MAXIMUM LIKELIHOOD ESTIMATE IN DISCRETE HIERARCHICAL LOG-LINEAR MODELS

NANWEI WANG

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

February 2017

Abstract

Hierarchical log-linear models are essential tools used for relationship identification between variables in complex high-dimensional problems. In this thesis we study two problems: the computation and the existence of the maximum likelihood estimate (henceforth abbreviated MLE) in high-dimensional hierarchical log-linear models.

When the number of variables is large, computing the MLE of the parameters is a difficult task to accomplish. A popular approach is to estimate the composite MLE rather than the MLE itself, that is, estimate the value of the parameter that maximizes the product of local conditional likelihoods. A more recent development is to choose the components of the composite likelihood to be local marginal likelihoods. We first show that the estimates obtained from local conditional and marginal likelihoods are identical. Second, we study the asymptotic properties of the composite MLE obtained by averaging the local estimates, under the double asymptotic regime, when both the dimension p and sample size N go to infinity. We compare the rate of convergence to the true parameter of the composite MLE with that of the global MLE under the same conditions. We also look at the asymptotic properties of the composite MLE when p is fixed and N goes to infinity and thus recover the same asymptotic results for p fixed as those of Liu and Ihler (2012).

The existence of the MLE in hierarchical log-linear models has important consequences for statistical inference: estimation, confidence intervals and testing as we shall see. Determining

whether this estimate exists is equivalent to finding whether the data belongs to the boundary of the marginal polytope of the model or not. Fienberg and Rinaldo (2012) gave a linear programming method that determines the smallest such face for relatively low-dimensional models. In this thesis, we consider higher-dimensional problems. We develop the methology to obtain an outer and inner approximation to the smallest face of the marginal polytope containing the data vector. Outer approximations are obtained by looking at submodels of the original hierarchical model, and inner approximations are obtained by working with larger models.

Acknowledgements

I would like to thank my parents for their nurturing support, developing my curiosity and giving me full freedom to explore the world.

I would also like to thank my PhD advisor, Professor Hélène Massam. Her invaluable advice will forever remind me that for great research to come to life, rigour, clarity and detail must follow spark and zest. I feel extraordinarily honoured to have had her there for me from the days of my first toddling attempts, all the way through to being able to demonstrate a matured research style.

To my thesis committee: I would like to thank Professor Xin Gao and Professor Huaxiong Huang, for their time, insightful comments and encouragement.

Table of Contents

List of Tables

List of Figures

1 Introduction

Hierarchical log-linear models are essential tools in the analysis of complex, high-dimensional categorical data of the types routinely encountered when analyzing multiple choice survey questions in social science or gene expression data in biology. Data points represent the values of the multivariate variable $X = (X_v, v \in V)$, where V is a finite set. Each variable X_v takes values in a finite set I_v. The N data points are classified according to the values of $X_v, v \in V$, in a $|V|$-dimensional array called a contingency table. There are $I = \prod_{v \in V} |I_v|$ cells $i = (i_v, v \in V)$ in this contingency table. The cell counts, that is, the total number of data points falling in cell $i, i \in I$ are denoted by $n(i)$, and the cell probabilities by $p(i)$. As we shall see in Section 2, the hierarchical log-linear model is defined by its generating set Δ, a subset of the power set of V, and the fact that $\log p(i)$ can be written as

$$\log p(i) = \theta_\emptyset + \sum_{D \in \Delta} \theta_D(i_D),$$

where $(\theta_D(i_D), D \in \Delta)$ are indicative of the relationship between variables $X_v, v \in D$. If, moreover, we assume that the cell counts $(n(i), i \in I)$ follow a multinomial distribution $\mathcal{M}(N, p(i), i \in I)$, then the density of cell counts, which is proportional to $\prod_{i \in I} p(i)^{n(i)}$, can be written under a natural exponential family form as

$$f(t; \theta)dt = \exp\{\langle \theta, t \rangle - Nk(\theta)\}\nu(dt), \tag{1.0.1}$$

where t is the sufficient statistic, $\langle \theta, t \rangle$ denotes the inner product of t and θ, and $\nu(dt)$ is a discrete

measure. The discrete graphical models class forms an important subclass of the hierarchical log-linear models class. Discrete graphical models are models for random variables $X = (X_v, v \in V)$ with distribution Markov with respect to an undirected graph G with vertices set V. In the case of discrete graphical models, the generating set is the set of complete induced subgraphs of G. More details will be given in Section 2.

