting model with $|E_a \cap E_t| < |E_t|$, and G_b is an overfitting model, with $|E_b \cap E_t| = |E_t|$, then for $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$ and positive constants c_m, c_2, and γ,*

$$PR_{G_a,G_b} < \exp\big\{ - c_m N^* + 2^6(q_N - 1)[c_2 + (N^* + 1)\epsilon_N] + (k_{b,N} - k_{a,N})\big[1/2 \log N^* + \gamma \log q_N + O(1)\big]\big\}$$

with probability greater than $1 - O(Q_N^{-Q_N})$ as $N \to \infty$.

Proof of Lemma 2.4.3. For a decomposable model G_a, let $\widehat{p}_a^*$ be the vector of frequency

estimates for the sum of the true and fictive marginal counts, let p_{0a} be the vector of true marginal cell probabilities, and let k_a be the number of parameters in the model. Similarly, for a decomposable model G_b. We assume G_a is an underfitting model and G_b is an overfitting model. By Lemma 2.3.2, we can write the logarithm of the Bayes factor comparing G_a with G_b as

$$\log BF_{G_a,G_b} = \ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*) + \frac{(k_{b,N} - k_{a,N})}{2}(\log N^* + O(1)) + C_{a,N} - C_{b,N}, \tag{2.30}$$

where $|C_{a,N} - C_{b,N}| \leqslant 2^6(q_N - 1)(c_2 + \epsilon_N)$ for $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$, with probability $1 - q_N O(Q_N^{-Q_N})$.

Since q_N increases with the sample size N, we require an expression for the log of posterior odds ratio (2.25) using our asymptotic approximation in Lemma 2.3.2. For two competing models G_a and G_b, by Assumption 7 with $\rho < 1/2$, we have that,

$$\log \frac{\pi(G_a|\rho)}{\pi(G_b|\rho)} \propto \log \frac{\rho^{k_{a,N}}(1-\rho)^{Q_{\max}-k_{a,N}}}{\rho^{k_{b,N}}(1-\rho)^{Q_{\max}-k_{b,N}}} = (k_{a,N} - k_{b,N}) \log \frac{\rho}{1-\rho} = -(k_{a,N} - k_{b,N})(\gamma \log q_N + O(1)),$$

for some constant $\gamma > 0$ and $Q_{\max} = 2^5(q_N - 1)$. Therefore, the log of the posterior odds ratio is

$$\log PR_{G_a,G_b} = \ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*) + \frac{(k_{b,N} - k_{a,N})}{2}(\log N^* + O(1)) + 2^6(q_N - 1)(c_2 + \epsilon_N) \\ - (k_{a,N} - k_{b,N})(\gamma \log q_N + O(1)). \tag{2.31}$$

Let $\mathcal{C}_a$ and $\mathcal{S}_a$ denote the set of cliques and the set of separators for model G_a, respectively. For N observations, we have

$$\ell(\widehat{p}_a^*) = N^* \sum_{C \in \mathcal{C}_a} \sum_{i_C \in I_C} \widehat{p}_a^{*C}(i_C) \log \widehat{p}_a^{*C}(i_C) - N^* \sum_{S \in \mathcal{S}_a} \nu(S) \sum_{i_S \in I_S} \widehat{p}_a^{*S}(i_S) \log \widehat{p}_a^{*S}(i_S),$$

and

$$\mathrm{E}_t \log f_a(x) = N^* \sum_{C \in \mathcal{C}_a} \sum_{i_C \in I_C} p_{0,a}^C(i) \log p_{0,a}^C(i) - N^* \sum_{S \in \mathcal{S}_a} \nu(S) \sum_{i_S \in I_S} p_{0,a}^S(i) \log p_{0,a}^S(i),$$

where $\log f_a$ represents the log density under graph G_a, and the expectation is under the true density $f_t(x)$. We also have the equivalent expressions for $\ell(\widehat{p}_b^*)$ and $\mathrm{E}_t \log f_b(x) = \mathrm{E}_t \log f_t(x)$, corresponding to the model G_b. Next, we write

$$\ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*) = \mathrm{E}_t \log f_a(x) - \mathrm{E}_t \log f_t(x) + \ell(\widehat{p}_a^*) - \mathrm{E}_t \log f_a(x) - \left[\ell(\widehat{p}_b^*) - \mathrm{E}_t \log f_t(x)\right].$$

Under Assumption 5, for all the underfitting models,

$$\mathrm{E}_t \log f_t(x) - \mathrm{E}_t \log f_a(x) > c_m N^*,$$

where c_m is the lower bound for the Kullback–Leibler divergence.

By Lemma 4.1.11 with $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$,

$$|\ell(\widehat{p}_a^*) - \mathrm{E}_t \log f_a(x)| < 2^5(q_N - 1)N^*\epsilon_N$$

and

$$|\ell(\widehat{p}_b^*) - \mathrm{E}_t \log f_t(x)| < 2^5(q_N - 1)N^*\epsilon_N$$

with probability $1 - 2^6(q_N - 1)O(Q_N^{-Q_N})$. Therefore,

$$\begin{aligned}
&\mathrm{P}\bigg(PR_{G_a,G_b} < \exp\big\{ - c_m N^* + 2^6(q_N - 1)N^*\epsilon_N + \frac{(k_{b,N} - k_{a,N})}{2}(\log N^* + O(1)) \\
&+ 2^6(q_N - 1)(c_2 + \epsilon_N) - (k_{a,N} - k_{b,N})(\gamma \log q_N + O(1))\big\}\bigg) \\
&= \mathrm{P}\bigg(PR_{G_a,G_b} < \exp\big\{ - c_m N^* + 2^6(q_N - 1)[c_2 + (N^* + 1)\epsilon_N] + (k_{b,N} - k_{a,N})[1/2 \log N^* \\
&+ \gamma \log q_N + O(1)]\big\}\bigg) \\
&> 1 - O(Q_N^{-Q_N}).
\end{aligned}$$

Notice that $q_N N^* \epsilon_N = q_N N \epsilon_N + q_N \alpha \epsilon_N$ and $q_N N \epsilon_N = (18C{q_N}^4 N \log q_N^2)^{1/2}$. Under Assump-

tion 2, $q_N^4 \log q_N^2 \propto q_N^4 \log q_N < N$; thus, $({q_N}^4 N \log q_N^2)^{1/2} < N < N^*$. Therefore, the leading term is $-c_m N^*$ and $PR_{G_a,G_b} \xrightarrow{\text{P}} 0$ for $q_N \to \infty$ as $N \to \infty$.

□

Lemma 2.4.4. *Let q_N be increasing with N, let α be the total fictive counts and let ω be a positive constant greater than 6. If G_a and G_b are both decomposable overfitting models, with $|E_a \cap E_t| = |E_b \cap E_t| = |E_t|$ and $k_{a,N} > k_{b,N}$, then for $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$, and positive constant $\gamma > (\omega/2M_1 - 2)$,*

$$PR_{G_a,G_b} < \exp\big\{ -(k_{a,N} - k_{b,N})(2 + \gamma - \omega/2M_1) \log q_N \{1 + o(1)\}\big\},$$

with probability greater than $1 - O({q_N}^{-\tilde{a}}) - O(Q_N^{-Q_N})$ as $N \to \infty$, where $\tilde{a} = (k_{a,N} - k_{b,N})\omega/6$ and M_1 is the lower bound for the smallest eigenvalue of the Fisher Information matrix under the true model.

Proof of Lemma 2.4.4. Let $\widehat{p}_t$ be the vector of marginal frequency estimators for the true model, and let the decomposable models G_a and G_b both be overfitting models with $k_{a,N} > k_{b,N}$. By Lemma 2.3.2, we can write the logarithm of the Bayes factor comparing G_a with G_b as

$$\log BF_{G_a,G_b} = \ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*) - \frac{(k_{a,N} - k_{b,N})}{2}(\log N^* + O(1)) + C_{a,N} - C_{b,N}, \tag{2.32}$$

where $|C_{a,N} - C_{b,N}| \leqslant 2^6(q_N - 1)(c_2 + \epsilon_N)$ for $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$, with probability $1 - q_N O(Q_N^{-Q_N})$.

Similar to the proof of Lemma 2.4.3, we require an expression for the log of posterior odds ratio (2.25). For $\gamma > (\omega/2M_1 - 2)$ and $\rho < 1/2$, where ω and M_1 are positive constants defined later in the proof, by Assumption 7

$$\log \frac{\pi(G_a|\rho)}{\pi(G_b|\rho)} < (k_{a,N} - k_{b,N}) \log \frac{\rho}{1-\rho} = -(k_{a,N} - k_{b,N})(\gamma \log q_N + O(1)).$$

Therefore, the log of the posterior odds ratio is

$$\begin{aligned} \log PR_{G_a,G_b} = \ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*) - \frac{(k_{a,N} - k_{b,N})}{2}(\log N^* + O(1)) + 2^6(q_N - 1)(c_2 + \epsilon_N) \\ - (k_{a,N} - k_{b,N})(\gamma \log q_N + O(1)). \end{aligned} \tag{2.33}$$

We can write

$$\begin{aligned} |\ell(\widehat{p}_a^*) - \ell(\widehat{p}_b^*)| &= |\ell(\widehat{p}_a^*) - \ell(\widehat{p}_a) - \{\ell(\widehat{p}_b^*) - \ell(\widehat{p}_b)\} + \ell(\widehat{p}_a) - \ell(\widehat{p}_b)| \\ &\leqslant |\ell(\widehat{p}_a^*) - \ell(\widehat{p}_a)| + |\ell(\widehat{p}_b^*) - \ell(\widehat{p}_b)| + |\ell(\widehat{p}_a) - \ell(\widehat{p}_b)| \end{aligned}$$

Recall that

$$|\ell(\widehat{p}_a^*) - \ell(\widehat{p}_a)| = \Bigg| \sum_{C\in\mathcal{C}} \sum_{i_C\in I_C} n^*_{a,C} \log \widehat{p}_a^{*C}(i_C) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S\in I_S} n^*_{a,S} \log \widehat{p}_a^{*S}(i_S) - \left(\sum_{C\in\mathcal{C}} \sum_{i_C\in I_C} n_{a,C} \log \widehat{p}_a^{C}(i_C) - \sum_{S\in\mathcal{S}} \sum_{i_S\in I_S} n_{a,S} \log \widehat{p}_a^{S}(i_S) \right) \Bigg| \tag{2.34}$$

If any cell i has zero counts, then the MLE $\widehat{p}(i) = 0$ and $\log \widehat{p}(i)$ is undefined, meaning $\ell(\widehat{p}_a)$ and $\ell(\widehat{p}_b)$ as also undefined. By Assumption 1, we have that $p_0(i)$ is bounded from below; that is, $p_0(i) > c' > 0$ for some constant c'. By Lemma 4.1.8,

$$|\widehat{p}(i) - p_0(i)| < (2N^{-1}Q_N \log Q_N)^{1/2}$$

with probability $1 - 2e^{-Q_N \log Q_N}$. If

$$(2N^{-1}Q_N \log Q_N)^{1/2} < \frac{c'}{2},$$

as $N \to \infty$, then

$$\widehat{p}(i) > p_0(i) - (2N^{-1}Q_N \log Q_N)^{1/2} > c' - \frac{c'}{2} > \frac{c'}{2} > 0.$$

Thus, we have

$$\mathrm{P}\left(\min \widehat{p}(i) > \frac{c'}{2}\right) = 1 - 2^{q_N} 2e^{-Q_N \log Q_N} = 1 - e^{(q_N+1)\log 2 - Q_N \log Q_N},$$

where 2^{q_N} is the number of cells. Therefore, with probability tending to 1, the MLE for each cell probability is non-zero, and hence the log-likelihoods evaluated at the MLE are well defined.

For some clique $C \in \mathcal{C}$ in the model G_a, let us consider (2.34) for a particular marginal cell $i_C \in I_C$. Then we have

$$\left|(n(i_{C)} + s(i_C)) \log\left(\frac{n(i_C) + s(i_C)}{N^*}\right) - n(i_C) \log \frac{n(i_C)}{N}\right|.$$

By the mean value theorem for functions with two variables, there exists a point (c, d) on the line segment from (x_1, y_1) to (x_2, y_2) such that $f(x_2, y_2) - f(x_1, y_1) = \frac{\partial f}{\partial x}(c, d)(x_2 - x_1) + \frac{\partial f}{\partial y}(c, d)(y_2 - y_1)$. Consider the function $f(x, y) = x \log \frac{x}{y}$, then $\frac{\partial f}{\partial x} = \log \frac{x}{y} + 1$ and $\frac{\partial f}{\partial y} = \frac{-x}{y}$. Hence,

$$\left|(n(i_{C)} + s(i_C)) \log\left(\frac{n(i_C) + s(i_C)}{N^*}\right) - n(i_C) \log \frac{n(i_C)}{N}\right| = \left(\log \frac{c}{d} + 1\right) s(i_C) - \frac{c}{d}\alpha,$$

where $c < d$ since (c, d) is a point on the line segment from $(n(i_C), N)$ to $(n(i_C)+s(i_C), N+\alpha)$. Then by Assumption 3,

$$\sum_{C\in\mathcal{C}}\sum_{i_C\in I_C}\left|\left(\log\frac{c}{d}+1\right)s(i_C)-\frac{c}{d}\alpha\right|\leqslant c_3(q_N-1)\alpha,$$

where $c_3 \geqslant 2^5$ is a constant. Therefore,

$$|\ell(\widehat{p}_a^*)-\ell(\widehat{p}_a)|+|\ell(\widehat{p}_b^*)-\ell(\widehat{p}_b)|\leqslant 2c_3(q_N-1)\alpha.$$

Next, since the log-linear parametrization (2.5) is a unique representation we can use $\ell(\widehat{\theta}_a)-\ell(\widehat{\theta}_b)$ in the place of $\ell(\widehat{p}_a)-\ell(\widehat{p}_b)$. By Lemma 4.2.1, we have

$$\begin{aligned}2\{\ell(\widehat{\theta}_a)-\ell(\widehat{\theta}_b)\} &= 2\{\ell(\widehat{\theta}_a)-\ell(\theta_0)\}-2\{\ell(\widehat{\theta}_b)-\ell(\theta_0)\}\\ &= \frac{1}{N}U_a(\theta_0)^T\left(H_a^{-1}(\theta_0)-D_aH_b^{-1}(\theta_0)D_a\right)U_a(\theta_0)\{1+o(1)\}\end{aligned}$$

with probability $1-2(k_a+k_b)O(Q_N^{-Q_N})$, where $(k_a+k_b)\leqslant 2^6(q_N-1)$, $D_a=\left(I_{k_{b,N}}, 0_{k_{b,N},k_{a,N}-k_{b,N}}\right)$ and $I_{k_{b,N}}$ is the identity matrix with dimension $k_{b,N}\times k_{b,N}$ and $0_{k_{b,N},k_{a,N}-k_{b,N}}$ is the matrix of zero with dimension $k_{b,N}\times(k_{a,N}-k_{b,N})$. To simplify notation, let $U_a = U_a(\theta_0)$ and $H_a^{-1/2} = H_a^{-1/2}(\theta_0)$. Then we standardize the score vector as follows,

$$\frac{1}{N}U_a^T\left(H_a^{-1} - D_a^T H_b^{-1} D_a\right) U_a$$
$$= \left(H_a^{-1/2}\frac{U_a}{\sqrt{N}}\right)^T \left[H_a^{1/2}\left(H_a^{-1} - D_a^T H_b^{-1} D_a\right) H_a^{1/2}\right] \left(H_a^{-1/2}\frac{U_a}{\sqrt{N}}\right).$$

We want to show that $N^{-1/2}H_a^{-1/2}U_a$ is a sub-Gaussian random vector. By definition, a random vector $X \in \mathbb{R}^d$ is said to be sub-Gaussian if it is centred and if, for any unit vector $u \in \mathbb{R}^d$, the random variable $u^T X$ is sub-Gaussian (Pauwels, 2020).

We know the score vector has the form $U = t - NP(\theta)$, where t is the vector of sufficient statistics defined in (2.6) and $P(\theta)$ is the vector of corresponding marginal probabilities as defined in (2.9), (2.10), and (2.11) - both with length $|J|$. Let $\tilde{t}$ denote an indicator vector, where it has a 1 in the component indicating one observation for a particular t_j for $j \in J$ and 0 in its other components. Since each component of $\tilde{t}$ and $P(\theta)$ are less or equal to 1, then $||\tilde{t} - 1 \cdot P(\theta)||_2 \leqslant ||\tilde{t}||_2 + ||P(\theta)||_2 \leqslant 2$. Thus, if $u \in \mathbb{R}^{|J|}$ is a unit vector such that $||u||_2^2 = 1$, then

$$u^T H_a^{-1/2}(\tilde{t} - P(\theta)) \leqslant ||u^T H_a^{-1/2}||_2 ||\tilde{t} - P(\theta)||_2 \leqslant ||u||_2 ||H_a^{-1/2}||_2 \cdot 2 = \lambda_{\max}(H_a^{-1}) \cdot 2,$$

where $\lambda_{\max}(H_a)$ denotes the largest eigenvalue of the matrix H_a. By Assumption 6, $\lambda_{\max}(H_a^{-1}) =$

$[\lambda_{\max}(H_a)]^{-1} \leqslant M_1^{-1}$ for a constant $M_1 > 0$ such that $M_1 < \lambda_{\min}(H_a)$. Thus,

$$u^T H_a^{-1/2}(\tilde{t} - P(\theta)) \leqslant \frac{2}{M_1},$$

meaning $u^T H_a^{-1/2}(\tilde{t} - \tilde{P}(\theta))$ is a strictly bounded random variable. Then, since

$$\mathrm{E}\left(u^T H_a^{-1/2}(\tilde{t} - P(\theta))\right) = 0,$$

we can apply Hoeffding's Lemma (Lemma 1.8 in Rigollet (2003)). Therefore, for all $s \in \mathbb{R}$,

$$\mathrm{E}\left(\exp\left\{s u^T H_a^{-1/2}(\tilde{t} - P(\theta))\right\}\right) \leqslant e^{s^2(2/M_1)^2/8} = e^{s^2/2M_1^2}.$$

Thus, by definition and by Hoeffding's Lemma, since $u^T H_a^{-1/2}(\tilde{t} - P(\theta)) \sim subGaussian(1/M_1^2)$, then $H_a^{-1/2}(\tilde{t} - P(\theta)) \sim subGaussian(1/M_1^2)$.

For N data points, $N^{-1/2} u^T H_a^{-1/2} U_a = N^{-1/2} \sum_{i=1}^{N} u^T H_a^{-1/2}(\tilde{t}_i - NP(\theta))$. Since the 2-norm of a vector of length N with components $N^{-1/2}$ is equal to 1, by Corollary 1.7 in Rigollet (2003), we have that

$$\mathrm{E}\left(\exp\left\{s N^{-1/2} u^T H_a^{-1/2} U_a\right\}\right) \leqslant e^{s^2/2M_1^2},$$

meaning $N^{-1/2}u^T H_a^{-1/2}U_a \sim subGaussian(1/M_1^2)$. Therefore, by definition, $N^{-1/2}H_a^{-1/2}U_a \sim subGaussian(1/M_1^2)$.

For the rest of the proof, we follow the proof of Lemma A.4 from Gao and Carroll (2017) which applies the large deviation result from Corollary 4.2 in Spokoiny and Zhilova (2013). Let $\xi = N^{-1/2}H_a^{-1/2}U_a$ and $A = H_a^{1/2}\left(H_a^{-1} - D_a^T H_b^{-1} D_a\right) H_a^{1/2}$. It can be shown that $tr(A) = k_{a,N} - k_{b,N}$, where $tr(\cdot)$ denotes the trace of a matrix. Then we define $\xi^* = M_1\xi$, because ξ^* satisfies the exponential moment condition

$$\log \mathrm{E}\left(\exp\{s^T\xi^*\}\right) \leqslant ||s||_2^2/2, \qquad s \in \mathbb{R}^{|J|},\ ||s||_2^2 \leqslant g$$

required for Corollary 4.2 (Spokoiny and Zhilova, 2013) for any real number $g > 0$. Corollary 4.2 states that

$$\mathrm{P}\left(|\xi^{*T}A\xi^*| \geqslant (k_{a,N} - k_{b,N}) + K\right) \leqslant 2\exp\{-K/6\} + 8.4\exp\{-x_c\},$$

with $6x_c > K > [2(k_{a,N} - k_{b,N})]^{1/2}/3$ and $x_c > g^2/4$ for large N. We choose $g = (4Q_N \log Q_N)^{1/2}$ and $K = (k_{a,N} - k_{b,N})\omega \log q_N$, for a positive constant $\omega > 6$. Therefore, $(k_{a,N} - k_{b,N})\omega \log q_N > [2(k_{a,N} - k_{b,N})]^{1/2}/3$ and

$$\mathrm{P}\left(|\xi^{*T}A\xi^*| \geqslant (k_{a,N} - k_{b,N})(1 + \omega \log q_N)\right) \leqslant 10.4\exp\{-(k_{a,N} - k_{b,N})\omega \log q_N/6\} = O({q_N}^{-\tilde{a}}),$$

where $\tilde{a} = (k_{a,N} - k_{b,N})\omega/6$.

The terms $C_{a,N}$ and $C_{b,N}$ result from our approximation in Lemma 2.3.2. In the underfitting case, in Lemma 2.4.3, the order of the difference $C_{a,N} - C_{b,N}$ was less than the leading term. However, for the overfitting case, we need to be more precise with this difference since both terms depend on the number of model parameters and the parameter for G_b are a subset of those for G_a. From Lemma 2.3.2, we see that $C_{a,N} - C_{b,N}$ is a bound for

$$\begin{aligned}
&\left| -\frac{1}{2}\left(\sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \log \widehat{p}_a^{*C}(i_C) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} \log \widehat{p}_a^{*S}(i_S) \right) \right| \\
&- \left| -\frac{1}{2}\left(\sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \log \widehat{p}_b^{*C}(i_C) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} \log \widehat{p}_b^{*S}(i_S) \right) \right|.
\end{aligned}$$

Therefore, by Lemma 4.1.9 with $\epsilon_N = (CN^{-1}Q_N \log Q_N)^{1/2}$,

$$\begin{aligned}
&\left\| -\frac{1}{2}\left(\sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \log \widehat{p}_a^{*C}(i_C) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} \log \widehat{p}_a^{*S}(i_S) \right) \right| \\
&- \left| -\frac{1}{2}\left(\sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \log \widehat{p}_b^{*C}(i_C) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} \log \widehat{p}_b^{*S}(i_S) \right) \right\| \\
&\leqslant \left| \left(\sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \log \widehat{p}_a^{*C}(i_C) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} \log \widehat{p}_a^{*S}(i_S) \right) - \left(\sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} \log \widehat{p}_b^{*C}(i_C) \right. \right. \\
&\left. \left. - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} \log \widehat{p}_b^{*S}(i_S) \right) \right| \\
&< \left| \left(\sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} (\log p_{0,a}^{C}(i_C) + \epsilon_N) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} (\log p_{0,a}^{S}(i_S) + \epsilon_N) \right) \right. \\
&\left. - \left(\sum_{C\in\mathcal{C}} \sum_{i_C \in I_C} (\log p_{0,b}^{C}(i_C) + \epsilon_N) - \sum_{S\in\mathcal{S}} \nu(S) \sum_{i_S \in I_S} (\log p_{0,b}^{S}(i_S) + \epsilon_N) \right) \right|,
\end{aligned}$$

with probability $1 - O(Q_N^{-Q_N})$. Under Assumptions 1 and 4, for a given clique $\left|\sum_{i_C \in I_C} \log p_0(i_C)\right| \leqslant 2^5 c_2$, and for a given separator $\left|\sum_{i_S \in I_S} \log p_0(i_S)\right| \leqslant 2^5 c_2$. Since G_a is an overfitting model, there are $(k_{a,N} - k_{b,N})$ extra parameters θ_j such that $j \in J_a \backslash J_b$, where J_a is the subset of I corresponding to the free parameters in G_a and similarly for J_b. By Assumption 3 each j has at most five 1's in it, meaning as $q_N \to \infty$, the addition of any θ_j for $j \in J_a \backslash J_b$ affects at most 5 cliques and 4 separators. Thus,

$$
\begin{aligned}
&\left| \left(\sum_{C \in \mathcal{C}} \sum_{i_C \in I_C} (\log p_{0,a}^C(i_C) + \epsilon_N) - \sum_{S \in \mathcal{S}} \nu(S) \sum_{i_S \in I_S} (\log p_{0,a}^S(i_S) + \epsilon_N) \right) \right. \\
&\left. - \left(\sum_{C \in \mathcal{C}} \sum_{i_C \in I_C} (\log p_{0,b}^C(i_C) + \epsilon_N) - \sum_{S \in \mathcal{S}} \nu(S) \sum_{i_S \in I_S} (\log p_{0,b}^S(i_S) + \epsilon_N) \right) \right| \\
&\leqslant 2 \cdot 5 \cdot 2^5 (c_2 + \epsilon_N)(k_{a,N} - k_{b,N}) + 2 \cdot 4 \cdot 2^5 (c_2 + \epsilon_N)(k_{a,N} - k_{b,N}) \\
&= 18 \cdot 2^5 (c_2 + \epsilon_N)(k_{a,N} - k_{b,N})
\end{aligned}
$$

We choose $\alpha = \frac{1}{(q_N - 1)}$. Therefore,

$$
\begin{aligned}
&\mathrm{P}\bigg(PR_{G_a,G_b} < \exp\big\{2c_3(q_N-1)\alpha - \frac{(k_{a,N}-k_{b,N})}{2}(\log N^* + O(1)) + \frac{(k_{a,N}-k_{b,N})}{2M_1}(1+\omega\log q_N) \\
&+ 18\cdot 2^5(c_2+\epsilon_N)(k_{a,N}-k_{b,N}) - (k_{a,N}-k_{b,N})(\gamma\log q_N + O(1))\big\}\bigg) \\
&> \mathrm{P}\bigg(PR_{G_a,G_b} < \exp\big\{2c_3 - (k_{a,N}-k_{b,N})(2+\gamma-\omega/2M_1)\log q_N + (k_{a,N}-k_{b,N})[18\cdot 2^5(c_2+\epsilon_N) \\
&+ 1/2M_1 + O(1)]\big\}\bigg) \\
&= \mathrm{P}\bigg(PR_{G_a,G_b} < \exp\big\{ -(k_{a,N}-k_{b,N})(2+\gamma-\omega/2M_1)\log q_N\{1+o(1)\}\big\}\bigg) \\
&> 1 - O({q_N}^{-\tilde{a}}) - O(Q_N^{-Q_N}).
\end{aligned}
$$

Since $\gamma > (\omega/2M_1 - 2)$ implies that $-(k_{a,N}-k_{b,N})(2+\gamma-\omega/2M_1) < 0$, then the leading term is $-(k_{a,N}-k_{b,N})(2+\gamma-\omega/2M_1)\log q_N$ and $PR_{G_a,G_b} \xrightarrow{\mathrm{P}} 0$ for $q_N \to \infty$ as $N \to \infty$.