Given a contingency table, we would like to explore the conditional independence relationships among the random variables, and to estimate the cell probabilities. The log-linear model is a generative model which learns the joint distribution $f(X_v, v \in V|\theta)$. In order to conduct some statistical inferences on $f(X_v, v \in V|\theta)$, we first take on the task of estimating the parameter θ. One of the most popular estimates of θ is the MLE. When p is large, however, evaluating the normalization constant $k(\theta)$ or even its approximation is $\mathbf{NP}-$hard, see Cooper (1990) and Roth (1996), and it is impossible to obtain the MLE of θ with a simple maximization of the likelihood function. Approximate techniques such as variational methods (see Jordan et al. (1999), Wainwright and Jordan (2008)) or MCMC techniques (see Geyer (1991)) have been developed in recent years. More recently still, work has been done on a third type of approximate techniques based on the maximization of composite likelihoods (see Besag (1975) and Lindsay (1988)). For a given data set $\{x^{(1)}, \ldots, x^{(N)}\}$ from a distribution with density $f(x|\theta)$, the likelihood function is $L(\theta) = \prod_{i=1}^{N} f(x^{(i)}|\theta)$. The composite likelihood is typically of the form $\prod_{i=1}^{N} \prod_{v \in V} f(x_v^{(i)}|x_{\mathcal{N}_v}^{(i)})$, where $\mathcal{N}_v$ is the set of neighbours of v in graph G. In other words, the composite likelihood is the product of the local conditional likelihoods.

In recent papers such as those of Ravikumar et al. (2010), Wiesel and Hero (2012), Liu and Ihler (2012), the estimate of θ is obtained from maximum likelihood estimates in the low dimensional local models, by combining the estimates to give a global estimate for θ. See Section 3 for more

details. In the case of statistical inference on Gaussian graphical models, Meng et al. (2014) consider local marginal models of $(X_v, X_{\mathcal{N}_v}), v \in V$, rather than the traditional local conditional models of X_v given $X_{\mathcal{N}_v}$.

In our work, we extended the estimates obtained from the local marginal likelihoods to discrete graphical models. Moreover, we show that the estimate obtained from the composite likelihood built on local marginal likelihoods is identical to the estimate obtained from the composite likelihood built on local conditional likelihoods. We therefore establish that one should use local conditional likelihoods instead of local marginal likelihoods, since the computational complexity of the former is much smaller than that of the latter.

MLE is a point estimation of the parameter θ, but to evaluate how good this estimate is, we need to study the asymptotic variance of the MLE. In this thesis, we extend the asymptotic analysis further, since we study the asymptotic properties of our estimate under both the classical and the double asymptotic regime, that is, when $|V| = p$ is fixed, and the number of data points N tends to infinity, but also when both p and N tend to infinity. The double asymptotic regime result is of greater interest in this big data era, as the dimension of a data set is no smaller than, or sometimes even larger than the number of data points.

The second main topic of this thesis is concerned with the existence of the MLE in the larger class of hierarchical log-linear models. The nonexistence of the MLE has problematic consequences for inference, clearly for estimation, but also for testing and model selection, see Fienberg and Rinaldo (2012). After we fit a statistical model on a dataset, it comes very naturally that we should test how well our model fits the data, or choose a better model from several candidates. In the literature, two popular summary statistics, the Pearson X^2 test and the likelihood ratio statistic G^2 test, are used for the goodness of fit test and model selection, see Bishop et al. (1975a) and Agresti

and Kateri (2011). If the MLE doesn't exist the standard regularity conditions for the asymptotic chi-square distribution no longer hold. Furthermore, as indicated in Geyer et al. (2009) the degrees of freedom used to approximate various measures of fit are incorrect in this case. The statistical implications of the nonexistence of the MLE on model selection in Bayesian inference are studied in further detail by Letac and Massam (2012).