□

Lemma 2.4.3 yields very similar results as Lemma 2.4.1. When the competing model is an underfitting model, the model prior does not affect the result since $\log q_N < N$. By Assumption 2, we have that $q_N^4 \log q_N < N^*$ and by Assumption 3, the upper bound for the number of parameters is a model is $2^5(q_N - 1)$, meaning $(k_{b,N} - k_{a,N})/2 \log N^* \propto q_N \log N^* < N^*$ since $\log N^* < (N^*)^{3/4}$. Thus, in Lemma 2.4.3, $-c_m N^*$ is the leading term and the posterior odds ratio converges to zero at an exponential rate. The model prior makes sure that strong model selection consistency will hold in the high-dimensional case for all overfitting models. Indeed, Lemma 2.4.4 holds for $\gamma > (\omega/2M_1 - 2)$, such that $\log \rho = \log {q_N}^{-\gamma}$, where ρ is the prior edge inclusion probability. In Lemma 2.4.4, the posterior odds ratio converges at a

polynomial rate. Therefore, we see that the removal of a true edge still has a stronger penalty than adding a false edge when q_N is increasing, but the posterior odds ratio also favours the model with less superfluous edges.

2.4.2 Strong model selection consistency for decomposable graphs

So far we have discussed in general the pairwise comparison of decomposable models. Now we must specify how the previous four lemmas from Section 2.4.1 are used when considering the true model, specifically when the true model is decomposable. Suppose G_a is any decomposable model such that $G_a \neq G_t$. When G_a is an underfitting model we have $G_a \subsetneq G_t$ and we can apply Lemma 2.4.1 if q is fixed or Lemma 2.4.3 if q_N is increasing, where we treat G_t as an overfitting model. Similarly, when G_a is an overfitting model, we have $G_a \not\subset G_t$ and we can apply Lemma 2.4.2 or Lemma 2.4.4, where we treat both G_t and G_a as overfitting models with $k_a > k_t$.

In Section 2.3.1, we defined the posterior probability (2.21) for any graph G. Thus, for the true graph G_t, we define

$$f(G_t|x) = \frac{I_{G_t}(t+s, N+\alpha)}{\sum_{G' \in \mathcal{D}_q} I_{G'}(t+s, N+\alpha)}. \tag{2.35}$$

In order to proof strong consistency, we show that the posterior probability $f(G_t|x)$ converges to 1 as $N \to \infty$. We do this by showing that the sum of all of the Bayes factors, for fixed q, or all of the posterior odds ratios, for increasing q_N, converge to 0.

Theorem 2.4.5 and Theorem 2.4.6 state the strong consistency results for when the true

graph G_t is decomposable for fixed q and increasing q_N, respectively.

Theorem 2.4.5. *Let q be fixed. If the true graph G_t is decomposable, then we have the following:*

1. *Let G_a be any decomposable graph which is not equal to G_t, then $BF_{G_a,G_t} \xrightarrow{P} 0$ as $N \to \infty$.*

2. *For all competing decomposable models, $f(G_t|x) \xrightarrow{P} 1$ as $N \to \infty$.*

Proof of Theorem 2.4.5. This proof requires two cases: when the competing model is an underfitting model, meaning $G_t \nsubseteq G_a$, and when the competing model is an overfiting model, meaning $G_t \subsetneq G_a$.

Part 1. In case 1, when $|E_a \cap E_t| < |E_t|$, G_a is an underfitting model. Since G_t is decomposable and has more edges than G_a, by Lemma 2.4.1, $BF_{G_a,G_t} \xrightarrow{\mathrm{P}} 0$.

Let $c_m > 0$ be the smallest Kullback-Leibler divergence between an underfitting model G_a and the true graph G_t, which by Assumption 5 is bounded. By Lemma 2.4.1, where $\epsilon_N = (18CN^{-1}q \log N)^{1/2}$, and C_1 and C_2 are positive constants,

$$\log BF_{G_a,G_t} < -c_m N^* + C_1 N^* \epsilon_N + C_2 \log N^* + O(1) < -c_m N^* + c' N^* = -(c_m - c')N^*,$$

where $c' > 0$ be a positive constant. Let δ_1 be the upper bound for $\log BF_{G_a,G_t}$. Therefore, for all of the competing underfitting models, we have

$$\mathrm{P}\left(\max_{G_a:G_t\not\subseteq G_a} BF_{G_a,G_t} \leqslant e^{-\delta_1}\right) > 1 - 2^q \exp\{-Cq\log N\}$$
$$> 1 - \exp\{q^2\log 2 - Cq\log N\} \to 1,$$

as $N \to \infty$, since q is fixed and C is a positive universal constant.

In case 2, when $|E_a \cap E_t| = |E_t|$ and $|E_a| > |E_t|$, then $k_a > k_t$. Since G_a is an overfitting model and G_t is decomposable, by Lemma 2.4.2, $BF_{G_a,G_t} \xrightarrow{\mathrm{P}} 0$.

By Lemma 2.4.2, where α is the total fictive counts, and C_1, and C_2 are positive constants,

$$\begin{aligned}\log BF_{G_a,G_t} &< -C_1(k_a - k_t)\log N^* + C_2(k_a - k_t)\log(\log N) + O(1)\\ &< -(C_1 - C_2)(k_a - k_t)\log N^* + O(1)\\ &= -C'(k_a - k_t)\log N^* + O(1),\end{aligned}$$

where C' is a constant such that $0 < C' < (C_1 - C_2)$. Let $m' = k_a - k_t$ and let δ_2 be the upper bound for $\log BF_{G_a,G_t}$. For fixed q, the number of model parameters is at most $2^q - 1$. Since $1 - O(N^{-\tilde{a}}) > 1 - O(N^{-\tilde{a}}) - O(N^{-q})$, using the binomial theorem, for all of the competing overfitting models, we have

$$\mathrm{P}\left(\max_{G_a:G_t\subsetneq G_a} BF_{G_a,G_t} \leqslant e^{-\delta_2}\right) > 1 - \left[\left(1 + N^{-\omega}\right)^{2^q} - 1\right] \to 1,$$

as $N \to \infty$, since q is fixed and $\lim_{N\to\infty} (1 + N^{-\omega})^{2^q} = 1$.

Part 2. From Part 1, we have that the largest upper bound for BF_{G_a,G_t} when $G_t \nsubseteq G_a$ is

$$BF_{G_a,G_t:G_t \nsubseteq G_a} < e^{-D_1 N^*},$$

where $D_1 = c_m - c' > 0$ is a constant. Let k'_a be the number of true parameters in k_a. Then for all underfitting competing models, we have

$$\begin{aligned}\sum_{G_a:G_t \nsubseteq G_a} BF_{G_a,G_t} &= \sum_{k'_a=0}^{k_t-1} \binom{k_t}{k'_a} \sum_{k_a-k'_a=0}^{(2^q-1)-k_t} \binom{(2^q-1)-k_t}{k_a-k'_a} BF_{G_a,G_t:G_t \nsubseteq G_a} \\ &< \exp\{q \log 2 - D_1 N^*\} \to 0,\end{aligned}$$

as $N \to \infty$. Also, we have that the largest upper bound for BF_{G_a,G_t} when $G_t \subsetneq G_a$ is

$$BF_{G_a,G_t:G_t \subsetneq G_a} < e^{-D_2 (k_a - k_t) \log N^*},$$

where $1 < D_2 < (C_1 - C_2)$ is a constant. Let $m' = k_a - k_t$. Then for all overfitting competing models, we have

$$\begin{aligned}
\sum_{G_a:G_t\subsetneq G_a} BF_{G_a,G_t} &= \sum_{k_a=k_t+1}^{2^p-1} \binom{(2^q-1)-k_t}{k_a-k_t} BF_{G_a,G_t:G_t\subsetneq G_a} \\
&< \sum_{m'=1}^{(2^q-1)-k_t} \binom{(2^q-1)-k_t}{m'} \left[e^{-D_2\log N^*}\right]^{m'} \\
&= \left[1+e^{-D_2\log N^*}\right]^{(2^q-1)-k_t} - 1 \\
&< \left[1+N^{*-D_2}\right]^{2^q} - 1 \\
&< \exp\{2^q N^{*-D_2}\} - 1 \to 0
\end{aligned}$$

as $N \to \infty$, since $1 + x < e^x$ for $x > 0$, q is finite and $D_2 > 0$.

From (2.35) and the proof of Theorem 2.4.6, we have

$$f(G_t|x) = \frac{1}{1+\sum_{G_a\neq G_t} BF_{G_a,G_t}}.$$

When q is fixed and G_t is decomposable, we have shown that for any competing decomposable graph G_a, the sum $\sum_{G_a\neq G_t} BF_{G_a,G_t} \xrightarrow{\mathrm{P}} 0$. Therefore, $f(G_t|x) \xrightarrow{\mathrm{P}} 1$ as $N \to \infty$.

□

Theorem 2.4.6. *Let q_N be increasing with N. If the true graph G_t is decomposable, then we have the following:*

1. *Let G_a be any decomposable graph which is not equal to G_t, then $PR_{G_a,G_t} \xrightarrow{P} 0$ as $N \to \infty$.*

2. For all competing decomposable models, $f(G_t|x) \xrightarrow{P} 1$ *as* $N \to \infty$.

To prove Part 2 of Theorem 2.4.6, we require a slightly stronger condition on γ; that is, $\gamma > (\omega/2M_1 - 1)$.

Proof of Theorem 2.4.6. This proof requires two cases: when the competing model is an underfitting model, meaning $G_t \nsubseteq G_a$, and when the competing model is an overfiting model, meaning $G_t \subsetneq G_a$.

Part 1. In case 1, when $|E_a \cap E_t| < |E_t|$, G_a is an underfitting model. Since G_t is decomposable and has more edges than G_a, by Lemma 2.4.3, $PR_{G_a,G_t} \xrightarrow{\mathrm{P}} 0$.

Let $c_m > 0$ be the smallest Kullback-Leibler divergence between an underfitting model G_a and the true model G_t, which by Assumption 5 is bounded. By Lemma 2.4.3, where $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$, and c_2, and γ are positive constants,

$$\begin{aligned}
\log PR_{G_a,G_t} &< -c_m N^* + 2^6(q_N - 1)[c_2 + (N^* + 1)\epsilon_N] \\
&\quad + (k_{t,N} - k_{a,N})\big[1/2 \log N^* + \gamma \log q_N + O(1)\big] \\
&< -c_m N^* + c' N^* \\
&= -(c_m - c')N^*,
\end{aligned}$$

where $c' > 0$ be a positive constant. Let δ_1 be the upper bound for $\log PR_{G_a,G_t}$. Therefore, for all of the competing underfitting models, we have

$$\mathrm{P}\left(\max_{G_a:G_t \nsubseteq G_a} PR_{G_a,G_t} \leqslant e^{-\delta_1}\right) > 1 - 2^{Q_N}\exp\{-CQ_N \log Q_N\}$$

$$> 1 - \exp\{Q_N \log 2 - CQ_N \log Q_N\} \to 1$$

as q_N increases with $N \to \infty$, where C is a positive universal constant.

In case 2, when $|E_a \cap E_t| = |E_t|$ and $|E_a| > |E_t|$, then $k_{a,N} > k_{t,N}$. Since G_a is an overfitting model and G_t is decomposable, by Lemma 2.4.4, $PR_{G_a,G_t} \xrightarrow{\mathrm{P}} 0$.

By Lemma 2.4.4, where $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$, and M_1, γ, and ω are positive constants,

$$\log PR_{G_a,G_t} < -(k_{a,N} - k_{t,N})(2 + \gamma - \omega/2M_1)\log q_N\{1 + o(1)\}$$

$$= -C'(k_{a,N} - k_{t,N})\log q_N,$$

where C' is a constant such that $0 < C' < (2 + \gamma - \omega/2M_1)$. Let $m' = k_{a,N} - k_{t,N}$ and let δ_2 be the upper bound for $\log PR_{G_a,G_t}$. By Assumption 3, the maximum order of any clique be 5, so the number of model parameters is at most $32(q_N - 1)$. Since $1 - O({q_N}^{-\tilde{a}}) > 1 - O({q_N}^{-\tilde{a}}) - O(Q_N^{-Q_N})$, using the binomial theorem, for all of the competing overfitting models we have

$$
\begin{aligned}
\mathrm{P}\left(\max_{G_a:G_t\subsetneq G_a} PR_{G_a,G_t} \leqslant e^{-\delta_2}\right) &> 1 - \sum_{\max_{G_a:G_t\subsetneq G_a}} 10.4\exp\left\{-m'\omega\log q_N/6\right\} \\
&> 1 - \sum_{k_{a,N}=k_{t,N}+1}^{32(q_N-1)} \binom{32(q_N-1)-k_{t,N}}{k_{a,N}-k_{t,N}} 10.4\exp\left\{-m'\omega\log q_N/6\right\} \\
&> 1 - \sum_{m'=1}^{32(q_N-1)-k_{t,N}} \binom{32(q_N-1)-k_{t,N}}{m'} \left[10.4\exp\left\{-\omega\log q_N/6\right\}\right]^{m'} \\
&> 1 - \left[\left(1+10.4{q_N}^{-\omega/6}\right)^{32q_N} - 1\right] \to 1,
\end{aligned}
$$

for q_N increasing with $N \to \infty$, since

$$
\begin{aligned}
\lim_{q_N\to\infty} \frac{\log\left(1+10.4{q_N}^{-\omega/6}\right)}{1/(32q_N)} &= \lim_{q_N\to\infty} \frac{\frac{\left(10.4(-\omega/6){q_N}^{-\omega/6-1}\right)}{1+10.4{q_N}^{-\omega/6}}}{-1/(32{q_N}^2)} \\
&= \lim_{q_N\to\infty} \left(\frac{10.4(-\omega/6)}{-1/32}\right) \frac{{q_N}^{1-\omega/6}}{1+10.4{q_N}^{-\omega/6}} = 0
\end{aligned}
$$

implies $\lim_{q_N\to\infty}\left(1+10.4{q_N}^{-\omega/6}\right)^{32q_N} = 1$, for $\omega > 6$.

Part 2. From Part 1, we have that the largest upper bound for PR_{G_a,G_t} when $G_t \nsubseteq G_a$ is

$$
PR_{G_a,G_t:G_t\nsubseteq G_a} < e^{-D_1 N^*},
$$

where $D_1 = c_m - c' > 0$ is a constant. Let $k'_{a,N}$ be the number of true parameters in $k_{a,N}$.

Then for all underfitting competing models, we have

$$\sum_{G_a:G_t \not\subset G_a} PR_{G_a,G_t} = \sum_{k_{a,N'}=0}^{k_t-1} \binom{k_{t,N}}{k_{a,N'}} \sum_{k_{a,N}-k'_{a,N}=0}^{32(q_N-1)-k_{t,N}} \binom{32(q_N-1)-k_{t,N}}{k_{a,N}-k'_{a,N}} PR_{G_a,G_t:G_t \not\subset G_a}$$

$$< \exp\{Q_N \log 2 - D_1 N^*\} \to 0,$$

as $N \to \infty$. Also, we have that the largest upper for PR_{G_a,G_t} when $G_t \subsetneq G_a$ is

$$PR_{G_a,G_t:G_t \subsetneq G_a} < e^{-D_2(k_{a,N}-k_{t,N})\log q_N},$$

where $1 < D_2 < (2+\gamma-\omega/2M_1)$ is a constant. Let $m' = k_{a,N} - k_{t,N}$. Then for all overfitting competing models, we have

$$\begin{aligned}
\sum_{G_a:G_t \subsetneq G_a} PR_{G_a,G_t} &= \sum_{k_{a,N}=k_{t,N}+1}^{32(q_N-1)} \binom{32(q_N-1)-k_{t,N}}{k_{a,N}-k_{t,N}} PR_{G_a,G_t:G_t \subsetneq G_a} \\
&< \sum_{m'=1}^{32(q_N-1)-k_t} \binom{32(q_N-1)-k_t}{m'} \left[e^{-D_2 \log q_N}\right]^{m'} \\
&= \left[1+e^{-D_2 \log q_N}\right]^{32(q_N-1)-k_t} - 1 \\
&< \left[1+{q_N}^{-D_2}\right]^{32q_N} - 1 \\
&< \exp\{32/5{q_N}^{1-D_2}\} - 1 \to 0
\end{aligned}$$

as $q_N \to \infty$, since $1 + x < e^x$ for $x > 0$ and $D_2 > 1$.

From (2.35) we have

$$f(G_t|x) = \frac{I_{G_t}(t+s, N+\alpha)\pi(G_t)}{\sum_{G' \in \mathcal{D}_q} I_{G'}(t+s, N+\alpha)\pi(G')} = \frac{1}{\sum_{G' \in \mathcal{D}_q} \frac{I_{G'}(t+s,N+\alpha)\pi(G')}{I_{G_t}(t+s,N+\alpha)\pi(G_t)}} = \frac{1}{1 + \sum_{G_a \neq G_t} PR_{G_a,G_t}}.$$

When q_N is increasing and G_t is decomposable, we have shown that for any competing decomposable graph G_a, the sum $\sum_{G_a \neq G_t} PR_{G_a,G_t} \xrightarrow{\text{P}} 0$. Therefore, $f(G_t|x) \xrightarrow{\text{P}} 1$ as $N \to \infty$. □

2.5 Theoretical results when the true graph is non-decomposable

In the case when the true graph is non-decomposable, numerical methods are required to compute the Bayes factor. To overcome this issue in the Gaussian case, Fitch et al. (2014) prove that when the true graph is non-decomposable, model selection procedures for decomposable graphs will favour a minimal triangulation of the true graph. In this section, we prove that this is indeed also the case for discrete graphical models. We show that the results from Section 2.4 can be extended to the non-decomposable case for both fixed q and q_N growing with N.

When G_t is non-decomposable, we denote $\mathcal{M}_t$ as the minimal triangulation space of G_t and we let $G_m = (V, E_m)$ be any minimum triangulation of G_t, where $E_m = E_t \cup F$ and

F is a non-empty set of fill-in edges. In this section, we use G_a to denote any competing decomposable graph that is not a minimal triangulation.

Theorem 2.5.1. *Let q be fixed. If the true graph G_t is non-decomposable, then we have the following:*

1. *Let G_m be a minimal triangulation of the true model G_t and G_a be any decomposable graph which is not a minimal triangulation, then $BF_{G_a,G_m} \xrightarrow{P} 0$ as $N \to \infty$.*

2. *If G_{m_1} and G_{m_2} are any two different minimal triangulations of G_t with the same number of fill-in edges, then the Bayes factor between them $BF_{G_{m_1},G_{m_2}}$ is stochastically bounded.*

3. *Let $\mathcal{M}_t$ be the minimal triangulation space of G_t. Then $\sum_{G_m \in \mathcal{M}_t} f(G_m|x) \xrightarrow{P} 1$ as $N \to \infty$.*

Proof of Theorem 2.5.1. Similar to Theorem 2.4.5, this proof requires two cases: when a minimal triangulation is competing with an underfitting model, meaning $G_m \nsubseteq G_a$, and when a minimal triangulation is competing with an overfitting model, meaning $G_m \supsetneq G_a$.

Part 1. In case 1, when $|E_a \cap E_t| < |E_m \cap E_t| = |E_t|$, G_a is an underfitting model and G_m is a minimal triangulation of the true model G_t, which is considered as an overfitting model. Then by Lemma 2.4.1, $BF_{G_a,G_m} \xrightarrow{\mathrm{P}} 0$.

Let $c_m > 0$ be the smallest Kullback-Leibler divergence between an underfitting model G_a and the minimal triangulation G_m, which by Assumption 5 is bounded. By Lemma 2.4.1,

where $\epsilon_N = (18CN^{-1}q \log N)^{1/2}$,

$$\begin{aligned}
\log BF_{G_a,G_m} &< -c_m N^* + C_1 N^* \epsilon_N + C_2 \log N^* + O(1) \\
&< -c_m N^* + c' N^* \\
&= -(c_m - c') N^*,
\end{aligned}$$

where $c' > 0$ be a positive constant. Let δ_1 be the upper bound for $\log BF_{G_a,G_m}$. Therefore, for all of the competing underfitting models, we have

$$\mathrm{P}\left(\max_{G_a : G_m \not\subseteq G_a} BF_{G_a,G_m} \leqslant e^{-\delta_1} \right) > 1 - \exp\{q^2 \log 2 - Cq \log N\} \to 1,$$

as $N \to \infty$, since q is fixed and C is a positive universal constant.

In case 2, when $|E_a \cap E_t| = |E_m \cap E_t| = |E_t|$ and $|E_a| > |E_m|$, then $k_a > k_m$. Since the competing model G_a and the minimal triangulation G_m are both decomposable overfitting models, by Lemma 2.4.2, $BF_{G_a,G_m} \xrightarrow{\mathrm{P}} 0$.

By Lemma 2.4.2, where α is the total fictive counts, and C_1, and C_2 are positive constants,

$$\begin{aligned}\log BF_{G_a,G_m} &< -C_1(k_a - k_m)\log N^* + C_2(k_a - k_m)(1 + \log(\log N)) + O(1) \\ &< -(C_1 - C_2)(k_a - k_m)\log N^* + O(1) \\ &= -C'(k_a - k_m)\log N^* + O(1).\end{aligned}$$

where C' is a constant such that $0 < C' < (C_1 - C_2)$. Let $m' = k_a - k_m$ and let δ_2 be the upper bound for $\log BF_{G_a,G_m}$. Since $1 - O(N^{-\tilde{a}}) > 1 - O(N^{-\tilde{a}}) - O(N^{-q})$, using the binomial theorem, for all of the competing overfitting models that are not a minimal triangulation we have

$$\mathrm{P}\left(\max_{G_a : G_m \subsetneq G_a} BF_{G_a,G_m} \leqslant e^{-\delta_2}\right) > 1 - \left[\left(1 + N^{-\omega}\right)^{2^q} - 1\right] \to 1,$$

as $N \to \infty$, since q is fixed and $\lim_{N\to\infty}\left(1 + N^{-\omega}\right)^{2^q} = 1$.