Given a contingency table with some zero cell counts, the MLE of the canonical parameter θ doesn't exist, and therefore a finite estimate cannot be found to maximize the log-likelihood function. Example 3.3-1 in Bishop et al. (1975a) provides us with a 2^3 contingency table example to illustrate a nonexistent MLE situation. When some of the cells have a zero count, the MLE of some of the cell probabilities may not be positive. When the MLE doesn't exist, part of the natural parameters go to infinity, so the Fisher information matrix is singular. To resolve this, Geyer proposed a one-side confidence interval in Geyer et al. (2009).

Nowadays, hierarchical log-linear models are used for the analysis of large sparse contingency tables where many, if not most of the entries are small or zero counts. These zero counts often cause the MLE not exist. It is therefore most important to know whether the MLE exists before we analyze the data and goodness-of-fit of log-linear models. In Section 8, we show how one can deal with these problems by using an adequate parametrization in a reduced model. We illustrate this strategy on a real data example, see 10.1.2.

The remainder of this thesis is organized as follows. In Chapter 2, we give preliminary results that we shall use in our work. In Chapter 3, we offer a brief review of the literature on contingency tables, hierarchical log-linear models, and the existence of MLE. In Chapter 4, we study the composite maximum likelihood estimate and show that the composite likelihood built from local marginal models yields the same estimates as that built from local conditional models. In Chapter 5, we

start working on the asymptotic properties of the maximum composite likelihood estimate. Both the classical asymptotic regime result (Section 5.1) and the double asymptotic regime result (Section 5.2) are given. In Chapters 6 to 10, we develop our methodology to approximate the smallest facial set containing sufficient statistic t: $F_t = F_\Delta(I_+)$, and illustrate with several examples of simulated and real data.

2 Preliminaries

In this chapter, we list the basic notations we use in this paper and give some background knowledge. First, we briefly introduce our parameterization of hierarchical log-linear models and the corresponding likelihood function. Second, we define the face of the convex hull of sufficient statistics, and talk about some properties of convex polytopes.

2.1 Hierarchical log-linear models

Let V denote a finite index set. Let $X = (X_v, v \in V)$ be a vector of discrete random variables. We will assume that each variable takes values from a finite set I_v, and then X takes its values from

$$I = \prod_{v=1}^{V} I_v$$

let $|I_v|$ denote the cardinality of the set I_v, then $|I| = \prod_{v=1}^{V} |I_v|$. We write $i = (i_v, v \in V)$ for an element of I, where $x_v = i_v$.

Definition 2.1.1. *Given V, X and I defined as above and given a sample $\{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$ from X, we cross-classify the sample points according to the value of each of the variables $X_v, v \in V$. Each sample then falls into a cell $i \in I$. This set of cells is called a contingency table, the cell count $n(i)$ is the number of sample points falling into cell i, $n(i) = \sum_{j=1}^{N} 1_{\{X=i\}}(x^{(j)})$, and we use $N = \sum_{i \in I} n(i)$ for the total sample size. We denote $p(i)$ as the probability of each sample falling in cell i.*

For $E \subset V$, let $i_E = (i_v, v \in E)$ denote the cells in the E-marginal table with cell counts

$$n_E(i_E) = \sum_{k \in I_{V \setminus E}} n(i_E, k),$$

A family Δ of subsets of V is called a simplicial complex if $D \in \Delta,\ D' \subset D,\ D' \neq \emptyset$ implies $D' \in \Delta$. We assume $\cup_{D \in \Delta} D = V$. We denote by Ω_Δ the linear subspace of $x \in R^I$ such that there exist functions $\theta_D \in R^I$ for $D \in \Delta$ depending only on i_D and such that $x = \sum_{D \in \Delta} \theta_D$, that is

$$\Omega_\Delta = \{x \in R^I : \exists \theta_D, D \in \Delta \text{ such that } \theta_D(i) = \theta_D(i_D) \text{ and } x = \sum_{D \in \Delta} \theta_D\} \tag{2.1.1}$$

The hierarchical log-linear model generated by Δ is the set of positive cell probabilities $p = (p(i))_{i \in I}$ over a contingency table such that $\log p \in \Omega_\Delta$. The simplicial complex Δ is also called the generating class of the hierarchical log-linear model. For each cell probability we can write

$$\log p(i) = \theta_\emptyset + \sum_{D \in \Delta} \theta_D(i_D), \tag{2.1.2}$$

where $\theta_\emptyset$ doesn't depend on i and is a constant. The parameterization (2.1.2) is not unique as there are more parameters than the number of cells. In order to make it unique, we need to impose certain constraints on the parameters $\theta_D(i_D)$. We first select one of the values in I_v and denote it 0. The cell with all its components equal to 0 is the zero cell:

$$i = 0 = (0, 0, \cdots, 0).$$

The choice of 0 is arbitrary. Changing the level of X_v that will be called 0 simply leads to an affine transformation of the parameters. This allows us to impose the so called "baseline" or "corner"constraints

$$\theta_D(i_D) = 0, \quad i_v = 0, \text{for some } v \in D \tag{2.1.3}$$

Using (2.1.3), equation (2.1.2) becomes

$$\log p(i) = \theta_\emptyset + \sum_{D \in \Delta, i_v \neq 0, \forall v \in D} \theta_D(i_D), \tag{2.1.4}$$

Constraints (2.1.3) also imply that

$$\log p(0) = \theta_\emptyset, \tag{2.1.5}$$

so we can change the notation to θ_0.

Now we change to a more concise notation. First we define the support of a cell as follows:

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

and the subset J of I:

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

With the constraints (2.1.3) and the definition of set J above, in Proposition 2.1 of Letac and Massam (2012), it is shown that for $i \notin J,\ \theta_i = 0$ and

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

i.e. the elements in set J index the parameters in the hierarchical log-linear model generate by Δ, so we name it the *parameter set*. Again, to simplify the notation, for any two cells $i \in I, j \in J$, we define a new notation

$$j \lhd i$$

to mean that $S(j)$ is contained in $S(i)$ and $j_{S(j)} = i_{S(j)}$, then the representation (2.1.4) of $\log p$ in terms of the free parameters $\theta = \{\theta_j,\ j \in J\}$ becomes

$$\log p(i) = \theta_0 + \sum_{j \in J, j \lhd i} \theta_j, \quad i \in I \tag{2.1.6}$$

where $\theta_0 = \log p(0)$ is the normalization constant and is determined by requirement $\sum_{i \in I} p(i) = 1$. Based on the Mobius inversion formula of (2.1.6), we can get

$$\theta_j = \sum_{j' \in J,\ j' \lhd j} (-1)^{|S(j)| - |S(j')|} \log \frac{p(j')}{p(0)},\ j \in J. \tag{2.1.7}$$

It is convenient to introduce the vectors

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

where $e_j, j \in J$ are the unit vectors in R^J. Then equation (2.1.6) becomes

$$\log p(i) = \theta_0 + \langle \theta, f_i \rangle = \langle \tilde{\theta}, \tilde{f}_i \rangle, \tag{2.1.8}$$

where $\tilde{\theta} = (\theta_0, \theta)$ and $\tilde{f}_i = (1, f_i)$. The log-linear model (2.1.6) can be rewritten in matrix form as

$$\begin{aligned} (\log \tfrac{p(i)}{p(0)}, i \in I) &= A^t \theta, \text{ or} \\ (\log p(i), i \in I) &= \tilde{A}^t \tilde{\theta}, \end{aligned} \tag{2.1.9}$$

where A is a $J \times I$ matrix whose columns are the f_i vectors and $\tilde{A}$ is a $(J+1) \times I$ matrix whose columns are the $\tilde{f}_i$ vectors. Both A and $\tilde{A}$ are called the design matrices of the log-linear model.

Here we give a hierarchical log-linear model example to help readers understand our notations.

Example 2.1.2. *Let X_a, X_b denote two binary random variables. The sample of X_a, X_b can be classified into a contingency table with cells $I = \{00; 01; 10; 11\}$. Here we consider two hierarchical log-linear models. One is the saturated model with the simplicial complex $\Delta_1 = \{ab, a, b\}$, the other one is the independent model with the simplicial complex $\Delta_2 = \{a, b\}$.*