Part 2. Let G_{m_1} and G_{m_2} be two different minimum triangulations of G_t with the same number of fill-in edges, with corresponding vectors of marginal frequency estimates $\widehat{p}^*_{m_1}$ and $\widehat{p}^*_{m_2}$. By Lemma 2.3.2, we have

$$\begin{aligned}\log BF_{G_{m_1},G_{m_2}} &= \ell(\widehat{p}_{m_1}) - \ell(\widehat{p}_{m_2}) + \frac{k_{m_2} - k_{m_1}}{2}\log N^* + O(1) \\ &= \ell(\widehat{p}_{m_1}) - \ell(\widehat{p}_{m_2}) + O(1),\end{aligned}$$

since $|E_{m_1}| = |E_{m_2}|$ and $k_{m_1} = k_{m_2}$.

Let G_{m_c} be the graph such that it contains the same edges as the true non-decomposable graph G_t and the cycles in G_t which are greater or equal to 3 are complete subgraphs. Therefore, $G_{m_1}, G_{m_2} \subsetneq G_{m_c} \subsetneq G_c$, where G_c is the complete graph, and $k_{m_c} - k_{m_1} = k_{m_c} - k_{m_2}$ is finite. For the model G_{m_1}, we have

$$\ell(\widehat{p}^*_{m_1}) = N^* \sum_{C \in \mathcal{C}} \sum_{i_C \in I_C} \widehat{p}^{*C}_{m_1}(i_C) \log \widehat{p}^{*C}_{m_1}(i_C) - N^* \sum_{S \in \mathcal{S}} \nu(S) \sum_{i_S \in I_S} \widehat{p}^{*S}_{m_1}(i_S) \log \widehat{p}^{*S}_{m_1}(i_S),$$

similarly for G_{m_2} and G_{m_c}. Since G_{m_1} and G_{m_2} are both overfitting models, we follow the proof of Lemma 2.4.2. We can write

$$\begin{aligned}
|\ell(\widehat{p}^*_{m_1}) - \ell(\widehat{p}^*_{m_2})| &= |\ell(\widehat{p}^*_{m_1}) - \ell(\widehat{p}_{m_1}) - \{\ell(\widehat{p}^*_{m_2}) - \ell(\widehat{p}_{m_2})\} - \{\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_1})\} \\
&\quad + \ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_2})| \\
&\leqslant |\ell(\widehat{p}^*_{m_1}) - \ell(\widehat{p}_{m_1})| + |\ell(\widehat{p}^*_{m_2}) - \ell(\widehat{p}_{m_2})| + |\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_1})| \\
&\quad + |\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_2})|
\end{aligned}$$

By Assumption 4,

$$|\ell(\widehat{p}^*_{m_1}) - \ell(\widehat{p}_{m_1})| < 2^5(q-1)C'N^* \frac{\alpha}{N^*} = 2^5(q-1)C'\alpha$$

and

$$\left|\ell(\widehat{p}^*_{m_2}) - \ell(\widehat{p}_{m_2})\right| < 2^5(q-1)C''N^*\frac{\alpha}{N^*} = 2^5(q-1)C''\alpha,$$

where C' and C'' are positive constants.

Let $\xi_1 = N^{-1/2}H_{m_c}U_{m_c}$ and $A_1 = H^{1/2}_{m_c}(H^{-1}_{m_c} - D^T_{m_c}H^{-1}_{m_1}D_{m_c})H^{1/2}_{m_c}$, where $D = (I_{k_{m_1}}, 0_{k_{m_1},k_{m_c}-k_{m_1}})$, H denotes the Hessian matrix and U denotes the score vector. Also, $tr(A_1) = (k_{m_c} - k_{m_1})$. As seen in the proof of Lemma 2.4.2, we know that $\xi_1 \sim subGaussian(1/M_1^2)$. By Lemma 4.2.1, we can write $|\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_1})|$ as $\frac{1}{2}\xi_1^T A_1\xi_1\{1 + o(1)\}$ with probability $O(N^{-q})$.

Let $\xi^* = M_1\xi$, then ξ^* satisfies the exponential moment condition

$$\log \mathrm{E}\left(\exp\{s^T\xi^*\}\right) \leqslant ||s||_2^2/2, \qquad s \in \mathbb{R}^{|J|},\ ||s||_2^2 \leqslant g$$

required for Corollary 4.2 (Spokoiny and Zhilova, 2013) for any real number $g > 0$. Corollary 4.2 states that

$$\mathrm{P}\left(|\xi_1^{*T}A_1\xi_1^*| \geqslant (k_{m_c} - k_{m_1}) + K\right) \leqslant 2\exp\{-K/6\} + 8.4\exp\{-x_c\},$$

with $6x_c > K > [2(k_{m_c} - k_{m_1})]^{1/2}/3$ and $x_c > g^2/4$ for large N. We choose $g = (4q\log N)^{1/2}$

and $K = (k_{m_c} - k_{m_1})C_\epsilon$, where $\sqrt{\frac{2}{9(k_{m_c}-k_{m_1})}} < C_\epsilon < -\frac{6}{(k_{m_c}-k_{m_1})}\log\frac{\epsilon}{20.8}$ for any $0 < \epsilon < 1$. Therefore, $(k_{m_c} - k_{m_1})C_\epsilon > [2(k_{m_c} - k_{m_1})]^{1/2}/3$ and $q\log N > (k_{m_c} - k_{m_1})C_\epsilon > \sqrt{\frac{2(k_{m_c}-k_{m_1})}{9}}$ for large N. Since $K/6 < x_c$ implies that $e^{-K/6} < e^{-x_c}$, we have

$$\begin{aligned}\mathrm{P}\left(|\xi_1^{*T} A_1 \xi_1^*| \geqslant (k_{m_c} - k_{m_1})(1 + C_\epsilon)\right) &\leqslant 10.4\exp\{-(k_{m_c} - k_{m_1})C_\epsilon/6\} \\ &< \epsilon/2.\end{aligned}$$

Let $\xi_2^{*T} A_2 \xi_2^*$ be the quadratic form of $\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_2})$ defined analogously to the quadratic form for G_{m_1}. Then

$$\begin{aligned}\mathrm{P}\left(|\xi_2^{*T} A_2 \xi_2^*| \geqslant (k_{m_c} - k_{m_2})(1 + C_\epsilon)\right) &\leqslant 10.4\exp\{-(k_{m_c} - k_{m_2})C_\epsilon/6\} \\ &< \epsilon/2.\end{aligned}$$

Since $k_{m_c} - k_{m_1} = k_{m_c} - k_{m_2}$, we can write

$$\begin{aligned}|\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_1})| + |\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_2})| &= |\xi_1^{*T} A_1 \xi_1^*| + |\xi_2^{*T} A_2 \xi_2^*| \\ &< 2(k_{m_c} - k_{m_1})(1 + C_\epsilon).\end{aligned}$$

Let $C_1 = 2^5(q-1)C' + 2^5(q-1)C''$ and $C_2 = \frac{(k_{m_c}-k_{m_1})}{2M_1}$ be constants because q is fixed. Since $|\log BF_{m_1,m_2}| \leqslant |\ell(\widehat{p}_{m_1}) - \ell(\widehat{p}_{m_2})| + O(1)$, then

$$\mathrm{P}\left(|\log BF_{m_1,m_2}| < C_1\alpha + C_2(1+C_\epsilon) + O(1)\right) > 1 - \epsilon/2 - \epsilon/2$$

Therefore,

$$\mathrm{P}\left(\exp\{-[C_1\alpha + C_2(1+C_\epsilon) + O(1)]\} < BF_{m_1,m_2} < \exp\{C_1\alpha + C_2(1+C_\epsilon) + O(1)\}\right)$$
$$> 1 - \epsilon.$$

Part 3. From Part 1, we have that the largest upper bound for BF_{G_a,G_m} when $G_m \nsubseteq G_a$ is

$$BF_{G_a,G_m:G_m \nsubseteq G_a} < e^{-D_1 N^*},$$

where $D_1 = c_m - c' > 0$ is a constant. Let k'_a be the number of true parameters in k_a. Then for all underfitting competing models, we have

$$\sum_{G_a:G_m \nsubseteq G_a} BF_{G_a,G_m} = \sum_{k'_a=0}^{k_t-1} \binom{k_m}{k'_a} \sum_{k_a-k'_a=0}^{(2^q-1)-k_m} \binom{(2^q-1)-k_m}{k_a-k'_a} BF_{G_a,G_t:G_m \nsubseteq G_a}$$
$$< \exp\{q^2 \log 2 - D_1 N^*\} \to 0,$$

as $N \to \infty$. Also, we have that the largest upper bound for BF_{G_a,G_m} when $G_m \subsetneq G_a$ is

$$BF_{G_a,G_m:G_m \subsetneq G_a} < e^{-D_2(k_a-k_m)\log N^*},$$

where $1 < D_2 < (C_1 - C_2)$ is a constant. Let $m' = k_a - k_m$. Then for all overfitting competing models, we have

$$\begin{aligned}
\sum_{G_a:G_m \subsetneq G_a} BF_{G_a,G_m} &= \sum_{k_a=k_m+1}^{2^q-1} \binom{(2^q-1)-k_m}{k_a-k_m} BF_{G_a,G_m:G_m \subsetneq G_a} \\
&< \sum_{m'=1}^{(2^q-1)-k_m} \binom{(2^q-1)-k_m}{m'} \left[e^{-D_2 \log N^*}\right]^{m'} \\
&= \left[1 + e^{-D_2 \log N^*}\right]^{(2^q-1)-k_m} - 1 \\
&< \left[1 + N^{*-D_2}\right]^{2^q} - 1 \\
&< \exp\{2^q N^{*-D_2}\} - 1 \to 0,
\end{aligned}$$

as $N \to \infty$, since $1 + x < e^x$ for $x > 0$, q is finite and $D_2 > 1$. Therefore,

$$\sum_{G_a:G_a \neq G_m} BF_{G_a,G_m} \xrightarrow{\mathrm{P}} 0, \quad \text{as} \quad N \to \infty \tag{2.36}$$

Let $G_{m_1}, G_{m_2}, ..., G_{m_l}$ be all the minimal triangulations of G_t, where l is a positive integer and $G_m \in \mathcal{M}_t$. Then similar to the proof of Theorem 2.5.2, we have

$$\sum_{G_m \in \mathcal{M}_t} f(G_m|x) = \frac{1}{1 + \sum_{G_a \notin \mathcal{M}_t} \frac{I_{G_a}(t+s, N+\alpha)}{\sum_{i=1}^{l} I_{G_{m_i}}(t+s, N+\alpha)}} \tag{2.37}$$

and $BF_{G_{m_i},G_a} \xrightarrow{\text{P}} \infty$, for $i = 1, 2, \ldots, l$. Therefore,

$$\sum_{G_m \in \mathcal{M}_t} f(G_m|x) = \frac{1}{1 + \sum_{G_a \notin \mathcal{M}_t} \frac{1}{\sum_{i=1}^{l} BF_{G_{m_i},G_a}}} \xrightarrow{\text{P}} 1 \tag{2.38}$$

for fixed q as $N \to \infty$.

□

Part 1 of Theorem 2.5.1 states that for fixed q when G_t is non-decomposable, the Bayes factor will favour a minimal triangulation over any other competing decomposable graph. This is compatible with our results from Section 2.4 since a minimal triangulation contains all of the true edges and the least number of possible false edges. Part 2 of the theorem states that the Bayes factor between two minimal triangulations with the same number of fill-in edges is stochastically bounded, meaning the Bayes factor is bounded by a constant C_ϵ, for any $0 < \epsilon < 1$. To prove this, we use the same approach as we do for comparing two overfitting models. The simulation results in Fitch et al. (2014) and Niu et al. (2021) suggest that one minimal triangulation being favoured over the other is data-dependent. Part 3 states that the posterior probability of competing models will eventually concentrate within the minimal triangulation space.

Theorem 2.5.2 provides the equivalent results for the high-dimensional case, when $q_N \to \infty$

as $N \to \infty$. Similar to Part 2 of Theorem 2.4.6, to prove Part 3 of Theorem 2.5.2, we require $\gamma > (\omega/2M_1 - 1)$.

Theorem 2.5.2. *Let q_N be increasing with N. If the true graph G_t is non-decomposable, then we have the following:*

1. *Let G_m be a minimal triangulation of the true model G_t and G_a be any decomposable graph which is not a minimal triangulation, then $PR_{G_a,G_m} \xrightarrow{P} 0$ as $N \to \infty$.*

2. *If G_{m_1} and G_{m_2} are any two different minimal triangulations of G_t with the same finite number of fill-in edges, then the posterior odds ratio between them $PR_{G_{m_1},G_{m_2}}$ is stochastically bounded.*

3. *Let $\mathcal{M}_t$ be the minimal triangulation space of G_t. Then $\sum_{G_m \in \mathcal{M}_t} f(G_m|x) \xrightarrow{P} 1$ as $N \to \infty$.*

Proof of Theorem 2.5.2. This proof requires two cases: when a minimal triangulation is competing with an underfitting model, meaning $G_m \not\subseteq G_a$, and when a minimal triangulation is competing with an overfitting model, meaning $G_m \subsetneq G_a$.

Part 1. In case 1, when $|E_a \cap E_t| < |E_m \cap E_t| = |E_t|$, G_a is an underfitting model and G_m is a minimal triangulation of the true model G_t, which is considered as an overfitting model. Then by Lemma 2.4.3, $PR_{G_a,G_m} \xrightarrow{\mathrm{P}} 0$.

Let $c_m > 0$ be the smallest Kullback-Leibler divergence between an underfitting model G_a and the minimal triangulation G_m, which by Assumption 5 is bounded. By Lemma 2.4.3,

where $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$, and c_2, and γ are positive constants,

$$\begin{aligned}\log PR_{G_a,G_m} &< -c_m N^* + 2^6(q_N - 1)[c_2 + (N^* + 1)\epsilon_N] \\ &\quad + (k_m - k_a)\big[1/2 \log N^* + \gamma \log q_N + O(1)\big] \\ &< -(c_m - c')N^*,\end{aligned}$$

where $c' > 0$ be a positive constant. Let δ_1 be the upper bound for $\log PR_{G_a,G_m}$. Therefore, for all of the competing underfitting models, we have

$$\begin{aligned}\mathrm{P}\left(\max_{G_a : G_m \not\subset G_a} PR_{G_a,G_m} \leqslant e^{-\delta_1}\right) &> 1 - 2^{Q_N} \exp\{-CQ_N \log Q_N\} \\ &> 1 - \exp\{Q_N \log 2 - CQ_N \log Q_N\} \to 1,\end{aligned}$$

as q_N increases with $N \to \infty$, where C is a positive universal constant.

In case 2, when $|E_a \cap E_t| = |E_m \cap E_t| = |E_t|$ and $|E_a| > |E_m|$, then $k_{a,N} > k_{m,N}$. Since the competing model G_a and the minimal triangulation G_m are both decomposable overfitting models, by Lemma 2.4.2, $PR_{G_a,G_m} \xrightarrow{\mathrm{P}} 0$.

By Lemma 2.4.4, where $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$, and M_1, γ and ω are positive constants,

$$\begin{aligned}\log PR_{G_a,G_m} &< -(k_{a,N} - k_{m,N})(2 + \gamma - \omega/2M_1) \log q_N \{1 + o(1)\} \\ &< -C'(k_{a,N} - k_{m,N}) \log q_N,\end{aligned}$$

where C' is a constant such that $0 < C' < (2 + \gamma - \omega/2M_1)$. Let $m' = k_{a,N} - k_{m,N}$ and let δ_2 be the upper bound for $\log PR_{G_a,G_m}$. Since $1 - O({q_N}^{-\tilde{a}}) > 1 - O({q_N}^{-\tilde{a}}) - O(Q_N^{-Q_N})$, using the binomial theorem, for all of the competing overfitting models that are not a minimal triangulation we have

$$\mathrm{P}\left(\max_{G_a : G_m \subsetneq G_a} PR_{G_a,G_m} \leqslant e^{-\delta_2}\right) > 1 - \left[\left(1 + 10.4{q_N}^{-\omega/6}\right)^{32 q_N} - 1\right] \to 1,$$

for q_N increasing with $N \to \infty$.

Part 2. Let G_{m_1} and G_{m_2} be two different minimum triangulations of G_t with the same number of fill-in edges, with corresponding vectors of marginal frequency estimates $\widehat{p}^*_{m_1}$ and $\widehat{p}^*_{m_2}$. By Lemma 2.3.2, we have

$$\begin{aligned}\log BF_{G_{m_1},G_{m_2}} &= \ell(\widehat{p}_{m_1}) - \ell(\widehat{p}_{m_2}) + \frac{k_{m_{2,N}} - k_{m_{1,N}}}{2}(\log N^* + O(1)) + C_{m_1,N} - C_{m_2,N} \\ &= \ell(\widehat{p}_{m_1}) - \ell(\widehat{p}_{m_2}) + C_{m_1,N} - C_{m_2,N},\end{aligned}$$

since $|E_{m_{1,N}}| = |E_{m_{2,N}}|$ and $k_{m_{1,N}} = k_{m_{2,N}}$, where $|C_{m_1,N} - C_{m_2,N}| \leqslant 2^6(q_N - 1)(c_2 + \epsilon_N)$ for $\epsilon_N = (18CN^{-1}Q_N \log Q_N)^{1/2}$, with probability $1 - q_N O(Q_N^{-Q_N})$.

Let G_{m_c} be the graph such that it contains the same edges as the true non-decomposable graph G_t and the cycles in G_t which are greater or equal to 3 are complete subgraphs. Therefore, $G_{m_1}, G_{m_2} \subsetneq G_{m_c} \subsetneq G_c$, where G_c is the complete graph, and $k_{m_c,N} - k_{m_1,N} = k_{m_c,N} - k_{m_2,N}$ is finite. For the model G_{m_1}, we have

$$\ell(\widehat{p}^*_{m_1}) = N^* \sum_{C \in \mathcal{C}} \sum_{i_C \in I_C} \widehat{p}^{*C}_{m_1}(i_C) \log \widehat{p}^{*C}_{m_1}(i_C) - N^* \sum_{S \in \mathcal{S}} \nu(S) \sum_{i_S \in I_S} \widehat{p}^{*S}_{m_1}(i_S) \log \widehat{p}^{*S}_{m_1}(i_S),$$

similarly for G_{m_2} and G_{m_c}. Since G_{m_1} and G_{m_2} are both overfitting models, we follow the proof of Lemma 2.4.2. We can write

$$\begin{aligned}
|\ell(\widehat{p}^*_{m_1}) - \ell(\widehat{p}^*_{m_2})| &= |\ell(\widehat{p}^*_{m_1}) - \ell(\widehat{p}_{m_1}) - \{\ell(\widehat{p}^*_{m_2}) - \ell(\widehat{p}_{m_2})\} - \{\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_1})\} \\
&\quad + \ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_2})| \\
&\leqslant |\ell(\widehat{p}^*_{m_1}) - \ell(\widehat{p}_{m_1})| + |\ell(\widehat{p}^*_{m_2}) - \ell(\widehat{p}_{m_2})| + |\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_1})| \\
&\quad + |\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_2})|
\end{aligned}$$

Similar to the proof of Lemma 2.4.4, for a constant $c_3 \geqslant 2^5$,

$$|\ell(\widehat{p}^*_{m_1}) - \ell(\widehat{p}_{m_1})| + \left|\ell(\widehat{p}^*_{m_2}) - \ell(\widehat{p}_{m_2})\right| \leqslant 2c_3(q_N - 1)\alpha.$$

Let $\xi_1 = N^{-1/2} H_{m_c} U_{m_c}$ and $A_1 = H^{1/2}_{m_c}(H^{-1}_{m_c} - D^T_{m_c} H^{-1}_{m_1} D_{m_c}) H^{1/2}_{m_c}$, where $D = (I_{k_{m_1,N}}, 0_{k_{m_1,N}, k_{m_c,N} - k_{m_1,N}})$, H denotes the Hessian matrix and U denotes the score vector. Also, $tr(A_2) = (k_{m_c,N} - k_{m_1,N})$. As seen in the proof of Lemma 2.4.2, we know that $\xi_1 \sim subGaussian(1/M_1^2)$. By Lemma 4.2.1, we can write $|\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_1})|$ as $\frac{1}{2}\xi_1^T A_1 \xi_1 \{1 + o(1)\}$ with probability $O(Q_N^{-Q_N})$.

Let $\xi_1^* = M_1 \xi_1$. We choose $g = (4q_N \log N)^{1/2}$ and $K = (k_{m_c,N} - k_{m_1,N})C_\epsilon$, where $\sqrt{\frac{2}{9(k_{m_c,N} - k_{m_1,N})}} < C_\epsilon < -\frac{6}{(k_{m_c,N} - k_{m_1,N})} \log \frac{\epsilon}{10.4}$ for any $0 < \epsilon < 1$. Therefore, $(k_{m_c,N} - k_{m_1,N})C_\epsilon > [2(k_{m_c,N} - k_{m_1,N})]^{1/2}/3$ and $q \log N > (k_{m_c<N} - k_{m_1,N})C_\epsilon > \sqrt{\frac{2(k_{m_c,N} - k_{m_1,N})}{9}}$ for large N. Since $K/6 < x_c$ implies that $e^{-K/6} < e^{-x_c}$, we have

$$\begin{aligned} \mathrm{P}\left(|\xi_1^{*T} A_1 \xi_1^*| \geqslant (k_{m_c,N} - k_{m_1,N})(1 + C_\epsilon)\right) &\leqslant 10.4 \exp\{-(k_{m_c,N} - k_{m_1,N})C_\epsilon/6\} \\ &< \epsilon/2. \end{aligned}$$

Let $\xi_2^{*T} A_2 \xi_2$ be the quadratic form of $\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_2})$ as defined analogously to the quadratic form for G_{m_1}. Then

$$\mathrm{P}\left(|\xi_2^{*T} A_2 \xi_2^*| \geqslant (k_{m_c,N} - k_{m_2,N})(1 + C_\epsilon)\right) \leqslant 10.4 \exp\{-(k_{m_c,N} - k_{m_2,N})C_\epsilon/6\}$$

$$< \epsilon/2.$$

Since $k_{m_c,N} - k_{m_1,N} = k_{m_c,N} - k_{m_2,N}$, we can write

$$|\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_1})| + |\ell(\widehat{p}_{m_c}) - \ell(\widehat{p}_{m_2})| = |\xi_1^{*T} A_1 \xi_1^*| + |\xi_2^{*T} A_2 \xi_2^*|$$

$$< 2(k_{m_c,N} - k_{m_1,N})(1 + C_\epsilon).$$

We can drop the term from the model prior distribution since $k_{m_1,N} = k_{m_2,N}$ implies that $-(k_{m_1,N} - k_{m_2,N})(\gamma \log q_N - \log 2) = 0$. Since $|\log BF_{m_1,m_2}| \leqslant |\ell(\widehat{p}_{m_1}) - \ell(\widehat{p}_{m_2})| + |C_{m_1,N} - C_{m_2,N}|$ where $C_{m_1,N} = C_{m_2,N}$, then choosing $\alpha = \frac{1}{(q_N - 1)}$, gives

$$\mathrm{P}\left(|\log BF_{m_1,m_2}| < 2c_3 + (k_{m_c,N} - k_{m_1,N})(1 + C_\epsilon)\right) < 1 - \epsilon/2 - \epsilon/2$$

Therefore,

$$\mathrm{P}\left(\exp\{-2c_3 - (k_{m_c,N} - k_{m_1,N})(1 + C_\epsilon)\} < BF_{m_1,m_2} < \exp\{2c_3 + (k_{m_c,N} - k_{m_1,N})(1 + C_\epsilon)\}\right)$$

$$< 1 - \epsilon.$$

Part 3. From Part 1, we know that the largest upper bound for PR_{G_a,G_m} when $G_m \nsubseteq G_a$ is

$$PR_{G_a,G_m:G_m \not\subset G_a} < e^{-D_1 N^*},$$

where $D_1 = c_m - c' > 0$ is a constant. Let ${k_{a,N}}'$ be the number of true parameters in $k_{a,N}$. Then for all underfitting competing models, we have

$$\begin{aligned}\sum_{G_a:G_m \not\subset G_a} PR_{G_a,G_m} &= \sum_{k_{a,N}'=0}^{k_{t,N}-1} \binom{k_{m,N}}{k_{a,N}'} \sum_{k_{a,N}-k_{a,N}'=0}^{32(q_N-1)-k_{m,N}} \binom{32(q_N-1)-k_{m,N}}{k_{a,N}-k_{a,N}'} PR_{G_a,G_t:G_m \not\subset G_a} \\ &< \exp\{Q_N \log 2 - D_1 N^*\} \to 0,\end{aligned}$$

as $N \to \infty$. Also, we have that the largest upper bound for PR_{G_a,G_m} when $G_m \subsetneq G_a$ is

$$PR_{G_a,G_m:G_m \subsetneq G_a} < e^{-D_2(k_{a,N}-k_{m,N}) \log N^*},$$

where $1 < D_2 < (2+\gamma-\omega/2M_1)$ is a constant. Let $m' = k_{a,N} - k_{m,N}$. Then for all overfitting competing models, we have

$$\begin{aligned}
\sum_{G_a:G_m\subsetneq G_a} PR_{G_a,G_m} &= \sum_{k_{a,N}=k_{m,N}+1}^{32(q_N-1)} \binom{32(q_N-1)-k_{m,N}}{k_{a,N}-k_{m,N}} PR_{G_a,G_m:G_m\subsetneq G_a} \\
&< \sum_{m'=1}^{32(q_N-1)-k_m} \binom{32(q_N-1)-k_{m,N}}{m'} \left[e^{-D_2\log q_N}\right]^{m'} \\
&= \left[1+e^{-D_2\log q_N}\right]^{32(q_N-1)-k_{m,N}} - 1 \\
&< \left[1+{q_N}^{-D_2}\right]^{32q_N} - 1 \\
&< \exp\{32{q_N}^{1-D_2}\} - 1 \to 0, \quad \text{as} \quad N\to\infty,
\end{aligned}$$

since $1+x<e^x$ for $x>0$ and $D_2>1$. Therefore,

$$\sum_{G_a:G_a\neq G_m} PR_{G_a,G_m} \xrightarrow{\text{P}} 0, \quad \text{as} \quad N\to\infty \tag{2.39}$$

Let $G_{m_1}, G_{m_2}, ..., G_{m_l}$ be all the minimal triangulations of G_t, where l is a positive integer and $G_m \in \mathcal{M}_t$. Then

$$\begin{aligned}
\sum_{G_m\in\mathcal{M}_t} f(G_m|x) &= \sum_{G_m\in\mathcal{M}_t} \frac{I_{G_m}(t+s,N+\alpha)\pi(G_m)}{\sum_{G'\in\mathcal{G}} I_{G'}(t+s,N+\alpha)\pi(G')} \\
&= \frac{\sum_{i=1}^{l} I_{G_{m_i}}(t+s,N+\alpha)\pi(G_{m_i})}{\sum_{i=1}^{l} I_{G_{m_i}}(t+s,N+\alpha)\pi(G_{m_i}) + \sum_{G_a\notin\mathcal{M}_t} I_{G_a}(t+s,N+\alpha)\pi(G_a)} \\
&= \frac{1}{1+\sum_{G_a\notin\mathcal{M}_t} \frac{I_{G_a}(t+s,N+\alpha)\pi(G_a)}{\sum_{i=1}^{l} I_{G_{m_i}}(t+s,N+\alpha)\pi(G_{m_i})}}.
\end{aligned} \tag{2.40}$$

From Part 1, we have that $PR_{G_a,G_{m_i}} \xrightarrow{\text{P}} 0$. Therefore,

$$\frac{I_{G_{m_i}}(t+s,N+\alpha)\pi(G_{m_i})}{I_{G_a}(t+s,N+\alpha)\pi(G_a)} = PR_{G_{m_i},G_a} \xrightarrow{\text{P}} \infty, \tag{2.41}$$

for $i = 1, 2, ..., l$. Thus, we have that

$$\sum_{G_m \in \mathcal{M}_t} f(G_m|x) = \frac{1}{1 + \sum_{G_a \notin \mathcal{M}_t} \frac{1}{\sum_{i=1}^{l} PR_{G_{m_i},G_a}}} \xrightarrow{\text{P}} 1 \tag{2.42}$$

for q_N increases with $N \to \infty$.