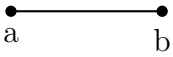


Figure 2.1: The simplicial complex Δ_1

Figure 2.2: The simplicial complex Δ_2

The parameter set for Δ_1 is $J_1 = [01; 10; 11]$, and the parameter set for Δ_2 is $J_2 = [01; 10]$. The absence of parameter θ_{ab} indicates that the two random variables are independent. The design matrix $\tilde{A}_1$ of Δ_1 is

$$\tilde{A}_1 = \begin{array}{c} \\ f_{00} \\ f_{01} \\ f_{10} \\ f_{11} \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \end{array}.$$

Discrete graphical models make up an important subset of the class of the hierarchical log-linear models. A graphical model is a hierarchical log-linear model whose simplicial complex Δ can be represented by an undirected graph such that all the elements of the simplicial complex are the complete induced sub-graphs. First we give some basic definitions from graph theory, and then we consider their Markov properties.

Let $G = (V, E)$ be an undirected graph where V is the vertex set and E is the set of edges. We write (a,b) for the undirected edge between two vertices a and b. We say that a, b are *adjacent* if $(a, b) \in E$. For a given vertex v, the set of its adjacent vertices is called the $neighbours$ of v, which we denote as N_v. If all the vertices are adjacent to each other, the graph is a *complete* graph. The sequence of vertices $\{a_1, a_2, \cdots, a_k\}$ form a *path* in G if $(a_i, a_i + 1) \in E, \forall i = 1, 2, \cdots, k - 1$. A graph is connected if every pair of distinct vertices is joined by a path, otherwise it is disconnected. When a graph is disconnected, we can study each component independently, so we only focus on connected graphs in this thesis. For a subset $A \subset V$, the induced sub-graph G_A is $G_A = (A, E_A)$ where E_A is the set of edges in E with both endpoints in A. We now provide definitions of three concepts that are fundamental to the theory we put forward in this thesis.

Definition 2.1.3. *For $G = (V, E)$ given, a subset $S \subset V$ is called a separator if there exist $A \subset V$, $B \subset V$, such that A, B, S are disjoint, $A \cup S \cup B = V$, $(A \cup S) \cap (B \cup S) = S$ and any path between $a \in A$ and $b \in B$ has to go through S. S is called a minimal separator if no non-trivial*

subset of separator S is a separator.

Definition 2.1.4. *Given $G = (V, E)$, we then say G can be decomposed into $G_{A\cup S}$ and $G_{B\cup S}$ if $S \subset V$ is a complete separator and S separates A from B.*

Definition 2.1.5. *The prime components of a given graph G are the induced sub-graphs that cannot be decomposed and that are maximum in the sense of inclusion. A prime component that is complete is called a maximal clique. From now on when we say clique, we mean a maximal clique unless otherwise specified. If all the prime components are cliques, then the graph is called decomposable. We denote the cliques in a decomposable graph as $\{C_1, C_2, \cdots, C_k\}$.*

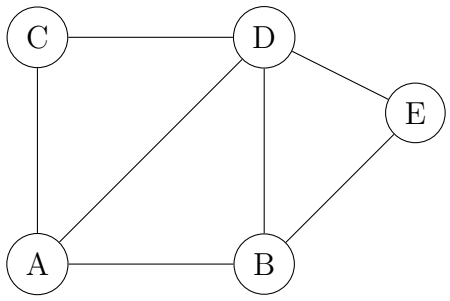


Figure 2.3: A decomposable undirected graph

We give an example of decomposable graph in Figure 2.3, which is decomposed into three cliques $\{ACD, ABD, BDE\}$. Set $\{AD, BD\}$ is a separator set.

When the dimension of the graphical model is high, we often have to work with graphs induced by the vertices $v \cap N_v$ for $v \in V$. We now define one-hop and two-hop neighborhoods of $v \in V$.

Definition 2.1.6. *For a given $v \in V$, we say that $\mathcal{M}_v$ is a one-hop neighborhood of v if it comprises v and its immediate neighbours in G, i.e. if $\mathcal{M}_v = \{v\} \cup \mathcal{N}_v$. We will say that $\mathcal{M}_v$ is a two-hop neighborhood if it comprises v, its immediate neighbours, and the neighbours of the immediate neighbours in G. For simplicity of notation, we will denote both the one-hop and two-hop neighborhoods*

by $\mathcal{M}_v$. We use the notation

$$\mathcal{N}_{2v} = \mathcal{M}_v \setminus \Big(\{v\} \cup \mathcal{N}_v \Big)$$

to denote the set of neighbours of the neighbours of v, as can be seen in Figure 2.4.