□

2.6 Simulations

In this section, we give our simulation results for strong model selection consistency when the true graph is decomposable and when the true graph is non-decomposable. We begin with an example of the behaviour of the pairwise Bayes factor for decomposable models with 5 vertices. Additionally, we give examples of pairwise comparisons between graphs with 100 vertices to demonstrate that our results hold in a high-dimensional case.

We assume all random variables are binary and sample the model parameters from a standard normal distribution and divide the values by 5 to ensure that none of the cell probabilities are too small. If the cell probabilities are too small it may affect the Bayes factor's ability to detect the differences between certain parameters in competing models. Then for each scheme, we use Gibbs sampling to generate a data set with 10,000 samples

and burn the first 1,000 samples. In each simulation, we compute the Bayes factor, or the logarithm of the Bayes factor for sample sizes 1000 to 8,000, incremented by 1,000, and take the mean (Ave.) and standard deviation (Sd.) of 100 replications across each sample size. For the pairwise examples, we compute the logarithm of the Bayes Factor because the gamma functions produce large outputs, which are treated as infinity in R. When we simulate the strong consistency results, we can compute the sum of the exponential logarithm of Bayes factors.

Example. Here we consider the behaviour of the pairwise Bayes factor for decomposable models with 5 vertices. In Figure 2.3a, we have the true model $G_t = \{ab, bc, cd, de\}$. We remove the edge $\{cd\}$ from the true model to obtain the competing underfitting model $G_1 = \{ab, bc, de\}$ in Figure 2.3b , and we add the edge $\{ac\}$ to the true model to obtain the competing overfitting model $G_2 = \{abc, cd, de\}$ in Figure 2.3c.

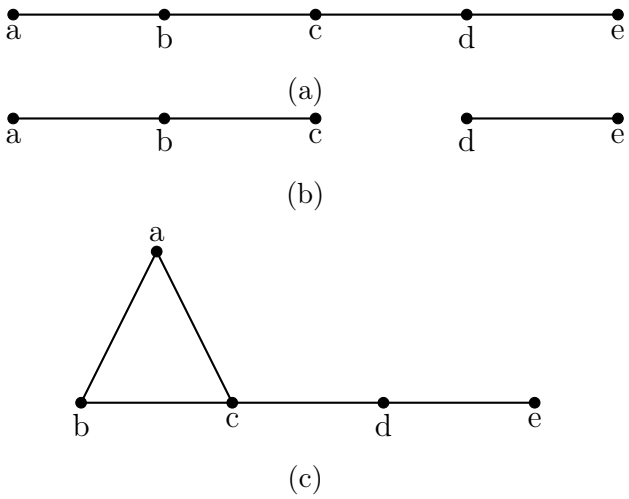


Figure 2.3: Image (a) is the graph G_t, (b) is the graph G_1 and (c) is the graph G_2.

The true model parameters are

$$\beta_0 = \begin{pmatrix} -0.143008907 \\ 0.005672983 \\ 0.174860886 \\ -0.275814717 \\ -0.206133244 \end{pmatrix},$$

where each element of β_0 corresponds to one of the random variables, and

$$\beta = \begin{pmatrix} 0 & 0.03497966 & 0 & 0 & 0 \\ 0.03497966 & 0 & -0.09529885 & 0 & 0 \\ 0 & -0.09529885 & 0 & -0.11242776 & 0 \\ 0 & 0 & -0.11242776 & 0 & -0.01575931 \\ 0 & 0 & 0 & -0.01575931 & 0 \end{pmatrix},$$

where the matrix β gives the values of the two-way interaction terms corresponding to the cliques $C_1 = \{ab\}$, $C_2 = \{bc\}$, $C_3 = \{cd\}$, and $C_4 = \{de\}$.

Table 2.1 gives the results of the pairwise comparisons between the underfitting model G_1 and the true model G_t, and between the overfitting model G_2 and G_t. We remark that in the first case when the competing model is missing a true edge, the Bayes factor converges much faster than the second case when the competing model has one false edge. This is consistent with our theoretical results, which find that the Bayes factor penalizes a missing true edge

more harshly than a false edge.

Table 2.1: Pairwise Bayes factor consistency results for graphs with 5 vertices.

Sample size	Avg. $\log BF_{G_1,G_t}$	Sd. $\log BF_{G_1,G_t}$	Avg. $\log BF_{G_2,G_t}$	Sd. $\log BF_{G_2,G_t}$
1,000	-90.2435	41.62552	-0.8100	0.8270
2,000	-210.6764	67.5859	-1.3042	0.9272
3,000	-347.8014	84.3046	-1.9635	0.7598
4,000	-483.1799	82.4035	-2.3295	0.6000
5,000	-625.8922	102.5758	-2.7459	0.4011
6,000	-752.7225	91.1773	-2.8961	0.4265
7,000	-914.7118	79.5608	-3.1285	0.3921
8,000	-1055.2129	63.9394	-3.2839	0.2811

Example. Here we examine the behaviour of pairwise Bayes factors when the true graph is non-decomposable. We assume the true graph is the smallest non-decomposable graph with 4 vertices; that is, $G_t = \{ab, ac, bd, cd\}$ which is represented in Figure 2.4. The true model parameters are

$$\beta_0 = \begin{pmatrix} -0.0155838 \\ -0.3616984 \\ -0.2489833 \\ 0.2883302 \end{pmatrix},$$

where each element of β_0 corresponds to one of the random variables, and

$$\beta = \begin{pmatrix} 0 & 0.3589437 & 0.1783851 & 0 \\ 0.3589437 & 0 & 0 & -0.3984551 \\ 0.1783851 & 0 & 0 & 0.2008634 \\ 0 & -0.3984551 & 0.2008634 & 0 \end{pmatrix},$$

where the matrix β gives the values of the two-way interaction terms corresponding to the pairs of vertices $\{ab\}$, $\{ac\}$, $\{bd\}$, and $\{cd\}$.

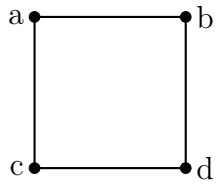
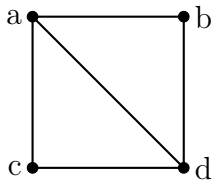
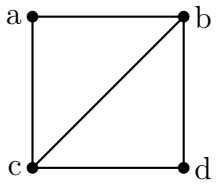


Figure 2.4: Visual representation of G_t.

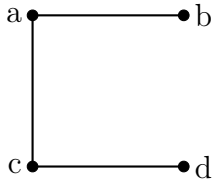
In Figures 2.5a and 2.5b we have the minimal triangulations $G_{m_1} = \{abd, acd\}$ and $G_{m_2} = \{abc, bcd\}$, respectively. Then in Figure 2.5c we have an underfitting model $G_a = \{ab, ac, cd\}$ with one missing true edge edge, and in Figure 2.5d we have an underfitting model $G_b = \{ab, bc, cd\}$ with two missing true edges and one false edge.



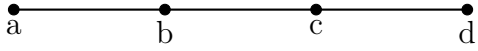
(a) Visual representation of G_{m_1}.



(b) Visual representation of G_{m_2}.



(c) Visual representation of G_a.



(d) Visual representation of G_b.

Figure 2.5: The two minimal triangulations of G_t and two competing models.

In Table 2.2 and Table 2.3, we have the results for each minimal triangulation competing with model G_a and G_b, respectively. We see that the pairwise Bayes factor converges to 0 in each case and it converges at similar rates. Finally, in Table 2.4, we observe that the Bayes factor between two minimal triangulations is bounded by constants close to 0.

Table 2.2: Comparing the underfitting model G_a to the minimal triangulations G_{m_1} and G_{m_2}.

Sample size	Avg. $\log BF_{G_a,G_{m_1}}$	Sd. $\log BF_{G_a,G_{m_1}}$	Avg. $\log BF_{G_a,G_{m_2}}$	Sd. $\log BF_{G_a,G_{m_2}}$
1,000	-3.0907	3.1218	-3.1660	2.9711
2,000	-7.1842	3.9829	-7.2119	4.8540
3,000	-11.3777	4.5142	-11.1513	5.0440
4,000	-14.8443	4.3048	-14.6982	4.4552
5,000	-19.2345	4.8599	-20.3070	4.4748
6,000	-23.8811	4.9440	-24.1890	4.3770
7,000	-29.1548	4.2772	-28.0810	4.3526
8,000	-33.4932	3.3035	-32.6125	2.9264

Table 2.3: Comparing the underfitting model G_b to the minimal triangulations G_{m_1} and G_{m_2}.

Sample size	Avg. $\log BF_{G_b,G_{m_1}}$	Sd. $\log BF_{G_b,G_{m_1}}$	Avg. $\log BF_{G_b,G_{m_2}}$	Sd. $\log BF_{G_b,G_{m_2}}$
1,000	-4.1105	3.0185	-3.8495	3.4017
2,000	-7.1983	4.3379	-7.8876	3.5872
3,000	-12.6721	5.4692	-11.5883	4.4302
4,000	-16.7995	4.4322	-16.1975	4.7767
5,000	-22.5098	5.8016	-21.0118	4.8093
6,000	-25.4583	4.6616	-25.6652	4.6348
7,000	-31.5930	4.4213	-31.0920	4.4580
8,000	-35.4698	2.8318	-35.8314	2.7732

Table 2.4: Comparison between the minimal triangulations G_{m_1} and G_{m_2}.

Sample size	Avg. $\log BF_{G_{m_1},G_{m_2}}$	Sd. $\log BF_{G_{m_1},G_{m_2}}$
1,000	-0.2058	1.7574
2,000	-0.1069	1.4370
3,000	0.2001	1.5596
4,000	0.2162	1.2121
5,000	0.1577	1.3021
6,000	0.0364	1.0890
7,000	0.1746	0.8681
8,000	0.1537	0.6723

Example. In this example, we demonstrate strong model selection consistency for decomposable models with 3 vertices. We arbitrarily chose $G_t = \{ac, bc\}$ to be the true graph, which is represented in Figure 2.6. There are 7 different competing models corresponding to decomposable graphs with 3 vertices. Namely, $G_1 = \{a, b, c\}$, $G_2 = \{ab, c\}$, $G_3 = \{bc, a\}$, $G_4 = \{ac, b\}$, $G_5 = \{ab, bc\}$, $G_6 = \{ab, ac\}$, and $G_7 = \{abc\}$. From the proof of Theorem 2.4.5,

we know that in order to prove that $f(G_t|x) \to 1$, we need to show that $\sum_{G_a \neq G_t} BF_{G_a,G_t} \to 0$.

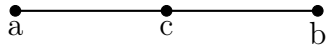


Figure 2.6: Visual representation of G_t.

The true model parameters are

$$\beta_0 = \begin{pmatrix} 0.03307030 \\ -0.09786483 \\ 0.29487145 \end{pmatrix},$$

where each element of β_0 corresponds to one of the random variables, and

$$\beta = \begin{pmatrix} 0 & 0 & 0.3885803 \\ 0 & 0 & 0.4318364 \\ 0.3885803 & 0.4318364 & 0 \end{pmatrix},$$

where the matrix β gives the values of the interaction terms corresponding to the cliques $C_1 = \{ab\}$ and $C_2 = \{bc\}$.

Table 2.5 gives the results for strong model selection consistency for decomposable graphs. We see that indeed the sum of the Bayes factors for all the competing models with 3 vertices converge to 0 as the sample size increases. Therefore, we can conclude that $f(G_t|x) \to 1$ as the sample size increases.

Table 2.5: Strong model selection consistency results for decomposable graphs with 3 vertices.

Sample size	Avg. $\sum_{G_a \neq G_t} BF_{G_a,G_t}$	Sd. $\sum_{G_a \neq G_t} BF_{G_a,G_t}$
1,000	3.1381	18.4994
2,000	0.6365	1.5304
3,000	0.2799	0.7748
4,000	0.1277	0.1418
5,000	0.0900	0.0763
6,000	0.0621	0.0378
7,000	0.0474	0.0164
8,000	0.0355	0.0077

Example. Next, we demonstrate strong model selection consistency for non-decomposable models with 4 vertices. We assume the true graph is the smallest non-decomposable graph with 4 vertices; that is, $G_t = \{ab, ac, bd, cd\}$ which is represented in Figure 2.4.

Since there are 61 competing decomposable graphs with 4 vertices, for this example, we arbitrarily chose 12 of the connected graphs to be the competing models; namely, $G_1 = \{ab, ac, cd\}$, $G_2 = \{acd, ab\}$, $G_3 = \{ac, bcd\}$, $G_4 = \{ab, bd, cd\}$, $G_5 = \{ac, bd, cd\}$, $G_6 = \{abcd\}$, $G_7 = \{bcd, ab\}$, $G_8 = \{acd, bd\}$, $G_9 = \{abd, ac\}$, $G_{10} = \{abd, cd\}$, $G_{11} = \{acd, bcd\}$, and $G_{12} = \{abc, abd\}$. From the proof of Theorem 2.5.1, we know that in order to prove that $\sum_{G_m \in \mathcal{M}_t} f(G_m|x) \to 1$, we need to show that $\sum_{i=1}^{l} BF_{G_{m_i},G_a} \to \infty$.

Table 2.6 gives the results for strong model selection consistency for non-decomposable graphs. The results show that for 12 out of the 61 possible competing decomposable models, the Bayes factor favours the minimal triangulations. Therefore, this indicates that $\sum_{G_m \in \mathcal{M}_t} f(G_m|x) \to 1$ as the sample size increases.

Table 2.6: Strong model selection consistency results for non-decomposable graphs with 4 vertices.

Sample size	Avg. $\sum_{G_a \notin \mathcal{M}_t} \frac{1}{\sum_{i=1}^{l} BF_{G_{m_i},G_a}}$	Sd. $\sum_{G_a \notin \mathcal{M}_t} \frac{1}{\sum_{i=1}^{l} BF_{G_{m_i},G_a}}$
1,000	0.5264	0.4868
2,000	0.3789	1.8577
3,000	0.0944	0.1772
4,000	0.0466	0.1222
5,000	0.0121	0.0278
6,000	0.0076	0.0210
7,000	0.0012	0.0019
8,000	0.0002	0.0003

Example. Here, to simplify the computations, we use *tree* graphs with 100 vertices to demonstrate the behaviour of the Bayes factor in the high-dimension setting. A tree is a connected acyclic undirected graph with q vertices and $q-1$ edges, and a disjoint union of trees is called a *forest*. In a forest graph, every pair of vertices are connected by at most one path. Consequently, the corresponding models consist of at most two-way interactions, which makes computing the Bayes factor more manageable. We use a built-in function in R to randomly generate tree graphs based on the Barabasi-Albert model which, is one of the algorithms commonly used to generate random scale-free networks.

To create the true model, we randomly generate a tree graph and remove the edge $\{24, 64\}$ so we can easily form a competing underfitting model and a competing overfitting model which are also tree graphs. The true model G_t is represented in Figure 2.7. The graph has two clusters, where most of the graph is connected in the large cluster and vertices {64}, {73},

{74}, and {95} are connected in the smaller cluster. In Figure 2.8, we have the competing underfitting model G_a, which is the same as the true graph, but we remove the edge $\{4, 36\}$. We see that the vertices {36}, {41}, {59}, and {82} now form another component. To form the competing overfitting model G_b, represented in Figure 2.9, we add the edge $\{24, 64\}$ to the true graph G_t.

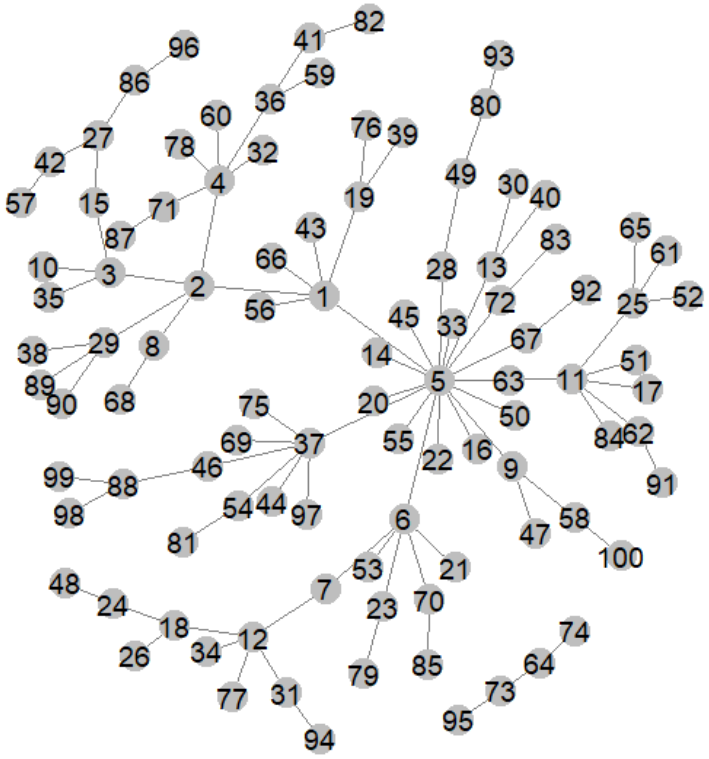


Figure 2.7: Visual representation of G_t.

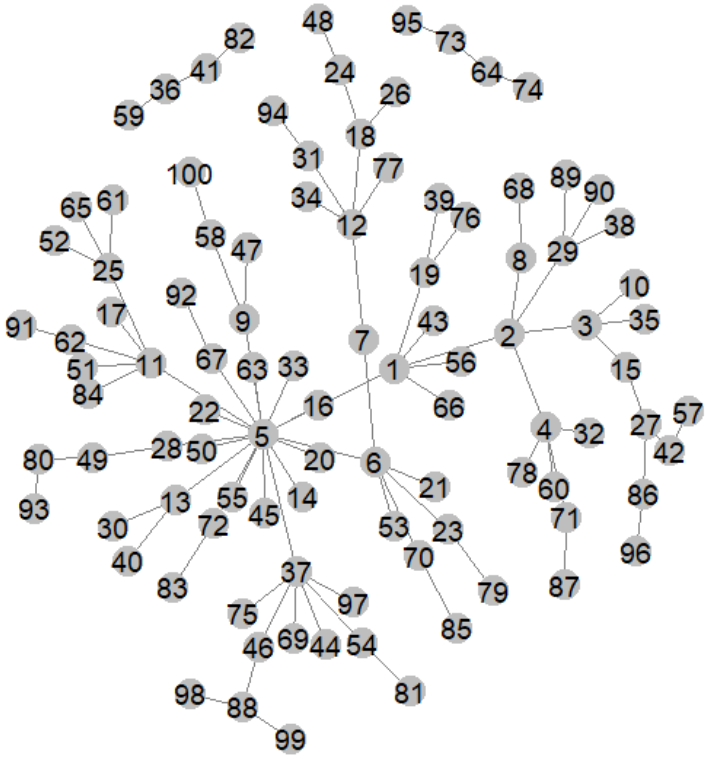


Figure 2.8: Visual representation of G_a.

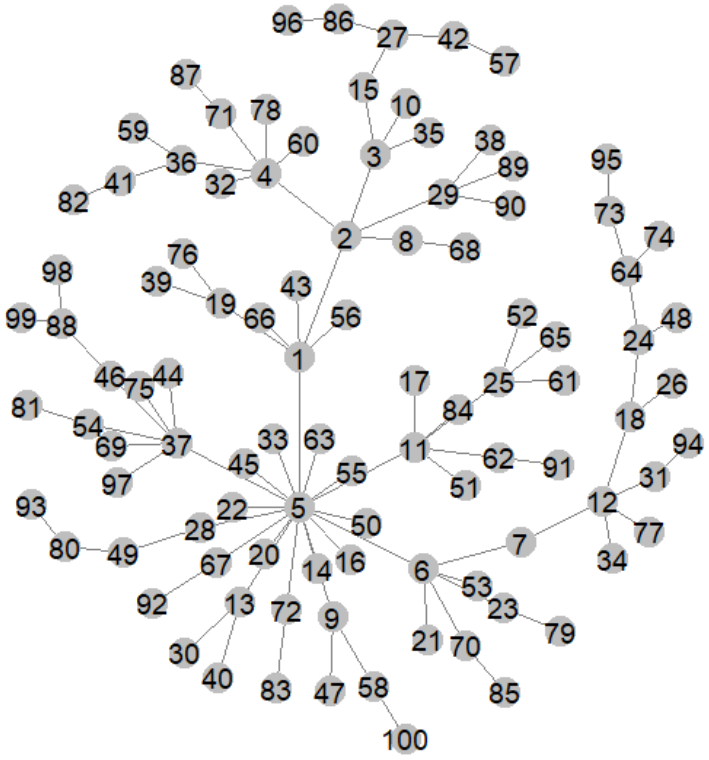


Figure 2.9: Visual representation of G_b.

In Table 2.7, we have the results for pairwise comparison of the underfitting model G_a with G_t, and the overfitting model G_b with G_t. Again, we see that in both cases the log of the Bayes factor tends to negative infinity and it penalizes a missing true more severely than an additional false edge.

Table 2.7: Pairwise Bayes factor consistency results comparing G_a to G_t, and G_b to G_t.