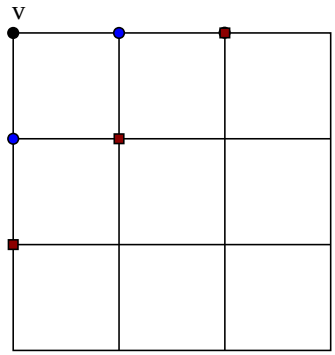


Figure 2.4: Neighbourhood structure in an undirected graph; blue vertices denote the neighbours of vertex v: $\mathcal{N}_v$, red nodes denote the neighbours of vertices in $\mathcal{N}_{2,v}$

Let us now recall Markov properties. Associated with an undirected graph $G = (V, E)$ and a collection of random variables $\{X_v, v \in V\}$ taking value from discrete set I, a probability measure P on I is said to obey

(P) *Pairwise Markov property*, if for any two random variable X_i, X_j,

$$X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}} \text{ if } (i,j) \notin E$$

(L) *Local Markov property*, if

$$X_v \perp\!\!\!\perp X_{V \setminus \{v \cup N_v\}} | X_{N_v},$$

(G) *Global Markov property*, if

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

where subsets A, B are separated by S.

Lauritzen (1996) (Proposition 3.4) showed that $(G) \Rightarrow (L) \Rightarrow (P)$. It's also well know that if the probability measure P is positive on I, the three Markov properties are equivalent. The hierarchical log-linear model we study in this thesis satisfies the positive probability measure condition, so we won't specify which Markov property we are using.

A hierarchical log-linear model is a graphical model Markov with respect to a graph G if its simplicial complex is the set of cliques of G. Here is an example of a model which is hierarchical but not graphical.

Example 2.1.7. *Assume random variables $X = (X_1, X_2, X_3)$ belong to a hierarchical log-linear model generated by $\Delta = \{12, 13, 23\}$. If we try to represent this simplicial complex by a graph, we get a triangle, but the clique (123) doesn't belong to Δ.*

If $\Delta = \{12, 13, 23, 123\}$ or $\Delta = \{12, 23\}$, the hierarchical log-linear model is a graphical model.

2.2 Exponential family and the maximum likelihood estimate

The probability distribution of log-linear model belongs to the natural exponential family $\mathcal{E}_A$ defined as follows

$$\mathcal{E}_A = \{p : p(i) = \exp(\langle \theta, f_i \rangle - k(\theta)),\ \theta \in R^J \text{ and } i \in I\} \tag{2.2.1}$$

where $k(\theta) = -\theta_0 = \log \sum_{i \in I} \exp(\langle \theta, f_i \rangle)$ is the normalization constant or the cumulative generating function, and A is the design matrix with column vectors $f_i, i \in I$.

We assume the cell counts $(n(i), i \in I)$ to follow a multinomial distribution with total counts N, then $\prod_{i \in I} p(i)^{n(i)}$ can be written under the form of exponential family:

$$\begin{aligned} \textstyle\prod_{i \in I} p(i)^{n(i)} &= \exp(\textstyle\sum_{i \in I} n(i) \log p(i)) = \exp(\sum_{i \in I} n(i)(\langle \theta, f_i \rangle - k(\theta))) \\ &= \exp\{\langle \theta, \textstyle\sum_{i \in I} n(i) f_i \rangle - \sum_{i \in I} n(i) k(\theta)))\} \end{aligned} \tag{2.2.2}$$

We write $t = \sum_{i \in I} n(i) f_i$. from the definition of f_i, t is a $|J|$-dimensional vector

$$t = \sum_{i \in I} n(i) \sum_{j \in J, j \lhd i} e_j = \sum_{j \in J} e_j (\sum_{i \in I, j \lhd i} n(i)) = \sum_{j \in J} e_j n_{S(j)}(j_{S(j)}),$$