Sample size	Avg. $\log BF_{G_a,G_t}$	Sd. $\log BF_{G_a,G_t}$	Avg. $\log BF_{G_b,G_t}$	Sd. $\log BF_{G_b,G_t}$
1,000	-22.5420	4.8036	-14.4729	2.4966
2,000	-41.4994	5.8613	-25.8421	2.9886
3,000	-63.2577	8.3956	-39.0688	4.5068
4,000	-79.2360	15.2566	-53.0429	3.7209
5,000	-97.2665	14.6249	-61.4124	4.5596
6,000	-118.4968	15.1862	-75.2714	4.0259
7,000	-137.6191	17.4938	-89.1145	4.9238
8,000	-156.4441	17.8156	-101.8007	4.4059

2.7 Conclusion

In our research, we introduce Bayesian model selection consistency results for high-dimensional discrete graphical models. We use the approximation of the logarithm of a gamma function to express the logarithm of the normalizing constant $I_G(t+s, N+\alpha)$, which is proportionate to the posterior probability. This permits us to derive a convenient expression for the logarithm of the Bayes factor, and to easily examine the behaviour of the Bayes factor when q is fixed and the behaviour of the posterior odds ratio when q_N is increasing. We establish pairwise consistency and strong model selection consistency for discrete graphical models, where the data is obtained from a contingency table, for both fixed dimensional models and for high-dimensional models. Moreover, our results demonstrate that when the true graph is non-decomposable, it can be reasonably approximated by one of its minimal triangulations and the difference between minimal triangulations with the same number of edges is negligible.

Many model selection algorithms are restricted to the class of decomposable models; however, our method allows these algorithms to also analyse non-decomposable models.

Chapter 3

Graphical Local Genetic Algorithm

Graphical log-linear models are an effective tool for representing complex structures that emerge from high-dimensional data. It is challenging to fit an appropriate model in the high-dimensional setting and many existing methods rely on a convenient class of models, called decomposable models, which lend well to a stepwise approach. However, these methods restrict the pool of candidate models from which they can search and these methods are difficult to scale. It can be shown that a non-decomposable model can be approximated by the decomposable model which is its minimal triangulation, thus extending the convenient computational properties of decomposable models to any model. In this chapter, we propose a local genetic algorithm with a crossover-hill-climbing operator (Lozano et al., (2004)) for log-linear graphical models. We show that the graphical local genetic algorithm can be used successfully to fit non-decomposable models for both a low number of variables and a high number of variables. We use an expression proportionate to the posterior probability as a

measure of fitness and parallel computing to decrease the computation time.

3.1 Introduction

The most commonly used graphical log-linear models are hierarchical models which are determined by their two-way interactions, meaning for every higher-order term in the model, the model also contains the corresponding lower-order terms. A log-linear model obtained from a q-dimensional contingency table can be represented by an undirected graph $G = (V, E)$ with vertex set $V = \{1, 2, \ldots, q\}$ and edge set $E \subseteq V \times V$. If the graphical log-linear model corresponds to a chordal (decomposable) graph, then it is called a decomposable model; otherwise, it is called nonchordal (non-decomposable). Frequently, graphical model selection consists of forward or backward elimination procedures on decomposable graphs due to the decomposable chain rule, that is, one can construct either an increasing or decreasing sequence of decomposable graphs differing by one edge (Lauritzen (1996)). However, for q variables there are 2^q possible models, thus these methods become computationally intensive for high-dimensional models.

In Gauraha (2016), and Gauraha and Parui (2020), they present a forward selection method for low-dimensional graphs, using the *mutual conditional independence* between vertices to reduce the search space and in turn reduces the computational complexity. Another popular method is the graphical lasso, which is the graphical version of the lasso introduced by Tibshirani (1996). The original approach was for Gaussian graphical models, but it has since been extended to log-linear models with many variations. For example, in

Allen and Liu (2012) they propose the Poisson graphical lasso, and in Dahinden et al. (2010), they offer a variation of the group lasso where they learn subsets of the graph then reconstruct the original graph.

In the high-dimensional setting, Petijean et al. (2013) present their approach, called *Chordalysis*. It is a forward selection method where they use data mining techniques to store and reuse the computed marginal likelihood ratios. They demonstrate that their method is efficient and effective for up to 150 variables. However, the efficiency of their algorithm relies on the decomposable property of the candidate graphs and the sensitivity of the algorithm decreases rapidly with sample size. Dobra and Mohammadi (2018) implement a Birth-Death Markov Chain Monte Carlo (BDMCMC) algorithm using a marginal posterior probability based on the marginal pseudo-likelihood with a Dirichlet prior to define the birth and death probabilities. To speed up their algorithm, they compute all of the possible edges using parallel computing.

Model selection for discrete variables can be particularly challenging because the typical optimization methods borrowed from calculus are not applicable. A popular approach for binary variables is to use a genetic algorithm, first introduced by Holland (1975), which in a metaheuristic procedure that aims to optimize some criterion by imitating Darwinian natural selection. The genetic algorithm is an iterative process where the binary elements represent chromosomes. During each iteration, or *generation*, two candidate *parents* are selected from the population using a measure of fitness. Then their chromosomes are combined using a crossover operation to produce offspring which are subject to random chromosomal mutation.

Finally, the new offspring are introduced into the population. Since its inception, many variations for each step of the genetic algorithm have been introduced.

Poli and Roverato (1998), and Blauth and Pigeot (2002) both proposed genetic algorithms for graphical models for a low number of variables - 10 variables and 6 variables, respectively. Poli and Roverato (1998) used the Akaike's Information Criterion as their measure of fitness and they used the elitism variation, meaning the top 5% of candidates of the current population are kept into the next generation. Their main contribution was how they exploited the hierarchical properties of the candidate models in the crossover step. The parents exchanged randomly selected subsets of their corresponding graphs and thus reducing the required computations. Blauth and Pigeot (2002) used the Bayesian Information Criterion as their measure of fitness and tournament selection which considers the ranks of the candidate chromosomes. Other variations of the genetic algorithm include local searches. Lozano et al. (2004) give a real-coded local genetic algorithm and García-Martínez (2006) give a binary-coded local genetic algorithm. The key in both these contributions is to balance the diversity of the global search while fine-tuning the local search. In the local search, they propose what they call the *crossover-hill-climbing operator*, meaning the most fit offspring replaces the worst parent and reproduces with the best parent for a predetermined number of iterations. In Lozano et al. (2004), they use negative assortment mating, meaning the candidate parents that are selected are the most different from each other. Conversely, in García-Martínez (2006), they use positive assortment mating.

In the following chapter, we present a local genetic algorithm which implements the

crossover-hill-climbing operator from Lozano et al. (2004) for log-linear graphical models, using the log of a nomalizing constant proportionate to the posterior probability as a measure of fitness. Since the genetic algorithm has no stepwise component, we are not constrained to the class of decomposable models. It can be shown that a model corresponding to a non-decomposable graph can be approximated by its minimal triangulation, which is by definition a decomposable graph. This allows us to benefit from convenient properties of decomposable graphs when computing the posterior probability, while also being able to consider a wider variety of candidate models. In order to focus the search, we use what we called the *edgewise* Bayes factor to initiate the candidate models. In the low-dimensional setting, we perform our algorithm on the entire graph and in the high-dimensional setting, we find appropriate candidate subsets of the graph, then reconstruct the full graph from a predetermined number of subsets. In Section 3.2, we give an overview of log-linear graphical models and define the posterior probability for a decomposable graph. Then we describe the graphical local genetic algorithm and how we use adjacency matrices to perform each step of the algorithm. In Section 3.3, we give our experiment results. We perform simulations for the number of variables $q \in \{6, 8, 12, 20, 50, 100\}$ in Section 3.3.1 and we apply our algorithm to a real data set in Section 3.3.2. We conclude in Section 3.4.

3.2 Materials and Methods

In this section, we first describe the log-linear graphical model, and give the necessary background from graph theory to illustrate how we compute the posterior probability cor-

responding to both decomposable and non-decomposable graphs. For additional details on graph theory and graphical models, see Lauritzen (1996). Then we explain the global and local components of the Graphical Local Genetic Algorithm (GLGA), and how we implement the algorithm in a high-dimensional setting. Finally, we discuss the complexity of the algorithm and the computing technique we used to speed up certain steps.

3.2.1 Log-linear Graphical Models

Consider a vector of random variables $X = (X_v, v \in V)$ indexed by the set $V = \{1, 2, \ldots, p\}$ such that each X_v takes values in the finite set I_v with $|I_v|$ levels. Then the resulting counts can be presented in a p-dimensional contingency table corresponding to

$$I = \bigtimes_{v \in V} I_v,$$

where I is the set of cells $i = (i_v, v \in V)$ and $i_v \in I_v$. The number of observations for cell i is denoted $n(i)$ and the probability of an object being observed in cell i is denoted $p(i)$. If $D \subset V$, the set of D-marginal cells is $i_D = (i_v, v \in D)$. For $N = \sum_{i \in I} n(i)$, we assume the cell counts follow a multinomial distribution and the cell probabilities are modelled by a hierarchical log-linear model. For simplicity, in this paper, we assume all random variables are binary.

The conditional independencies between the random variables X_v can be read off an undirected graph $G = (V, E)$ with vertex set V and edge set $E \subseteq V \times V$, that is, if X_a is independent of X_b given $X_{V \setminus \{a,b\}}$, whenever (a, b) is not an edge in E. A graph is *complete*

if every pair of vertices has an edge. The discrete graphical model for X is said to be *decomposable* or Markov with respect to G if it corresponds to a *chordal* or *triangulated* undirected graph, meaning every cycle of length greater or equal to 4 has a chord. Furthermore, a collection of random variables $(X_v)_{v \in V}$ with associated graph G are said to be Markov relative to G if for any triple of disjoint sets (A, B, S),

$$X_A \perp\!\!\!\perp X_B | X_S,$$

where $V = A \cup B \cup S$ and S is a complete subset.

For a graph G and any of its decompositions (A, B, S), we call the subsets A and B cliques, and the subset S a separator. The advantage of using a decomposable model is that the probability distribution of its variables can be written as a product of factors over the cliques $C \in \mathcal{C}$ and the separators $S \in \mathcal{S}$ of the corresponding decomposable graph. This allows for many convenient computational properties. If a graph is non-decomposable, it has been shown that its *minimal triangulation* can be used as a reasonable proxy. The minimal triangulation of a non-decomposable graph is the graph made up of the least number of *fill-in edges* which results in a decomposable graph. Since the minimal triangulation of a non-decomposable graph is by definition decomposable, all of the computational advantages will apply.

For example, the graph in Figure 3.1a is the smallest non-decomposable graph and it has three possible triangulations. Graph 3.1d is the complete graph on four vertices and it is a triangulation of 3.1a; however, it is not minimal. Graphs 3.1b and 3.1c are minimal

since removing the edge (b, c) from 3.1b, or the edge (a, d) from 3.1c will result in a non-decomposable graph. Therefore, if we want to consider the non-decomposable graph 3.1a as a candidate model, then we would calculate the required computations corresponding to either the minimal triangulation 3.1b or 3.1d and use the result to compare graph 3.1a to other competing models.

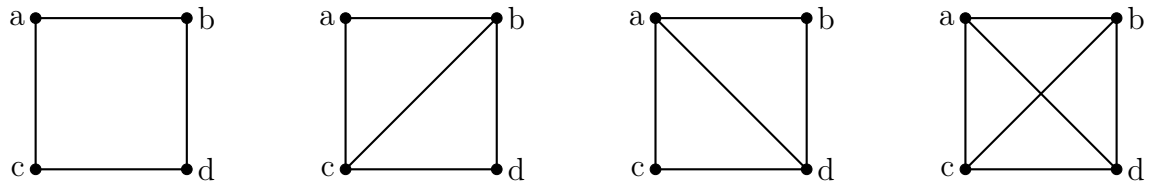


Figure 3.1: The smallest non-decomposable graph and its triangulations.

The implementation of a genetic algorithm requires a measure of fitness. We use an expression proportionate to the posterior probability, $f(G|x)$, which requires an appropriate prior distribution. We choose the Dirichlet distribution on the log-linear parameters because it is the Diaconis-Ylvisaker (DY) conjugate prior (Diaconis and Ylvisaker (1979)). The Dirichlet distribution is parametrized by fictive counts (or pseudocounts), denoted $s(i)$ for $i \in I$, which sum up to α. In Dawid and Lauritzen (1993), they develop the hyper Dirichlet conjugate prior which exhibits the same Markov properties when corresponding to a decomposable model. It can be shown that a posterior probability which is Markov with respect to a decomposable graph G is proportionate to a normalizing constants, denoted $I_G(n + s, N + \alpha)$, and can be written as the product of gamma functions indexed over the cliques and separators of G, that is,

$$f(G|x) \propto I_G(n+s, N+\alpha) = \frac{\prod_{C\in\mathcal{C}} \prod_{i_C \in I_C} \Gamma(n_C(i_C) + s_C(i_C))}{\Gamma(N+\alpha) \prod_{S\in\mathcal{S}} \left[\prod_{i_S \in I_S} \Gamma(n_S(i_S) + s_S(i_S))\right]^{\nu(S)}}, \tag{3.1}$$

where n is the vector of true cell counts, s is the vector of fictive cell counts, and $\nu(S)$ is the multiplicity of separator S. If a model corresponds to a non-decomposable graph, then we compute the the normalizing constant (3.1) for its minimal triangulation.

We denote the number of free parameters in a decomposable model by k and it can be expressed as

$$k = -1 + \sum_{C\in\mathcal{C}} |I_C| - \sum_{S\in\mathcal{S}} \nu(S) \cdot |I_S|. \tag{3.2}$$

Since we assume that all variables are binary, in our simulations we use $|I_C| = 2^{|C|}$ and $|I_S| = 2^{|S|}$. However, the algorithm can be implemented for variables with more than 2 levels using equation (3.2).

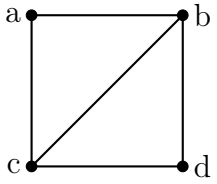
3.2.2 Graphical Local Genetic Algorithm

Genetic algorithms (GAs) belong to the class of evolutionary algorithms and are used to solve optimization and search problems. They mimic the evolutionary process of natural selection, popularized by Charles Darwin. In the original algorithm, developed by Holland (1975), each candidate solution corresponds to an individual in the population which is assumed to have one chromosome. A chromosome is represented by a string of 0's and 1's and each

element of the string is called an allele. In the first generation, the population is randomly initiated and the fitness of each candidate in the population is measured. Then two parents are selected from the population, often the most 'fit' are selected, and they produce *offspring* by a crossover operation. The simplest crossover operation is the one-point crossover where the chromosome of each parent it randomly cut into two segments and switched with one of the segments of the other parent to create offspring. Finally, the offspring are subject to random mutations in one of more of the alleles in their chromosome. Depending on the fitness of the offspring, they may replace existing members of the population or they may simply be added to the population. The algorithm iterates until a predetermined stopping criterion is met. There are many variations of each step in the algorithm. For more details on these variations, see Givens and Hoeting (2013).

Since we are interested in graphical models, we will use adjacency matrices instead of strings of 0's and 1's. An undirected graph $G = (V, E)$ with $|V| = q$ can be represented by a $q \times q$ matrix $A = (a_{ij})$, where $a_{ij} = 1$ if $(i, j) \in E$ and $i \neq j$, and $a_{ij} = 0$ otherwise. For example, consider a graph G_1 with vertices $V = \{a, b, c, d\}$ and cliques $G_1 = \{abc, bcd\}$, as seen in Figure 3.2a. In Figure 3.2b, we have its adjacency matrix with 1's where the element of the matrix corresponds to an edge in the graph, and 0's where the element corresponds to no edge in the graph and 0's on the diagonal.

All of our computations are in R, where it is more practical to use the $\log \Gamma(\cdot)$ function instead of $\Gamma(\cdot)$, so we use the logarithm of the normalizing constant (2.21) to measure the fitness of each model. We use the *igraph* package in R to obtain the cliques and separators



(a)

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

(b)

Figure 3.2: A visual representation of the graph G_1 and the corresponding adjacency matrix.

from the adjacency matrix of a decomposable graph or of a minimal triangulation of a graph, and compute the sum of the log of gamma functions indexed according to the graph's factorization. In the case that a candidate model is non-decomposable and we compute the log of the normalizing constant for its minimal triangulation, we do not consider the minimal triangulation to be an updated version of the candidate model. We use the minimal triangulation only for its convenient computational properties. It has been shown that when comparing overfitting models, the log posterior probability will favour the model with less superfluous edges. However, when comparing underfitting models, it will favour the candidate model with most true edges. This cause the genetic algorithm to be prone to keeping false edges if it means having more true edges. Therefore, we add a penalty term to our measure of fitness to prevent the algorithm from selecting a model with too many unnecessary edges. We use the penalty $-k \log \Gamma(N + \alpha)$, where k is the number of free parameters in the model (3.2).

Since the genetic algorithm does not have the same restrictions as some other model selection methods, it is sensitive to its initial conditions and can consider any possible candidate model. Seeing as our goal is to implement the GLGA in the high-dimensional

setting, we must direct the algorithm towards the more suitable models. To do so, we use the Bayes factor (3.3) to compare the presence of each edge versus no edges. The Bayes factor is the ratio of posterior probabilities (2.21) for two models and it is commonly used in model selection to compare candidate models. When comparing two models G_a and G_b with equal probability, the Bayes factor is a ratio of normalizing constants (2.21), that is,

$$BF_{G_a,G_b} = \frac{I_{G_a}(t+s, N+\alpha)}{I_{G_b}(t+s, N+\alpha)}. \tag{3.3}$$

To initialize the population, we compute the *edgewise* Bayes factor for each possible edge, that is, we compare a candidate model with the single edge in question to the model with no edges. We use ranges of the values of the edgewise Bayes factors to determine the probability of including that edge or not in the otherwise randomly generated initial candidate models, where these ranges depend on the sample size. One of the convenient properties of the Bayes factor is that, for two decomposable models, the cliques or separators common to both models will cancel out and hence not need to be computed. Thus, even if we have a high-dimensional data set, we only need to compute the Bayes factor for the given edge. For example, even if $q = 100$, to compute the Bayes factor for the edge $\{ab\}$, we simply need to compare the existence of the edge $\{ab\}$ versus no edge $\{a, b\}$. Therefore, is it not computationally intensive to compute the Bayes factor for each edge.

In order to perform the crossover step and the mutation step, we use the upper triangular

adjacency matrix. For the crossover step, we randomly select a cut-point and we interchange the rows above and below the cut-point between the two parent matrices. We do this three times using three different cut-points to create six offspring at each crossover step. In Figure 3.3a, we have the upper triangular matrix of the adjacency matrix which represents the graph G_1 from Figure 2.5a in Section 2.2. To continue our example, consider a graph $G_2 = \{ac, ad, cd\}$ with upper triangular adjacency matrix seen in Figure 3.3b. Say we randomly choose to cut G_1 and G_2 between row 1 and row 2, as seen in Figure 3.3a and Figure 3.3b. Then to complete the crossover, we take row 1 from G_1 and rows 2-4 from G_2 to form one offspring (Figure 3.3c), and we take row 1 from G_2 and rows 2-4 from G_1 to form a second offspring (Figure 3.3d). In each crossover step, we do three unique cuts to obtain six offspring. In the mutation step, since the genetic algorithm tends to pick up extra edges when there are missing true edges, we simply take the upper triangular adjacency matrix for each offspring and with a small probability we change 0 to 1. We do not allow for random mutations from 0 to 1.

$$\left(\begin{array}{cccc} 0 & 1 & 1 & 0 \\ \hline 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right) \qquad \left(\begin{array}{cccc} 0 & 0 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right) \qquad \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(a) (b) (c) (d)

Figure 3.3: Example of creating two offspring with one cut-point in the crossover step.

Once we have the six offspring obtained from what we refer to as the global crossover, post-mutation step, if certain conditions are met we perform a local search. We implement the

crossover-hill-climbing step from Lozano et al. (2004), but applied to graphical models. We refer to crossover operations preform during the crossover-hill-climbing step as local crossover. Let p_1 = the current best parent, p_2 =the best current offspring, $n_t = 3$, and $n_{off} = 6$. The pseudo code for the crossover-hill-climbing operator can be found in Figure 3.4.

Crossover-hill-climbing(p_1, p_2, n_{off}, n_t)

1. Select parents p_1 and p_2.
2. Repeat n_t times.
   (a) Generate n_{off} offspring by performing local crossover on p_1 and p_2.
   (b) Evaluate the fitness of each n_{off} offspring.
   (c) Find the offspring with the highest fitness value, o_{best}.
   (d) If o_{best} is better than the parent with the lowest fitness score, either p_1 or p_2, then replace that parent with o_{best}.
   (e) If $p_1 = p_2$, then exit iteration.
3. Return parent with highest fitness score, either p_1 or p_2.

Figure 3.4: Pseudocode for crossover-hill-climbing step.

Since the local search can add unnecessary computational expense, Lozano et al. (2004) only carry out the crossover-hill-climbing step with probability 1 if the best new offspring is better than the worst member of the population, or with probability 0.0625 otherwise. We opt to carry out the local search only if the best new offspring is better than the worst member of the population. Lozano et al. (2004) implement specific mating, global crossover, mutation and replacement strategies, which keeps the population diverse; however, we choose more targeted strategies. Only our local crossover operation follows their approach. Our

global crossover step and our mutation step are similar to the generic genetic algorithm, with the probability of mutation 0.0005.

In the low-dimensional case, we initiate $m_1 = 20$ models then for each of the $n_{it} = 20$ global iterations, we add $m_2 = 5$ new randomly generated models to add diversity to the population. Thus, the regular GLGA considers a total of 120 models, not including the offspring. We only do at most 3 local iteration because since we control the initialization of the matrices with the edgewise Bayes factor, the local searches converge quickly. We select the matrix with the highest edgewise values up until a cut-off, determined by the sample size, as one parent which mates in every iteration with a second parent which is the model with the highest fitness score out of the current population. We only keep the offspring from the global crossover step and the offspring from the local crossover step if they have higher fitness scores than the member of the population with the current highest fitness score. The pseudo code for the graphical local genetic algorithm can be found in Table 3.5. Note that in our simulation results we use $m_1 = 20$, $m_2 = 5$, $n_{it} = 20$, $n_t = 3$, and $x = 3$.

Graphical Local Genetic Algorithm

1. Initialize population of m_1 matrices.
2. Repeat n_{it} times.
   (a) Evaluate fitness of current population and select best two matrices from the current population as parent models.
   (b) Perform one-point global crossover x times and mutation to produce $2x$ offspring.
   (c) Evaluate fitness of the offspring and find best offspring, o_{best}.
   (d) If the top offspring is better than the worst individual in the general population, then
      i. Find the best parent, c_{best}, and perform Crossover-hill-climbing(c_{best}, o_{best}, $2x$, n_t).
      ii. Once the termination condition is met, if final best parent is better than the best individual in the general population, then replace the worst individual with result from Crossover-hill-climbing step.
   (e) Generate m_2 new random matrices to add diversity to the general population, then continue to next iteration.

Figure 3.5: Pseudocode for graphical local genetic algorithm.

Initiating the Matrices

The advantage of the graphical local genetic algorithm is its flexibility; however, it is sensitive to its initialization. Thus, we use the edgewise Bayes factor (2.22) and the sample size to guide the search. Once we compute the Bayes factor for each edge, we use the distribution of these edgewise values as a guideline to choose two cut-off points. We choose two cut-off points: the first to indicate that edges with a corresponding value greater or equal to this cut-off will be initialized with a high probability, and the second to indicate that edges with

a corresponding value lesser or equal to this cut-off will be initialized with a low probability. The value of these cut-off points will depend on the sample size.

We use the first and third quartile as guidelines for the cut-off values. Edges corresponding to Bayes factor values greater than the third quartile are included with probability 0.9, and edges corresponding to Bayes factor values between the first and the third quartile are included with probability 0.4. To be conservative, for sample sizes over 1,000, we round up the third quartile value and we round down the first quartile value to the nearest number in the set $\{0, 5, 10, 100, 500, 1,000, 5,000, 10,000\}$. Therefore, there are few edges considered with high probability and there are few edges that are not considered. In general, for sample sizes over 100,000 we take the cut-off points $\{100, 1,000\}$, between 5,000 and 100,000 we take $\{10, 100\}$, and under 1,000 we take $\{7, 10\}$. For example, if we have a sample size of 6,000, then we compute (2.22) for each edges. The edges with corresponding edgewise Bayes factor values greater or equal to 100 will be initialized with probability 0.9, the edges with values between 10 and 100 will be initialized with probability 0.4, and the edges with values less than 10 will not be considered.

These are general guidelines for fitting a predictive model. If it is desired to obtain a sparse model for interpretation, then the cut-off points can be made more conservative and the edge probabilities can be decreased. A histogram showing the distribution of the edgewise Bayes factor values can help make the decision.

High-Dimensional Setting

The regular local graphical genetic algorithm we just described works well for up to 20 variables. For larger number of variables, we randomly select overlapping subsets of 8 variables and perform the algorithm as usual. We store the resulting top submodels as an array of adjacency matrices, then we select the union of all the edges to reconstruct the full model. We choose subsets of 8 because the algorithm works well with the number of variables for a variety of edge densities. The number of subsets is chosen depending on the number of variables; however, it is preferable to fit too many submodels than too few. In the high-dimensional case, we do not need to initiate as many matrices and we do not need as many global iterations because the subsets will be relatively sparse. We initiate $m_1 = 10$ models then for each of the $n_{it} = 3$ global iterations, we add $m_2 = 5$ newly generated models. Thus, the high-dimensional GLGA considers a total of 25 models, not including the offspring.

Since we are computing the log of the normalizing constant for subsets of the graph, it can occur that false edges are retained in the model. To combat this we, for sample sizes of 5,000 or more, follow the general guidelines for cut-off points in Section 3.2.2, and for sample sizes under 5,000, we take the upper cut-off as 10 and the lower cutoff as the second lowest edgewise Bayes factor value. The edges corresponding to high Bayes factor values are included with probability 0.9; however, the edges with corresponding values between the two cut-offs are included with probability 0.1. We lower this second probability since each subset is less dense than the full graph. Furthermore, to reduce superfluous edges, after the final graph is constructed if there are 3-cycles such that two of the edges have correspond

to high edgewise Bayes factor values and the third edges corresponds to a lower value, we delete the third edge with probability 0.8. This reduces the number of false edges that have accumulated over the course of the algorithm. As stated earlier, it is preferable to fit too many submodels than too few. This adjustment to remove extra edges means that we do not need to worry about searching too many subsets. In Section 3.3.1, we use 600 subsets of 8 variables to fit two models with $q = 100$ variables with two different densities. We use the same set up to fit both models and obtain favourable results, which demonstrates that this adjustment corrects for taking *too many* subsets. If we want the resulting model to be sparse for the interpretation purposes, then we will adjust the initial settings instead of taking less subsets.

Scalability of Algorithm

Most of the computations are performed on arrays of matrices and can be done quickly. The bottleneck of the algorithm is computing the fitness of each model. Both the log of the normalizing constant and the penalty slow down the computation because we must iterate over all the cliques and separators. Since the fitness of each model can be computed separately, we can use parallel computing to decrease the computing time.

In R, the packages *foreach* and *doParallel* allow us to use parallel execution on seven cores of the computer, which significantly reduces the computing time. For example, to compute the log of the normalizing constant for 120 models with 8 vertices, the regular 'for' loop takes 4.2213 seconds and the 'foreach' loop takes 1.3114 seconds. For 120 models with

20 vertices, the regular 'for' loop takes 1.3405 minutes and the 'foreach' loop takes 31.5809 seconds. In the low-dimensional setting, meaning up to 20 variables, the time it takes to run the algorithm is still manageable with $q = 20$ taking 36.1979 seconds.

We also use parallel computing to compute the edgewise Bayes factor when generating the initial model population. For $q = 100$, there are 4950 edges we need to consider. It takes the regular 'for' loop 2.8219 minutes to compute the Bayes factor indicating the presence or absence of each edge, and the 'foreach' loops takes 1.4356 minutes.

3.3 Experiments

In Section 3.3.1, we give the results of experiments with simulated data sets for $q \in \{6, 8, 12, 20, 50, 100\}$ for various sample sizes from 100 to 500,000. We use Gibbs sampling to generate each data set and we burn the first 1,000 samples. Then in Section 3.3.2, we implement the GLGA on a real world data set with $q = 32$.

3.3.1 Simulated Data Sets

Here we demonstrate the capabilities of the GLGA using simulated data from known graphs for various q and various sample sizes and compare it to the Chordalysis approach by Petijean et al. (2013). In Petijean et al. (2013), they illustrate the advantages of the Chordalysis approach; however, the method can only return a decomposable and it looses accuracy for smaller sample sizes.

To evaluate the performance of each algorithm, we use the

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \text{ specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \text{ and } F_1\text{-score} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}},$$

where TP, TN, FP and FN are the number of true positives, true negative, false positives and false negative, respectively. Each score is between 0 and 1, and a higher score implies better accuracy. The sensitivity is the proportion of true edges correctly identified, and the specificity is the proportion of absent edges correctly identified. We aim to have both sensitivity and specificity be 0.70 or above. The F_1-score measures the balance between precision= TP/(TP + FP) and recall= TP/(TP + FN), thus is measures the model selection method's ability to both identify true edges and false edges. In general, an F_1-score over 0.9 is very strong, between 0.9-0.8 is strong, between 0.5-0.8 is okay, and below 0.5 is weak.

Low-Dimensional Data Sets

First we give our simulation results for known graphs with $q \in \{6, 8, 12, 20\}$ in Tables 3.1, 3.2, 3.3, and 3.4, respectively. For each graph, we give the average sensitivity, specificity, and F_1 scores, and standard deviations (Sd.) of 20 runs and compare the results to the Chordalysis algorithm. Chordalysis returns the same graph every run, thus we do not have any standard deviation to report. For each sample size, the first row is the GLGA results, and the second row is the Chordalysis results.

In general, the GLGA performs well for sample size 5,000 or more, and it outperforms Chordalysis for small sample sizes. Moreover, since Chrordalysis must return a decomposable

model, only the GLGA is able to select the true model.

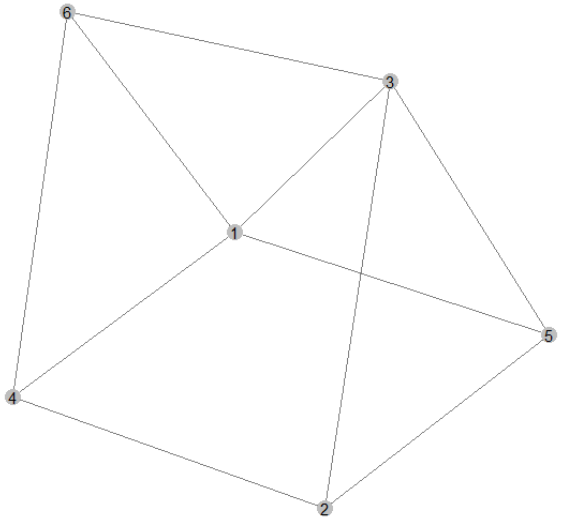


Figure 3.6: True non-decomposable graph with $q = 6$.

Table 3.1: Results from simulated data set with $q = 6$. The first row gives the results using GLGA and the second row is using Chordalysis.

Sample Size	Sensitivity	Sd.	Specificity	Sd.	F_1-Score	Sd.
100	0.63	0.0707	1.00	0.0000	0.77	0.0512
	0.20	-	1.00	-	0.33	-
500	0.50	0.0000	1.00	0.0000	0.67	0.0000
	0.50	-	1.00	-	0.67	-
1,000	0.60	0.0000	1.00	0.0000	0.75	0.0000
	0.60	-	1.00	-	0.75	-
5,000	0.78	0.0167	0.96	0.0179	0.87	0.0106
	0.60	-	1.00	-	0.75	-
10,000	0.80	0.0000	0.80	0.0000	0.84	0.0000
	0.80	-	0.80	-	0.84	-
50,000	0.80	0.0000	1.00	0.0000	0.89	0.0000
	0.90	-	0.60	-	0.86	-
100,000	0.96	0.0110	0.92	0.0219	0.96	0.0045
	1.00	-	0.60	-	0.91	-
500,000	1.00	0.0000	0.92	0.0219	0.98	0.0052
	1.00	-	0.20	-	0.83	-

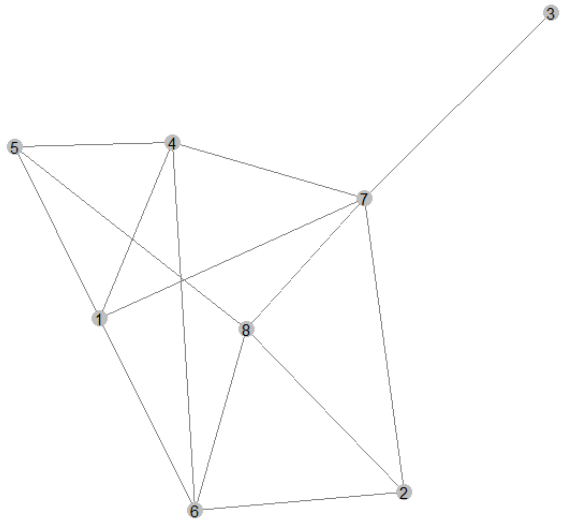


Figure 3.7: True non-decomposable graph with $q = 8$.

Table 3.2: Results from simulated data set with $q = 8$. The first row gives the results using GLGA and the second row is using Chordalysis.

Sample Size	Sensitivity	Sd.	Specificity	Sd.	F_1-Score	Sd.
100	0.50	0.0000	0.86	0.0000	0.61	0
	0.07	-	1.00	-	0.13	-
500	0.65	0.0000	1.00	0.0000	0.78	0
	0.50	-	1.00	-	0.67	-
1,000	0.50	0.0000	1.00	0.0000	0.67	0
	0.50	-	1.00	-	0.67	-
5,000	0.71	0.0000	0.93	0.0175	0.80	0.0076
	0.64	-	1.00	-	0.78	-
10,000	0.77	0.0064	0.94	0.0156	0.84	0.0102
	0.79	-	0.86	-	0.81	-
50,000	0.83	0.0078	0.81	0.0239	0.82	0.0114
	1.00	-	0.57	-	0.81	-
100,000	0.83	0.0217	0.87	0.0120	0.84	0.0163
	0.86	-	0.71	-	0.80	-
500,000	0.97	0.0078	0.81	0.0128	0.90	0.0077
	1.00	-	0.50	-	0.80	-

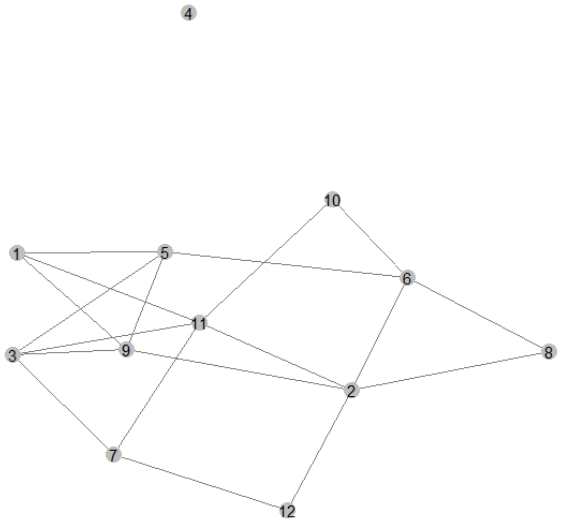


Figure 3.8: True non-decomposable graph with $q = 12$.

Table 3.3: Results from simulated data set with $q = 12$. The first row gives the results using GLGA and the second row is using Chordalysis.

Sample Size	Sensitivity	Sd.	Specificity	Sd.	F_1-Score	Sd.
100	0.57	0.0263	0.85	0.0000	0.58	0.0195
	0.26	-	0.98	-	0.40	-
500	0.77	0.0141	0.91	0.0049	0.77	0.0123
	0.53	-	1.00	-	0.69	-
1,000	0.72	0.0288	0.96	0.0000	0.79	0.0193
	0.58	-	0.98	-	0.71	-
5,000	0.78	0.0047	0.96	0.0000	0.83	0.0030
	0.68	-	0.98	-	0.79	-
10,000	0.67	0.0047	0.98	0.0000	0.78	0.0034
	0.63	-	0.98	-	0.75	-
50,000	0.89	0.0000	0.98	0.0000	0.92	0.0000
	0.68	-	0.98	-	0.79	-
100,000	0.75	0.0115	0.97	0.0055	0.83	0.0040
	0.79	-	0.94	-	0.81	-
500,000	0.91	0.0047	0.79	0.0019	0.75	0.0025
	0.89	-	0.85	-	0.79	-

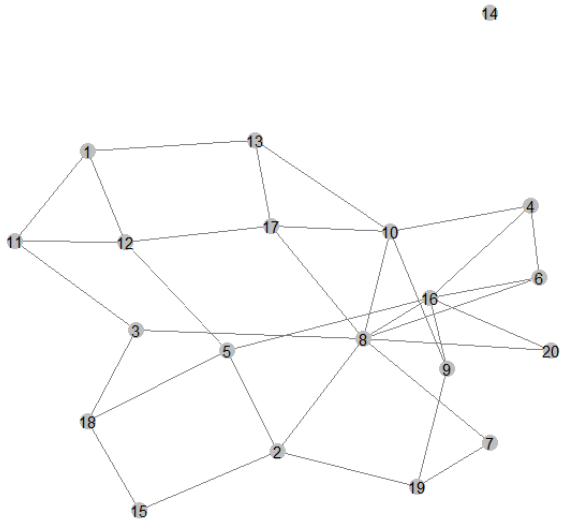


Figure 3.9: True non-decomposable graph with $q = 20$.

Table 3.4: Results from simulated data set with $q = 20$. The first row gives the results using GLGA and the second row is using Chordalysis.

Sample Size	Sensitivity	Sd.	Specificity	Sd.	F_1-Score	Sd.
100	0.65	0.0716	0.75	0.0670	0.48	0.0244
	0.24	-	1.00	-	0.38	-
500	0.61	0.0161	0.97	0.0000	0.70	0.0122
	0.41	-	1.00	-	0.58	-
1,000	0.75	0.0161	0.96	0.0035	0.77	0.0152
	0.56	-	1.00	-	0.72	-
5,000	0.82	0.0110	0.90	0.0021	0.72	0.0132
	0.59	-	1.00	-	0.74	-
10,000	0.70	0.0026	0.96	0.0006	0.75	0.0023
	0.53	-	1.00	-	0.69	-
50,000	0.82	0.0026	0.94	0.0006	0.79	0.0020
	0.76	-	0.98	-	0.83	-
100,000	0.73	0.0026	0.96	0.0007	0.76	0.0025
	0.71	-	0.98	-	0.79	-
500,000	0.91	0.0000	0.93	0.0006	0.82	0.0010
	0.74	-	0.92	-	0.69	-

High-Dimensional Data Sets

Here we start with results for the same graph with 20 variables as in Figure 3.9; however, we use the modified version of the GLGA with subsets. Note that we used different data sets, so the results using Chordalysis are different in each table. We use 50 subsets of 8 variables when using the high-dimensional method for $q = 20$. The results in Table 3.5 show that even though we are fitting subsets of the model, we can still obtain favourable results.

In Tables 3.6, 3.7, and 3.8, we have the results for $q = 50$, $q = 100$ with an edge density of 0.02, and $q = 100$ with an edge density of 0.05, respectively. For $q = 50$, we take 300 subsets of 8 and for $q = 100$, we take 600 subsets of 8. In Figures 3.11 and 3.12, we see that there is a noticeable difference in density between the two graphs. However, we use the same number of subsets for both graphs. This justifies the use of the adjustment we describe in Section 3.2.2. Again, it is better to take more subsets, than to take too few. The adjustment will control the number of false edges.

We see that the GLGA has sensitivity and specificity scores over 0.7, and F_1-score over 0.5 for sample size 5,000 or more. For the smaller sample sizes, it does not give as strong results; however, it is able to find the same or more edges than using Chordalysis. Note that the Chordalysis algorithm had an error when running on a data set with sample size 500,000 due to lack of memory, this is why its sensitivity score for the simulation is so low.

Table 3.5: Results from simulated data set with $q = 20$. The first row gives the results using GLGA and the second row is using Chordalysis.

Sample Size	Sensitivity	Sd.	Specificity	Sd.	F_1-Score	Sd.
100	0.84	0.0572	0.68	0.0262	0.50	0.0064
	0.09	-	1.00	-	0.16	-
500	0.80	0.0147	0.70	0.037	0.50	0.0190
	0.29	-	1.00	-	0.46	-
1,000	0.89	0.0449	0.77	0.0482	0.60	0.0337
	0.56	-	1.00	-	0.71	-
5,000	0.79	0.0417	0.83	0.0154	0.62	0.0173
	0.53	-	0.99	-	0.67	-
10,000	0.71	0.0322	0.94	0.0197	0.71	0.0390
	0.56	-	1.00	-	0.71	-
50,000	0.85	0.0000	0.91	0.0059	0.75	0.0091
	0.65	-	0.99	-	0.77	-
100,000	0.75	0.0573	0.95	0.0187	0.76	0.0332
	0.68	-	0.97	-	0.75	-
500,000	0.78	0.0332	0.91	0.0199	0.71	0.0325
	0.79	-	0.94	-	0.76	-

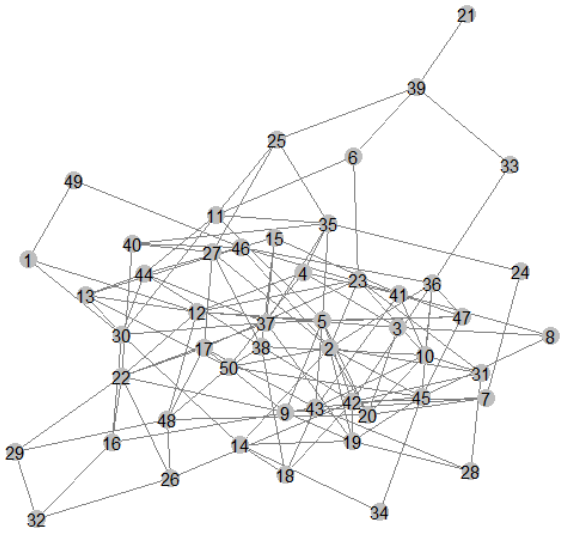


Figure 3.10: True non-decomposable graph with $q = 50$.

Table 3.6: Results from simulated data set with $q = 50$. The first row gives the results using GLGA and the second row is using Chordalysis.

Sample Size	Sensitivity	Sd.	Specificity	Sd.	F_1-Score	Sd.
100	0.61	0.0153	0.63	0.0035	0.26	0.0225
	0.05	-	1.00	-	0.10	-
500	0.60	0.0220	0.69	0.0031	0.23	0.0195
	0.31	-	1.00	-	0.48	-
1,000	0.76	0.0229	0.83	0.0032	0.48	0.0142
	0.37	-	1.00	-	0.54	-
5,000	0.73	0.0236	0.75	0.0019	0.57	0.0365
	0.42	-	0.99	-	0.57	-
10,000	0.61	0.0289	0.93	0.0050	0.55	0.01418
	0.48	-	0.99	-	0.62	-
50,000	0.81	0.0165	0.91	0.0071	0.63	0.0196
	0.47	-	0.98	-	0.58	-
100,000	0.62	0.0116	0.98	0.0036	0.69	0.0138
	0.52	-	0.98	-	0.60	-
500,000	0.77	0.0137	0.92	0.0104	0.62	0.0189
	0.56	-	0.94	-	0.54	-

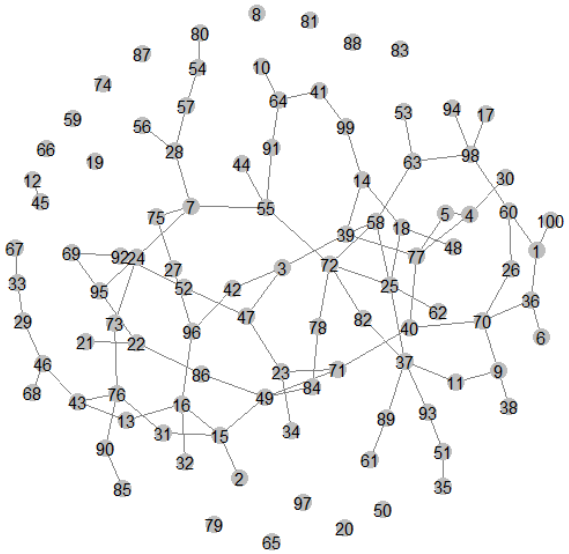


Figure 3.11: True non-decomposable graph with $q = 100$ and edge density 0.02.

Table 3.7: Results from simulated data set with $q = 100$ and edge density 0.02. The first row gives the results using GLGA with 600 subsets of 8 variables and the second row is using Chordalysis.

Sample Size	Sensitivity	Sd.	Specificity	Sd.	F_1-Score	Sd.
100	0.41	0.0306	0.80	0.0024	0.07	0.0059
	0.04	-	1.00	-	0.08	-
500	0.66	0.0222	0.92	0.0040	0.23	0.0134
	0.36	-	1.00	-	0.53	-
1,000	0.78	0.0243	0.90	0.0008	0.24	0.0077
	0.60	-	1.00	-	0.75	-
5,000	0.82	0.0231	0.98	0.0007	0.59	0.0153
	0.72	-	1.00	-	0.84	-
10,000	0.78	0.0523	0.98	0.0069	0.54	0.0153
	0.77	-	1.00	-	0.87	-
50,000	0.88	0.0206	0.98	0.0020	0.57	0.0202
	0.84	-	0.99	-	0.90	-
100,000	0.80	0.0200	0.99	0.0007	0.78	0.0169
	0.84	-	1.00	-	0.91	-
500,000	0.88	0.0230	0.99	0.0007	0.75	0.0174
	0.03	-	0.98	-	0.03	-

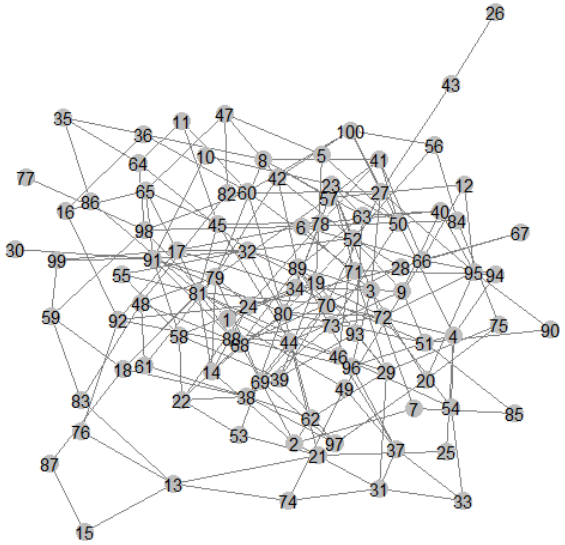


Figure 3.12: True non-decomposable graph with $q = 100$ and edge density 0.05.

Table 3.8: Results from simulated data set with $q = 100$ and edge density 0.05. The first row gives the results using GLGA with 600 subsets of 8 variables and the second row is using Chordalysis.

Sample Size	Sensitivity	Sd.	Specificity	Sd.	F_1-Score	Sd.
100	0.33	0.0050	0.88	0.0056	0.17	0.0063
	0.06	-	0.99	-	0.11	-
500	0.53	0.0291	0.94	0.0026	0.38	0.0163
	0.28	-	1.00	-	0.43	-
1,000	0.61	0.0116	0.94	0.0011	0.41	0.0071
	0.38	-	1.00	-	0.55	-
5,000	0.71	0.0470	0.95	0.0077	0.50	0.0146
	0.44	-	0.99	-	0.60	-
10,000	0.79	0.0231	0.94	0.0010	0.52	0.0144
	0.43	-	1.00	-	0.60	-
50,000	0.82	0.0126	0.94	0.0035	0.52	0.0155
	0.47	-	0.99	-	0.63	-
100,000	0.72	0.0124	0.99	0.0010	0.74	0.0163
	0.46	-	0.99	-	0.63	-
500,000	0.75	0.0163	0.97	0.0024	0.61	0.0161
	0.46	-	0.99	-	0.61	-

3.3.2 Application on Real Data Set

In this section, we apply the GLGA to a real-world data set. We apply our algorithm to the *Movies Dataset* collected by TMDB and GroupLens (`https://grouplens.org/datasets/movielens/latest/`). The original data set contains over 280,000 movie titles with reviews from over 50,000 individual viewers. The ratings are on a scale from 0 to 5 with intervals of 0.5. We select 32 movies which were reviewed by the same 353 individuals and if they rated a movie 4 or more then we encoded that observation as '1' to mean they like the movie, and if they rated a movie 3.5 or less we encoded that observation as '0' to mean they do not like the movie.

Since the same size a relatively small for this number of variables and the purpose of the model is for interpretation, we use conservative cut-off points and initial edge probabilities when generating the initial populations of submatrices. Edges with a Bayes factor value of 30 or over are initialized with probability 0.8, and edges with values of 15 or lower are initialized with probability 0.05. We used a histogram of the edgewise Bayes factors to decide these cut-offs. Moreover, since in our simulations we used 50 subsets for 20 variables, here we used 60 subsets. Figure 3.13 shows the graph representing selected model with the numeric labels given in the original data set. A legend for the titles, genre and year of the movies is provide in Appendix 4.2.

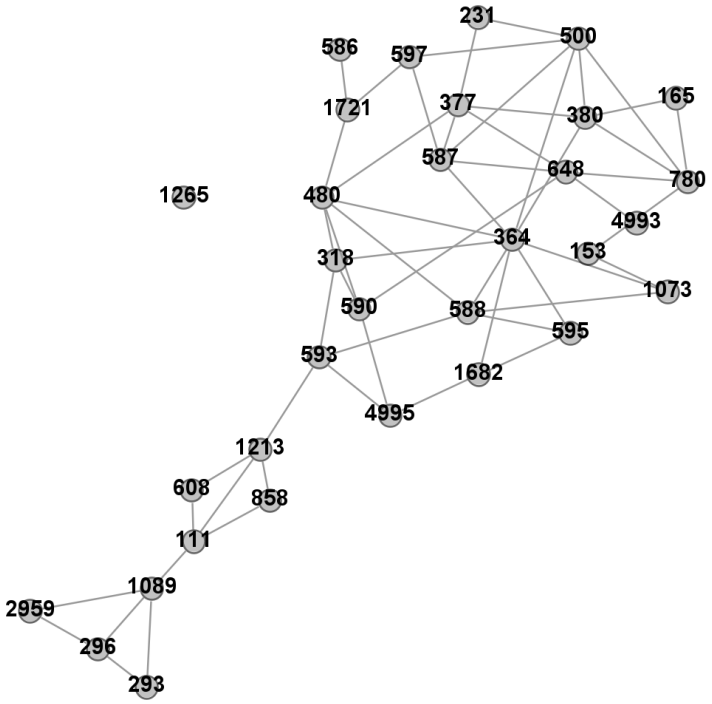


Figure 3.13: Graphical representation of model selected for the Movie Dataset.

The graph in Figure 3.13 show connections between movies which are liked by the same viewers. First, we notice that *Bridge to Terabithia* (1265) is the only movie not connected any other movie in the graph. This is because it is the only family movie we considered; therefore, we do not expect it to have been viewed by the same demographic as the other movies. The movie with the most connections is *Batman Returns* (364). This movie is apart of the well-known Batman franchise, so it was likely viewed by different demographics. It is considered an action movie and it is connected to movies with related genres drama, thriller, and horror; as well as, action. *Batman Returns* was directed by Tim Burton who is known for his quirky gothic fantasy and horror style. We notice that *Batman Returns* is connected to *Silent Hill* (588) and *Motha vs Godvilla* (1682) which are both categorized as horror. It is

also connected to *Big Fish* (587) which is a fantasy-drama, also directed by Tim Burton.

This type of model can be used to guide movie recommendations to users based on the movies they have previously viewed. The user would be recommended the movies connected to a movie that have viewed and if they like the original movie, they would select a recommended movie and if they did not like the original movie, they know not to select the new recommendations. For first time viewers, the recommendation system should start with a movie with many connections because those are likely to have been enjoyed by a diverse audience, for example, *Batman Returns*.

3.4 Conclusion

Graphical log-linear models are an effective tool for modelling complex interactions between discrete variables; however, model selection for high-dimensional data is a difficult task. In this chapter, we introduce the Graphical Local Genetic Algorithm, which is an extension of the graphical genetic algorithm to the high-dimensional setting with the crossover-hill-climbing operator from Lozano et al. (2004).

First, we successfully apply the GLGA to graphs up to 20 variables, then we modify the GLGA by implementing the algorithm for subsets of 8 variables and reconstructing the final model using the resulting subgraphs. We are able to fit data sets with up to 100 variables using the GLGA. Previously, the graphical genetic algorithm had only been implemented for graphs with a low number of variables. Many competing model selection methods are stepwise methods which rely on the properties of decomposable graphs. Our simulation results show

that the GLGA is flexible in that it can fit non-decomposable models with varying densities by taking advantage of the convenient properties of minimal triangulations. Moreover, we use the GLGA to analyse a real-world data set containing movie reviews for 32 movies from 353 individuals. The resulting model exhibits valuable connections between movies, which can be used for a movie recommendation system.

Chapter 4

Conclusion and Future Work

In this thesis, we studied the Bayes factor as a model selection criterion for high-dimensional discrete graphical models. Our main topic was proving that the Bayes factor is strong model selection consistent for non-decomposable models when the number of variable is increasing with the sample size. Our primary contribution is our approximation of log of the normalizing constant proportionate to the posterior probability. We derive a convenient expression approximate to log of the posterior probability that is comparable to the BIC. While the BIC is suitable for fixed dimension, our approximation is preferable for increasing dimension because it allows us to avoid high-dimensional integration and to control each individual error term. First, we examine the behaviour of the pairwise Bayes factor, then we establish the conditions for strong model selection consistency for both decomposable and non-decomposable graphs. Our theoretical results demonstrate that when the true graph is non-decomposable, it can be reasonably approximated by one of its minimal triangulations

and the difference between minimal triangulations with the same number of edges is negligible. We bolster these results with simulations using both a low number of variables and a high number of variables.

After establishing our theoretical results for the Bayes factor, we propose a model selection algorithm for high-dimensional discrete data as our second topic. We introduce the Graphical Local Genetic Algorithm, which is an extension of the graphical genetic algorithm to the high-dimensional setting with the crossover-hill-climbing operator from Lozano et al. (2004). In general, the genetic algorithm is a metaheuristic approach that aims to optimize some criterion and is not constrained to any stepwise procedure. We apply the GLGA to graphs up to 20 variables, then we modify the algorithm to take subsets of 8 variables and reconstructing the final model using the resulting subgraphs. Using this method, we are able to fit data sets with up to 100 variables. Our experiment results show that the GLGA is flexible in that it can fit non-decomposable models with varying densities, and it can be used to analyse real-world data sets.

In our research, we use the DY conjugate prior and the model prior defined in Section 2.3.2, where the edge probability is inversely proportional to the number of variables. Future research could include investigating the behaviour of the Bayes factor under other prior distributions, and there is a need to study the Bayes factor under more flexible settings, that is, when q_N increasing at a rate similar to N or faster than N. Moreover, the GLGA is adaptable and it would be interesting to see it applied to other types of graphical models, such as continuous graphical models and directed graphical models. The GLGA can easily be

modified to use other model selections criterion when computing the fitness of the candidate models.

Bibliography

Allen, G.I. and Liu, Z. (2012). *A Log-Linear Graphical Model for Inferring Genetic Networks from High-Throughput Sequencing Data.* In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, USA. 1-6.

Batir, N. (2008). Inequalities for the Gamma Function. *Archiv der Mathematik.* **91**. 554-563.

Blauth, A. and Pigeot, I. (2002). Using Genetic Algorithms for Model Selection in Graphical Models. *Collaborative Research Center 386.* Discussion Paper 278.

Cao, X., Khare, K., and Ghosh, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *The Annals of Statistics.* **47**. 319–348.

Cao, X., Khare, K., and Ghosh, M. (2020). Consistent Bayesian Sparsity Selection for High-dimensional Gaussian DAG Models with Multiplicative and Beta-mixture. *Journal of Multivariate Analysis.* **179**. 104-628.

Dahinden, C., Kalisch, M., and Bühlmann, P. (2010). Decomposition and Model Selection for Large Contingency Tables. *Biometrical Journal*. **25**. 233-252.

Darroch, J. N. and Speed, T. P. (1983). Additive and multiplicative models and interactions. *The Annals of Statistics*. **11**. 724–738.

Dawid, A. P., and Lauritzen, S. L. (1993). Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics*. **21**. 1272-1317.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate Priors for Exponential Families. *The Annals of Statistics*. **7**. 269-281.

Dobra, A. and Mohammadi, A. (2018). Loglinear Model Selection and Human Mobility. *The Annals of Applied Statistics*. **12**. 815-845.

Fitch, A. M., Jones, M. B., and Massam, H. (2014). The performance of covariance selection methods that consider decomposable models only. *Bayesian Analysis*. **9**. 659–684.

Gao, X. and, Carroll, R. J. (2017). Data Integration with High Dimensionality. *Biometrika*. **104**. 251-272.

Gao, X., Pu, D. Q., Wu, Y., and Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statistica Sinica*. **22**. 1123-1146.

García-Martínez, C., Lozano M., and Molina, D. (2006). *A Local Genetic Algorithm for Binary-Coded Problems.* In Parallel Problem Solving from Nature. Runarsson, T.P., Beyer, H.-G., Burke, E., Merelo-Guervós, J. J., Whitley, L.D., Yao, X.. Springer-Verlag: Berlin, Heidelberg. 192-201.

Gauraha, N. (2016). Model Selection for Graphical Log-Linear Models: A Forward Model Selection Algorithm based on Mutual Conditional Independence. *Methodology.* n. pag.

Gauraha, N. and Parui, S. K. (2020). Mutual Conditional Independence and its Applications to Model Selection in Markov Networks. *Annals of Mathematics and Artificial Intelligence.* **88**. 951-972.

Heggernes, P. (2006). Minimal triangulations of graphs: A survey. *Discrete Mathematics.* **306**. 297-317.

Givens G. H. and Hoeting J. A. (2013). *Genetic Algorithms.* In Computational Statistics. Giudici, P., Givens, G.H., Mallick, B.K.. John Wiley & Sons Inc. Publication: Hoboken, New Jersey. 75-84.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems.* University of Michigan Press: Ann Arbor, Michican.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford Science Publications.

Lee, K., Lee, J., Lin, L. (2019). Minimax posterior convergence rates and model

selection consistency in high-dimensional DAG models based on sparse Cholesky factors. The Annals of Statistics. **47**. 3413–3437.

Letac, G. and Massam, H. (2012). Bayes Factors and the Geometry of Discrete Hierarchical Loglinear Models. *The Annals of Statistics*. **40**. 1-30.

Lozano, M., Herrera F., Krasnogor, N. and Molina, D. (2004). Real-Coded Memetic Algorithms with Crossover Hill-Climbing. *Evolutionary Computation*. **12**. 273-302.

Massam, H., Liu J., and Dobra A. (2009). A Conjugate Prior for Discrete Hierarchical Log-Linear Models. *The Annals of Statistics*. **37**. 3431-3467.

Meinshausen, N., Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*. **34**. 1436–1462.

Niu, Y., Debdeep P. and Mallick, B. K. (2021). Bayesian Graph Selection Consistency Under Model Misspecification. *Bernoulli*. **27**. 637-672.

Pauwels, E. (2020). Sub Gaussian random variables. *Lecture notes: Statistics optimization and algorithms in high dimension*. Université Toulouse III - Paul Sabatier, delivered February 6 2020.

Petijean, F., Webb G. I., and Nicholson, A. E. (2013). *Scaling Log-Linear Analysis to High-Dimensional Data.* In Proceedings of the IEEE 13th International Conference on Data Mining. 597-606.

Poli, I. and Roverato A. A Genetic Algorithm for Model Selection. *Journal of the Italian Statistical Society*. **7**. 197-208.

Raskutti, G., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2009). Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of ℓ_1-regularized MLE. *Advances in Neural Information Processing Systems*. 1329–1336.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics* **38**. 1287–1319.

Rigollet, P. (2015). *Sub-Gaussian random variables.* Lecture notes: High dimensional statistics. Massachusetts Institute of Technology, delivered Spring 2015.

Rose, D. J., Tarjan, R. E. and Lueker, G. S. (1976). Algorithmic aspects of vertex elimination on graphs. *SIAM Journal on computing*. **5**. 266-283.

Spokoiny, V. and Zhilova, M. (2013). Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics*. **22**. 100-113.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*. **58**. 267–288.

Wainwright, M. J., Ravikumar, P., and Lafferty, J. D. (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. *Advances in*

Neural Information Processing Systems (B. Schölkopf, J. Platt and T. Hoffman, eds.) **19**. 1465–1472.

YUAN, M. AND LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*. **94**. 19–35.

Appendix A

In Sections 2.3.3, 2.4 and 2.5, we use a combination of the weak law of large numbers, the continuous mapping theorem and the triangle inequality to acquire the necessary large deviation results. We prove these lemmas for one component of a vector of probabilities, meaning in the following lemmas, π_0, $\widehat{\pi}$, and $\widehat{\pi}^*$ correspond to cell probabilities for a Bernoulli random variable. When we apply these lemmas, we use the Bonferroni correction to account for the differences between the estimates and the true probabilities for all of the marginal cell probabilities in a given model. In our main results, we use p_0, $\widehat{p}$, and $\widehat{p}^*$ to denote vectors of marginal probabilities.

Most of our theoretical results are in terms of the parametrization with respect to the cell probabilities. The exception is when we are comparing overfitting models, which requires the log of the likelihood ratio. In Appendix 4.2, we provide the quadratic form of the log of the likelihood ratio using the log-linear parametrization.

Here we provide the proofs for Lemmas 4.1.1-4.2.1, where Lemmas 4.1.3-4.1.7 are for a fixed number of variables, and Lemmas 4.1.8-4.1.12 are for a number of variables increasing with the sample size. More details are given in the proofs which cover increasing q_N than are

given in the proofs for fixed q. Lemma 4.2.1 shows the proof for $q_N \to \infty$.

4.1 Results regarding large deviation bounds

Lemma 4.1.1. *Let $Y_i \sim Bernoulli(1, \pi_0)$, $i = 1, ..., N$. By Assumption 1, π_0 is bounded away from 0 and 1. Define $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \widehat{\pi}$ and $E(Y_i) = \pi_0$. Then*

$$\widehat{\pi}\log\widehat{\pi} + (1-\widehat{\pi})\log(1-\widehat{\pi}) \xrightarrow{P} \pi_0\log\pi_0 + (1-\pi_0)\log(1-\pi_0),$$

where the term $\pi_0\log\pi_0 + (1-\pi_0)\log(1-\pi_0)$ is the expected log-likelihood of Y_i.

Proof of Lemma 4.1.1. Let $g(\pi_0) = \pi_0\log\pi_0 + (1-\pi_0)\log(1-\pi_0)$, where $g(\pi_0)$ is continuous on the interval $(0, 1)$. By weak law of large numbers $\widehat{\pi} \xrightarrow{\mathrm{P}} \pi_0$. Furthermore, by the continuous mapping theorem $g(\widehat{\pi}) \xrightarrow{\mathrm{P}} g(\pi_0)$. □

Lemma 4.1.2. *Let $Y_i \sim Multinomial(N, \tilde{\pi}_0)$, $i = 1, ..., N$, where $Y_i = (Y_1, \ldots, Y_q)^T$ and $\tilde{\pi}_0 = (\pi_{01}, \ldots, \pi_{0q})^T$ such that $\sum_{j=1}^{q} y_j = N$ and $\sum_{j=1}^{q} \pi_{0j} = 1$. By Assumption 1, each π_{0j} is bounded away from 0 and 1. Define $\widehat{\pi}_j = \frac{y_j}{N}$ and $E(Y_i) = N\pi_{0i}$. Then*

$$\sum_{j=1}^{q} \widehat{\pi}_j \log\widehat{\pi}_j \xrightarrow{P} \sum_{j=1}^{q} \pi_{0j}\log\pi_{0j},$$

where the term $\sum_{j=1}^{q} \pi_{0j}\log\pi_{0j}$ is the expected log-likelihood of Y_i.

Proof of Lemma 4.1.2. Let $g(\pi_{0j}) = \pi_{0j}\log\pi_{0j}$, where $g(\pi_{0j})$ is continuous on the interval $(0, 1)$. By weak law of large numbers $\widehat{\pi}_j \xrightarrow{\mathrm{P}} \pi_{0j}$. Furthermore, the sum of a finite number of

continuous functions is itself a continuous function, so by the continuous mapping theorem, $\sum_{j=1}^{q} g(\widehat{\pi}_j) \xrightarrow{\mathrm{P}} \sum_{j=1}^{q} g(\pi_{0j})$. □

Lemma 4.1.3. *Under the same setting as Lemma 4.1.1,*

$$P\left(|\widehat{\pi} - \pi_0| > (2N^{-1}q \log N)^{1/2}\right) = O\left(N^{-q}\right),$$

where the dimension q is fixed.

Proof of Lemma 4.1.3. Under the same setting as Lemma 4.1.1, we have that $Y \sim Bernoulli(1, \pi_0)$. We let $Z = Y - \mathrm{E}(Y)$, then for N independent Bernoulli random variables $Y_1, \ldots, Y_N$ we can write $|\widehat{\pi} - \pi_0| = |N^{-1} \sum_{i=1}^{N} [Y_i - \mathrm{E}(Y_i)]|$. Following the proof of Lemma 4.1.7, we know that

$$\mathrm{P}\left(|\widehat{\pi} - \pi_0| > t\right) \leqslant 2e^{-Nt^2/2}.$$

Letting $t = (2N^{-1}q \log N)^{1/2}$,

$$\mathrm{P}\left(|\widehat{\pi} - \pi_0| > t\right) \leqslant 2\exp\{-N[(2N^{-1}q \log N)^{1/2}]^2/2\} = 2\exp\{-q \log N\} = O(N^{-q}).$$

□

Lemma 4.1.4. *Under the same setting as Lemma 4.1.1,*

$$P\left(|g(\widehat{\pi}) - g(\pi_0)| > (CN^{-1}q \log N)^{1/2}\right) = O\left(N^{-q}\right),$$

where the dimension q is fixed, $g(\cdot)$ is the continuous function defined in Lemma 4.1.1 and C is a positive universal constant.

Proof of Lemma 4.1.4. From Lemma 4.1.1, we have $g(\pi_0) = \pi_0 \log \pi_0 + (1-\pi_0)\log(1-\pi_0)$. Following the proof of Lemma 4.1.8, by the mean value theorem and Lemma 4.1.3 which gives

$$|g(\widehat{\pi}) - g(\pi_0)| = |g'(k)||\widehat{\pi} - \pi_0| < |g'(k)|(2N^{-1}q\log N)^{1/2} = (CN^{-1}q\log N)^{1/2},$$

where C is a positive universal constant. Therefore,

$$\mathrm{P}\left(|g(\widehat{\pi}) - g(\pi_0)| > (CN^{-1}q\log N)^{1/2}\right) = O(N^{-q}),$$

for fixed dimension q. □

Lemma 4.1.5. *Let $\widehat{\pi}^*$ be the empirical frequency of the sum of true cell counts and fictive cell counts for a fixed marginal cell probability. Under Assumption 4, by Lemma 4.1.3,*

$$P\left(|\widehat{\pi}^* - \pi_0| > (18N^{-1}q\log N)^{1/2}\right) = O\left(N^{-q}\right),$$

where the dimension q is fixed.

Proof of Lemma 4.1.5. Under the conditions as Lemma 4.1.10, we define $\widehat{\pi} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and

$\widehat{\pi}^* = \frac{1}{N+\alpha}(\sum_{i=1}^{N} Y_i + \tilde{s})$. Then following the proof of Lemma 4.1.10, under Assumption 4, by the triangle inequality and Lemma 4.1.3 where $\epsilon_N = (2N^{-1}q \log N)^{1/2}$, we have

$$|\widehat{\pi}^* - \pi_0| \leqslant |\widehat{\pi}^* - \widehat{\pi}| + |\widehat{\pi} - \pi_0| < \frac{2\alpha}{N} + \epsilon_N < 2\epsilon_N + \epsilon_N = 3\epsilon_N.$$

Therefore,

$$\mathrm{P}\left(|\widehat{\pi}^* - \pi_0| > (18N^{-1}q \log N)^{1/2}\right) = O\left(N^{-q}\right),$$

for fixed q.

□

Lemma 4.1.6. *Let $\widehat{\pi}^*$ be the empirical frequency of the sum of true cell counts and fictive cell counts for a fixed marginal cell probability. Then by Lemma 4.1.4 and Lemma 4.1.5,*

$$P\left(|g(\widehat{\pi}^*) - g(\pi_0)| > (18CN^{-1}q \log N)^{1/2}\right) = O\left(N^{-q}\right),$$

where the dimension q is fixed, $g(\cdot)$ is the continuous function defined in Lemma 4.1.1 and C is a positive universal constant.

Proof of Lemma 4.1.6. In Lemma 4.1.1, we define $g(\pi_0) = \pi_0 \log \pi_0 + (1 - \pi_0)\log(1 - \pi_0)$ and in Lemma 4.1.5, we define $\widehat{\pi} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and $\widehat{\pi}^* = \frac{1}{N+\alpha}(\sum_{i=1}^{N} Y_i + \tilde{s})$.

Following the proof of Lemma 4.1.11, by the triangle inequality and Lemma 4.1.4 where

$\epsilon_N = (CN^{-1}q \log N)^{1/2}$, we have

$$|g(\widehat{\pi}^*) - g(\pi_0)| < |g'(k)|2\epsilon_N + \epsilon_N = C\epsilon_N,$$

where C is a positive universal constant. Therefore,

$$\mathrm{P}\left(|g(\widehat{\pi}^*) - g(\pi_0)| > (18CN^{-1}q \log N)^{1/2}\right) = O\left(N^{-q}\right),$$

for fixed dimension q.

□

Lemma 4.1.7. *Let $\theta_{[t]}$ denote the t^{th} component of the log-linear parameter vector $\theta = (\theta_j, j \in J)$, where θ_j has the representation (2.5). Let $\widehat{\theta}$ denote the MLE and let θ_0 denote the true value of the log-linear parameter vector. Under the same setting as Lemma 4.1.1,*

$$P\left(|(\widehat{\theta} - \theta_0)_{[t]}| > (CN^{-1}q \log N)^{1/2}\right) = O\left(N^{-q}\right),$$

where the dimension q is fixed.

Proof of Lemma 4.1.7. Following the proof of Lemma 4.1.12, we know that any $\theta_{[t]}$ is a linear combination of the log of the cell probabilities.

Let $\widehat{\pi}$ and π_0 denote the MLE and the true value of any cell probability, respectively. Let $h(\cdot)$ be any continuous function. Then by the mean value theorem and by Lemma 4.1.3

where $\epsilon_N = (2N^{-1}q \log N)^{1/2}$, we have

$$|h(\widehat{\pi}) - h(\pi_0)| < (CN^{-1}q \log N)^{1/2},$$

where C is a positive universal constant.

It is known that the sum of a finite number of continuous functions is also a continuous function. Since $\log(x)$ is a continuous function for $x > 0$ and the dimension q is finite, then we can apply the continuous mapping theorem. Thus, by Lemma 4.1.3 and by the continuous mapping theorem, we have

$$\mathrm{P}\left(|(\widehat{\theta} - \theta_0)_{[t]}| > (CN^{-1}q \log N)^{1/2}\right) = O\left(N^{-q}\right),$$

for fixed dimension q.

□

Lemma 4.1.8. *Under the same setting as Lemma 4.1.1,*

$$P\left(|\widehat{\pi} - \pi_0| > (2N^{-1}Q_N \log Q_N)^{1/2}\right) = O\left({Q_N}^{-Q_N}\right),$$

where $Q_N = {q_N}^2$ and the dimension q_N increases as $N \to \infty$.

Proof of Lemma 4.1.8. From the conditions stated in Lemma 4.1.1, we have that $Y \sim$ *Bernoulli*$(1, \pi_0)$ with $Y \in [0, 1]$ and $\mathrm{E}(Y) = \pi_0$. Let $Z = Y - \mathrm{E}(Y)$, then $\mathrm{E}(Z) = 0$, where $Z \in (-1, 1)$ since

$$0 \leqslant Y \leqslant 1 \implies -\mathrm{E}(Y) \leqslant Z \leqslant 1 - \mathrm{E}(Y) \implies -1 < -\pi_0 \leqslant Z \leqslant 1 - \pi_0 < 1,$$

and by assumption π_0 is bounded away from 0 and 1. By Hoeffding's lemma, Lemma 1.8 in Rigollet (2003),

$$\mathrm{E}\left(e^{tZ}\right) \leqslant e^{t^2/2},$$

for $t \in \mathbb{R}$, thus $Z \sim subGaussian(1)$. Since $E(Z) = 0$, by Hoeffding's inequality, Theorem 1.9 in Rigollet (2003), for N independent Bernoulli random variables $Y_1, \ldots, Y_n$ we have

$$\mathrm{P}\left(\frac{1}{N}\sum_{i=1}^{N}[Z_i - \mathrm{E}(Z_i)] > t\right) = \mathrm{P}\left(\frac{1}{N}\sum_{i=1}^{N}[Y_i - \mathrm{E}(Y_i)] > t\right) \leqslant e^{-2N^2t^2/4N} = e^{-Nt^2/2}.$$

Also from Lemma 4.1.1, we can write $|\widehat{\pi} - \pi_0| = |N^{-1}\sum_{i=1}^{N}[Y_i - \mathrm{E}(Y_i)]|$. Therefore,

$$\mathrm{P}\left(|\widehat{\pi} - \pi_0| > t\right) \leqslant 2e^{-Nt^2/2}.$$

Letting $t = (2Q_N \log Q_N)^{1/2}$ with $Q_N = {q_N}^2$, then

$$\mathrm{P}\left(|\widehat{\pi}-\pi_0|>t\right) \leqslant 2\exp\{-N[(2N^{-1}Q_N\log Q_N)^{1/2}]^2/2\} = 2\exp\{-Q_N\log Q_N\} = O(Q_N^{-Q_N}).$$

□

Lemma 4.1.9. *Under the same setting as Lemma 4.1.1,*

$$P\left(|g(\widehat{\pi})-g(\pi_0)|>(CN^{-1}Q_N\log Q_N)^{1/2}\right) = O\left(Q_N^{-Q_N}\right),$$

where $Q_N = {q_N}^2$, the dimension q_N is increasing as $N\to\infty$, $g(\cdot)$ is the continuous function defined in Lemma 4.1.1 and C is a positive universal constant.

Proof of Lemma 4.1.9. From Lemma 4.1.1, we have $g(\pi_0) = \pi_0\log\pi_0 + (1-\pi_0)\log(1-\pi_0)$. By the mean value theorem, we have

$$g'(k) = \frac{g(\widehat{\pi})-g(\pi_0)}{\widehat{\pi}-\pi_0},$$

for some k between π_0 and $\widehat{\pi}$. By Assumption 1, π_0 is bounded away from 0 and 1; therefore, the neighbourhood of k is also bounded away from 0 and 1. Then we can assume $|g'(k)|$ is bounded by a positive constant. Then if we apply the result from Lemma 4.1.8, where $\epsilon_N = (2N^{-1}Q_N\log Q_N)^{1/2}$, we have

$$|g(\widehat{\pi}) - g(\pi_0)| = |g'(k)||\widehat{\pi} - \pi_0| < |g'(k)|(2N^{-1}Q_N \log Q_N)^{1/2} = (CN^{-1}Q_N \log Q_N)^{1/2},$$

where C is a positive universal constant. Therefore,

$$\mathrm{P}\left(|g(\widehat{\pi}) - g(\pi_0)| > (CN^{-1}Q_N \log Q_N)^{1/2}\right) = O(Q_N^{-Q_N}),$$

where $Q_N = {q_N}^2$ and the dimension q_N is increasing as $N \to \infty$.

□

Lemma 4.1.10. *Let $\widehat{\pi}^*$ be the empirical frequency of the sum of true cell counts and fictive cell counts for a fixed marginal cell probability. Under Assumption 4, by Lemma 4.1.8,*

$$P\left(|\widehat{\pi}^* - \pi_0| > (18N^{-1}Q_N \log Q_N)^{1/2}\right) = O\left(Q_N^{-Q_N}\right),$$

where $Q_N = {q_N}^2$ and the dimension q_N is increasing as $N \to \infty$.

Proof of Lemma 4.1.10. Under the same setting as Lemma 4.1.1, we define $\widehat{\pi} = \frac{1}{N}\sum_{i=1}^{N} Y_i$, where $\sum_{i=1}^{N} Y_i$ is the true count for a fixed marginal cell with $\sum_{i=1}^{N} Y_i < N$. Then let $\widehat{\pi}^* = \frac{1}{N+\alpha}(\sum_{i=1}^{N} Y_i + \tilde{s})$, where $\tilde{s}$ is the marginal fictive count corresponding to the appropriate marginal cell and α is the total fictive counts with $\tilde{s} < \alpha$.

By the triangle inequality and Lemma 4.1.8, where $\epsilon_N = (2N^{-1}Q_N \log Q_N)^{1/2}$, we have

$$
\begin{aligned}
|\widehat{\pi}^* - \pi_0| &\leqslant |\widehat{\pi}^* - \widehat{\pi}| + |\widehat{\pi} - \pi_0| \\
&< \left| \frac{(\sum_{i=1}^{N} Y_i + \tilde{s})}{N + \alpha} - \frac{\sum_{i=1}^{N} Y_i}{N} \right| + \epsilon_N \\
&< \left| \frac{2N\alpha}{N(N+\alpha)} \right| + \epsilon_N \\
&< \frac{2\alpha}{N} + \epsilon_N.
\end{aligned}
$$

By Assumption 4, we have that $\frac{2\alpha}{N} < 2\epsilon_N$, then

$$
|\widehat{\pi}^* - \pi_0| < \frac{2\alpha}{N} + \epsilon_N < 2\epsilon_N + \epsilon_N = 3\epsilon_N.
$$

Therefore,

$$
\mathrm{P}\left(|\widehat{\pi}^* - \pi_0| > (18N^{-1}Q_N \log Q_N)^{1/2}\right) = O\left(Q_N^{-Q_N}\right),
$$

where $Q_N = {q_N}^2$ and the dimension q_N is increasing as $N \to \infty$.

□

Lemma 4.1.11. *Let $\widehat{\pi}^*$ be the empirical frequency of the sum of true cell counts and fictive cell counts for a fixed marginal cell probability. Then by Lemma 4.1.9 and Lemma 4.1.10,*

$$
P\left(|g(\widehat{\pi}^*) - g(\pi_0)| > (18CN^{-1}Q_N \log Q_N)^{1/2}\right) = O\left(Q_N^{-Q_N}\right),
$$

where $Q_N = {q_N}^2$, the dimension q_N is increasing as $N \to \infty$, $g(\cdot)$ is the continuous function defined in Lemma 4.1.1 and C is a positive universal constant.

Proof of Lemma 4.1.11. In Lemma 4.1.1, we define $g(\pi_0) = \pi_0 \log \pi_0 + (1-\pi_0)\log(1-\pi_0)$, and in Lemma 4.1.10, we define $\widehat{\pi} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and $\widehat{\pi}^* = \frac{1}{N+\alpha}(\sum_{i=1}^{N} Y_i + \tilde{s})$.

By the triangle inequality and Lemma 4.1.9, where $\epsilon_N = (CN^{-1}Q_N \log Q_N)^{1/2}$, we have

$$|g(\widehat{\pi}^*) - g(\pi_0)| \leqslant |g(\widehat{\pi}^*) - g(\widehat{\pi})| + |g(\widehat{\pi}) - g(\pi_0)| < |g(\widehat{\pi}^*) - g(\widehat{\pi})| + \epsilon_N,$$

and by the mean value theorem,

$$g(\widehat{\pi}^*) = g(\widehat{\pi}) + g'(k)(\widehat{\pi}^* - \widehat{\pi}),$$

for some k between $\widehat{\pi}$ and $\widehat{\pi}^*$. By Assumption 1, π_0 is bounded away from 0 and 1; therefore, the neighbourhood of k is also bounded away from 0 and 1. Then we can assume $|g'(k)|$ is bounded by a positive constant. By the mean value theorem and Lemma 4.1.10, we have

$$|g(\widehat{\pi}^*) - g(\widehat{\pi})| = |g'(k)||\widehat{\pi}^* - \widehat{\pi}| < |g'(k)|\frac{2\alpha}{N} < |g'(k)|2\epsilon_N.$$

Then we can write

$$|g(\widehat{\pi}^*) - g(\pi_0)| < |g'(k)|2\epsilon_N + \epsilon_N = C\epsilon_N,$$

where C is a positive universal constant. Therefore,

$$\mathrm{P}\left(|g(\widehat{\pi}^*) - g(\pi_0)| > (18CN^{-1}Q_N \log Q_N)^{1/2}\right) = O\left(Q_N^{-Q_N}\right),$$

where $Q_N = {q_N}^2$ and the dimension q_N is increasing as $N \to \infty$.

□

Lemma 4.1.12. *Let $\theta_{[t]}$ denote the t^{th} component of the log-linear parameter vector $\theta = (\theta_j, j \in J)$, where θ_j has the representation (2.5). Let $\widehat{\theta}$ denote the MLE and let θ_0 denote the true value of the log-linear parameter vector. By Assumption 3 and under the same setting as Lemma 4.1.1,*

$$P\left(|(\widehat{\theta} - \theta_0)_{[t]}| > (2^{10}CN^{-1}Q_N \log Q_N)^{1/2}\right) = O\left(Q_N^{-Q_N}\right),$$

where $Q_N = {q_N}^2$ and the dimension q_N is increasing as $N \to \infty$.

Proof of Lemma 4.1.12. From (2.5) we see that if $\tilde{\pi}$ denotes the vector of all cell probabilities and A denotes a matrix with components -1, 0, or 1, then

$$\theta = A \log \tilde{\pi}.$$

Therefore, any t^{th} component of the log-linear parameter vector θ; that is, $\theta_{[t]}$ is a linear combination of log of the cell probabilities.

Let $\widehat{\pi}$ and π_0 denote the MLE and the true value of any cell probability, respectively. Let $h(\cdot)$ be any continuous function. Then the mean value theorem gives,

$$h(\widehat{\pi}) = h(\pi_0) + h'(k)(\widehat{\pi} - \pi_0),$$

for some k between π_0 and $\widehat{\pi}$. By Assumption 1, π_0 is bounded away from 0 and 1, thus there exists constants $0 < c_1 < c_2 < 1$ such that $\pi_0 \in (c_1, c_2)$. If $k \in [c_1, c_2]$, then $|h'(k)|$ is bounded. Otherwise, we can choose an ϵ, such that $0 < \epsilon << 1$ and $k \in (c_1 - \epsilon, c_2 + \epsilon)$. Then by Lemma 4.1.8, where $\epsilon_N = (2N^{-1}Q_N \log Q_N)^{1/2}$, we have

$$|k - \pi_0| < |\widehat{\pi} - \pi_0| < \epsilon_N$$

with probability $1 - O(Q_N^{-Q_N})$. Therefore, $|h'(k)|$ is bounded with probability $1 - O(Q_N^{-Q_N})$. By the mean value theorem,

$$|h(\widehat{\pi}) - h(\pi_0)| = |h'(k)||\widehat{\pi} - \pi_0| < |h'(k)|(2N^{-1}Q_N \log Q_N)^{1/2} = (CN^{-1}Q_N \log Q_N)^{1/2},$$

where C is a positive universal constant.

When we consider the log-linear parametrization (2.5), we see that for a particular $j \in J$,

this parametrization is defined by the summation over $j' \in J : j' \lhd j$. Therefore, the number of summations to express any $\theta_{[t]}$ is at most $2^{\max|C|}$ for $C \in \mathcal{C}$, since the number of elements in j is at most the cardinality of the clique with the highest order. Under Assumption 3, the highest order of any clique or separator is 5, thus for $q_N \to \infty$ as $N \to \infty$, the number of log of the cell probabilities that are summed is at most 2^5. Since $\log(x)$ is a continuous function for $x > 0$, then the sum of 2^5 continuous functions is also a continuous function. Thus, by Lemma 4.1.8 and by the continuous mapping theorem, we have

$$\mathrm{P}\left(|(\widehat{\theta} - \theta_0)_{[t]}| > (2^{10} C N^{-1} Q_N \log Q_N)^{1/2}\right) = O\left(Q_N^{-Q_N}\right),$$

where $Q_N = {q_N}^2$ and the dimension q_N is increasing as $N \to \infty$.

□

4.2 Quadratic form of log of the likelihood ratio for over-fitting models

Here we use the marginal probabilities, and first, second, and third derivatives of the log-likelihood function (2.8), that is, (2.9), (2.10), (2.11), (2.12), (2.13), (2.14), and (2.15) found in Section 2.2.2.

Lemma 4.2.1. *Let $U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$ denote the score vector and let $H(\theta) = E\left(-\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T}\right)$ denote the Fisher Information under the true model, where the log-likelihood $\ell(\theta)$ has the form (2.8). Under Assumptions 2 and 4, for a model with k parameters and $Q_N = {q_N}^2$, as $N \to \infty$*

$$2\{\ell(\widehat{\theta}) - \ell(\theta_0)\} = \frac{1}{N} U_N(\theta_0)^T \{H(\theta_0)\}^{-1} U_N(\theta_0)\{1 + o(1)\},$$

with probability $1 - 2kO(N^{-q})$ *when* q *is fixed, and with probability* $1 - 2k_N O(Q_N^{-Q_N})$ *when* q_N *is increasing with* N.

Proof of Lemma 4.2.1. Here we prove the lemma for $q_N \to \infty$ as $N \to \infty$. We use the notation and follow the proof of Lemma A2 in Gao and Carroll (2017).

Recall that k is the number of free parameters in a hierarchical log-linear model. For a fixed component r and $r, t, u \in \{1, \ldots, k\}$, let $\ell_r^{(1)} = \frac{\partial \ell}{\partial \theta_r}$, $\ell_{rt}^{(2)} = \frac{\partial^2 \ell}{\partial \theta_r \partial \theta_t}$, and $\ell_{rtu}^{(3)} = \frac{\partial^3 \ell}{\partial \theta_r \partial \theta_t \partial \theta_u}$, where each derivative has the form (2.13), (2.14), and (2.15), respectively. Let $H_{rt}(\theta_0) = N^{-1}\mathrm{E}\left(-\frac{\partial^2 \ell(\theta_0)}{\partial \theta_r \partial \theta_t^T}\right)$ denote the $(r,t)^{th}$ entry of the Hessian matrix evaluated at the true parameter θ_0. The Taylor expansion for one component of the score vector at $\widehat{\theta}$ is

$$0 = \frac{1}{N}\ell_r^{(1)}(\widehat{\theta}) = \frac{1}{N}\ell_r^{(1)}(\theta_0) + \sum_t \frac{1}{N}\ell_{rt}^{(2)}(\theta_0)(\widehat{\theta} - \theta_0)_{[t]} + \sum_{tu} \frac{1}{2N}\ell_{rtu}^{(3)}(\tilde{\theta})(\widehat{\theta} - \theta_0)_{[t]}(\widehat{\theta} - \theta_0)_{[u]},$$

for some $\tilde{\theta}$ between $\widehat{\theta}$ and θ_0.

From (2.14) we know that $\frac{1}{N}\ell_{rt}^{(2)}(\theta_0) = P_r(\theta_0)P_t(\theta_0) - P_{rt}(\theta_0) = O(1)$ is bounded since it is a continuous function of marginal cell probabilities of the form (2.9) and (2.10). Furthermore, since there is no random variable in (2.14), we can write $\frac{1}{N}\ell_{rt}^{(2)}(\theta_0) = -H_{rt}(\theta_0)$. Hence,

$$\sum_t \frac{1}{N}\ell^{(2)}_{rt}(\theta_0)(\widehat{\theta}-\theta_0)_{[t]} = -\sum_t H_{rt}(\theta_0)(\widehat{\theta}-\theta_0)_{[t]}.$$

Similarly, from (2.15) we know that $\frac{1}{N}\ell^{(3)}_{rtu}(\theta_0) = P_{rt}(\theta_0)P_u(\theta_0) + P_{ru}(\theta_0)P_t(\theta_0)P_{tu}(\theta_0)P_r(\theta_0) - P_{rtu}(\theta_0) - P_r(\theta_0)P_u(\theta_0)P_t(\theta_0) - P_t(\theta_0)P_u(\theta_0)P_r(\theta_0) = O(1)$ is bounded since it is a continuous function of marginal cell probabilities of the form (2.9), (2.10) and (2.11). The function $\ell^{(3)}_{rtu}(\cdot)$ is by definition bounded because it is comprised of the marginal cell probabilities (2.9), (2.10) and (2.11), meaning it is made up of terms where the denominator is greater or equal to the numerator. Then we can say that $\left|\frac{1}{N}\ell^{(3)}_{rtu}(\tilde{\theta})\right|$ is bounded by a constant, where $\tilde{\theta}$ is in the neighbourhood of θ_0. Since $H_{rt}(\theta_0)$ is also bounded, we can say that $\left|\frac{1}{N}\ell^{(3)}_{rtu}(\tilde{\theta})/H_{rt}(\theta_0)\right| < C_1$ for some constant $C_1 > 0$. Moreover, there is no random variable in (2.15), so we can write $\frac{1}{N}\ell^{(3)}_{rtu}(\tilde{\theta}) = \mathrm{E}\left(\frac{1}{N}\ell^{(3)}_{rtu}(\tilde{\theta})\right)$. Then by Assumptions 2 and 3, and by Lemma 4.1.12,

$$\begin{aligned}
&-\sum_t H_{rt}(\theta_0)(\widehat{\theta}-\theta_0)_{[t]} + \sum_{tu}\frac{1}{2N}\mathrm{E}\left(\ell^{(3)}_{rtu}(\tilde{\theta})\right)(\widehat{\theta}-\theta_0)_{[t]}(\widehat{\theta}-\theta_0)_{[u]} \\
&= -\sum_t H_{rt}(\theta_0)(\widehat{\theta}-\theta_0)_{[t]}\left\{1 + \sum_u(\theta_0-\widehat{\theta})_{[u]}\left[\frac{\mathrm{E}\left(\ell^{(3)}_{rtu}(\tilde{\theta})\right)(\widehat{\theta}-\theta_0)_{[t]}}{2NH_{rt}(\theta_0)(\widehat{\theta}-\theta_0)_{[t]}}\right]\right\} \\
&\leqslant -\sum_t H_{rt}(\theta_0)(\widehat{\theta}-\theta_0)_{[t]}\left\{1 + \frac{C_1 k_N}{2}\left(\frac{2^{10}CQ_N\log Q_N}{N}\right)^{1/2}\right\} \\
&\leqslant -\sum_t H_{rt}(\theta_0)(\widehat{\theta}-\theta_0)_{[t]}\left\{1 + \frac{C_2}{2}\left(\frac{2^{10}C(q_N-1)^2Q_N\log Q_N}{N}\right)^{1/2}\right\} \\
&= -\sum_t H_{rt}(\theta_0)(\widehat{\theta}-\theta_0)_{[t]}\left\{1+o(1)\right\}
\end{aligned}$$

with probability $1 - q_N O(Q_N^{-Q_N})$, where $Q_N = {q_N}^2$ and $C_2 = 2^5 C_1$. Thus, we can write

$$0 = \frac{1}{N}\ell_r^{(1)}(\widehat{\theta}) = \frac{1}{N}\ell_r^{(1)}(\theta_0) - \sum_t H_{rt}(\theta_0)(\widehat{\theta} - \theta_0)_{[t]}\{1 + o_p(1)\}.$$

Then in matrix form, we obtain the result

$$\frac{1}{N}U_N(\theta_0) = H(\theta_0)(\widehat{\theta} - \theta_0)\{1 + o_p(1)\} \implies (\widehat{\theta} - \theta_0) = \frac{1}{N}\{H(\theta_0)\}^{-1}U_N(\theta_0)\{1 + o_p(1)\}. \tag{4.1}$$

Next, we consider the Taylor series expansion of the log-likelihood $\ell(\theta)$ at $\widehat{\theta}$

$$\begin{aligned}\ell(\widehat{\theta}) - \ell(\theta_0) = U_N(\theta_0)^T(\widehat{\theta} - \theta_0) + \sum_{rt}\frac{1}{2}(\widehat{\theta} - \theta_0)_{[r]}(\widehat{\theta} - \theta_0)_{[t]}\ell_{rt}^{(2)}(\theta_0) \\ + \sum_{rtu}\frac{1}{6}(\widehat{\theta} - \theta_0)_{[r]}(\widehat{\theta}_0 - \theta_0)_{[t]}(\widehat{\theta}_0 - \theta_0)_{[u]}\ell_{rtu}^{(3)}(\tilde{\theta}),\end{aligned}$$

for some $\tilde{\theta}$ between $\widehat{\theta}$ and θ_0. Since $\ell_{rt}^{(2)}(\theta_0) = -NH_{rt}(\theta_0)$ and $\ell_{rtu}^{(3)}(\tilde{\theta}) = \mathrm{E}\left(\ell_{rtu}^{(3)}(\tilde{\theta})\right)$, using the same reasoning as above, we have

$$
\begin{aligned}
&- \sum_{rt} \frac{1}{2} (\widehat{\theta} - \theta_0)_{[r]} (\widehat{\theta} - \theta_0)_{[t]} N H_{rt}(\theta_0) + \sum_{rtu} \frac{1}{6} (\widehat{\theta} - \theta_0)_{[r]} (\widehat{\theta} - \theta_0)_{[t]} (\widehat{\theta} - \theta_0)_{[u]} \mathrm{E}\left(\ell^{(3)}_{rtu}(\tilde{\theta}) \right) \\
&= - \sum_{rt} \frac{1}{2} (\widehat{\theta} - \theta_0)_{[r]} (\widehat{\theta} - \theta_0)_{[t]} N H_{rt}(\theta_0) \left\{ 1 + \frac{\sum_u \mathrm{E}\left(\ell^{(3)}_{rtu}(\tilde{\theta}) \right) (\theta_0 - \widehat{\theta})_{[u]}}{3 N H_{rt}(\theta_0)} \right\} \\
&\leqslant - \sum_{rt} \frac{1}{2} (\widehat{\theta} - \theta_0)_{[r]} (\widehat{\theta} - \theta_0)_{[t]} N H_{rt}(\theta_0) \left\{ 1 + \frac{C_2}{3} \left(\frac{2^{10} C (q_N - 1)^2 Q_N \log Q_N}{N} \right)^{1/2} \right\} \\
&= - \sum_{rt} \frac{1}{2} (\widehat{\theta} - \theta_0)_{[r]} (\widehat{\theta} - \theta_0)_{[t]} N H_{rt}(\theta_0) \left\{ 1 + o(1) \right\}
\end{aligned}
$$

with probability $1 - q_N O(Q_N^{-Q_N})$. From (4.1), we see that we obtain

$$
\begin{aligned}
\ell(\widehat{\theta}) - \ell(\theta_0) &= \frac{1}{N} U_N(\theta_0)^T \{H(\theta_0)\}^{-1} U_N(\theta_0) \{1 + o_p(1)\} \\
&\quad - \frac{1}{2N^2} \left\{ \{H(\theta_0)\}^{-1} U_N(\theta_0) \right\}^T N H(\theta_0) \{H(\theta_0)\}^{-1} U_N(\theta_0) \{1 + o_p(1)\} \\
&= \frac{1}{N} U_N(\theta_0)^T \{H(\theta_0)\}^{-1} U_N(\theta_0) \{1 + o_p(1)\} \\
&\quad - \frac{1}{2N} U_N(\theta_0)^T \{H(\theta_0)\}^{-1} U_N(\theta_0) \{1 + o_p(1)\} \\
&= \frac{1}{2N} U_N(\theta_0)^T \{H(\theta_0)\}^{-1} U_N(\theta_0) \{1 + o_p(1)\}.
\end{aligned}
$$

Therefore,

$$
2\{\ell(\widehat{\theta}) - \ell(\theta_0)\} = \frac{1}{N} U_N(\theta_0)^T \{H(\theta_0)\}^{-1} U_N(\theta_0) \{1 + o(1)\},
$$

with probability $1 - 2k_N O(Q_N^{-Q_N})$, where $k_N \leqslant 2^5(q_N - 1)$.

□

Appendix B

Table 4.1: Legend for the labels of the movies in Figure 3.13.

Label	**Movie Title**	**Genre**	**Year**
111	Scarface	Crime/Drama	1983
153	Lost in Translation	Romance/Drama	2003
165	Back to the Future Part II	Sci-fi/Comedy	1989
231	Syriana	Drama/Political Thriller	2005
293	A River Runs Through It	Drama	1992
296	Terminator 3: Rise of the Machines	Action/Sci-fi	2003
318	The Million Dollar hotel	Drama/Mystery	2000
364	Batman Returns	Action/Adventure	1992
377	Nightmare on Elm Street	Horror/Mystery	1984
380	Rain Man	Drama	1988
480	Monsoon Wedding	Comedy/Drama/Romance	2001
500	Resevoir Dogs	Action/Adventure	1992
586	Wag the Dog	Comedy/Political Cinema	1997
587	Big Fish	Drama/Fantasy	2003
588	Silent Hill	Supernatural/Horror	2006
590	The Hours	Drama/Romance	2002
593	Solaris	Sci-fi/Drama/Mystery	1972

Label	Movie Title	Genre	Year
597	Titanic	Romance/Drama	1997
608	Men in Black II	Sci-fi/Action	2002
648	Beauty and the Beast	Fantasy/Romance	1946
780	The Passion of Joan of Arc	Drama/Silent	1928
858	Sleepless in Seattle	Romance/Comedy	1993
1073	Arlington Road	Thriller/Crime	1999
1089	Point Break	Action/Crime	1991
1213	The Talented Mr. Ripley	Thriller/Drama	1999
1265	Bridge to Terabithia	Family/Fantasy	2007
1682	Mothra vs Godzilla	Sci-fi/Horror	1964
1721	All the Way Boys	Action/Comedy	1972
2959	License to Wed	Romance/Comedy	2007
4993	5 Card Stud	Western/Drama	1968
4995	Boogie Nights	Comedy/Drama	1997