Since e_j is the unit vector in R^J, $t_j = n_{S(j)}(j_{S(j)})$, i.e. t_j is the $j_{S(j)}$-marginal cell count, which is also the sufficient statistics of the contingency table. We can rewrite the equation (2.2.2) as follows

$$\prod_{i \in I} p(i)^{n(i)} = \exp\{\langle \theta, t \rangle - Nk(\theta)\}.$$

then the log-likelihood function of the contingency table is

$$l(\theta|t) = \langle \theta, t \rangle - Nk(\theta). \tag{2.2.3}$$

In a natural exponential family of the form $\exp\{\langle \theta, t \rangle - Nk(\theta)\}$, the first derivative of the cumulative generating function $Nk(\theta)$ equals the expectation of sufficient statistics t: $E(t) = Nk^{'}(\theta)$, where

$$E(t_j) = Nk_j^{'}(\theta) = N \frac{\sum_{i \in I, j \lhd i} \exp(\theta, f_i)}{\sum_{i \in I} \exp(\langle \theta, f_i \rangle} = N \sum_{i \in I, j \lhd i} p(i) = N p(j_{S(j)})$$

the notation $p(j_{S(j)})$ denotes the marginal probability of cell $j_{S(j)}$ and we denote the vector of marginal probability of cell set $J_{S(J)}$ as $P(\theta) = (p(j_{S(j)}), j \in J)$. Taking the second derivative, we obtain

$$l^{''}(\theta|t) = -Nk^{''}(\theta) = -N(\sum_{i \in I} \frac{\exp\langle \theta, f_i \rangle}{L(\theta)} f_i \otimes f_i - P(\theta) \otimes P(\theta)),$$

where $\otimes$ denotes the outer product. The Fisher information matrix is

$$F = E(-l^{''}(\theta|t)) = N(\sum_{i \in I} \frac{\exp\langle \theta, f_i \rangle}{L(\theta)} f_i \otimes f_i - P(\theta) \otimes P(\theta)).$$

Definition 2.2.1. *A finite parameter value $\hat{\theta}$ is a maximum likelihood estimate(MLE) if it is a global maximum of $l(\theta|t)$:*

$$\hat{\theta} = \arg\max_{\hat{\theta} \in R^J} l(\theta|t)$$

Computing the MLE of the log-likelihood (2.2.3) becomes intractable in the high-dimensional log-linear model because of the complexity of the partition function $k(\theta)$. Later in this thesis, we will consider several composite likelihood methods to approximate the maximum likelihood estimate(MLE).

2.3 The Marginal Polytope and Its Faces

We now define the marginal polytope, a central object for hierarchical log-linear models.

Definition 2.3.1. *Given a log-linear model with design matrix A, the convex hull of the columns $\{f_i, i \in I\}$ is called the marginal polytope of the log-linear model, and denoted by $\mathbf{P}_\Delta$ or $\mathbf{P}_A$,*

$$\mathbf{P}_A = \{x = \sum_{i=1}^{I} \lambda_i f_i, \forall \lambda_i \geq 0 \text{ and } \sum \lambda_i = 1\}$$

Since $\frac{t}{N} = \sum_{i \in I} \frac{n(i)}{N} f_i$, $\frac{t}{N} \in \mathbf{P}_A$. As a result, the marginal polytope comprises the set of all possible observable sufficient statistics. Lemma 3.2.2 of the following section shows that the MLE of the parameters θ in (2.2.3) doesn't exist if and only if the sufficient statistics lie on a face of the marginal polytope $\mathbf{P}_A$. We now consider the notation and concept of face of a polytope.

Definition 2.3.2. *A set $\mathbf{P} \subset \mathbf{R}^h$ is a* (convex) polytope *if $\mathbf{P}$ is the convex hull of a finite subset of $\mathbf{R}^h$. Equivalently, a polytope can be defined as a bounded subset of $\mathbf{R}^h$ defined by linear inequalities.*

Definition 2.3.3. *For any vector $g \in \mathbf{R}^h$ and any constant $c \in \mathbf{R}$, define three sets $H_{g,c} = \{x \in \mathbf{R}^h : \langle g, x \rangle = c\}$, $H^+_{g,c} = \{x \in \mathbf{R}^h : \langle g, x \rangle \geq c\}$ and $H^-_{g,c} = \{x \in \mathbf{R}^h : \langle g, x \rangle \leq c\}$. If $g \neq 0$, then $H_{g,c}$ is an* (affine) hyperplane*, and $H^+_{g,c}$ and $H^-_{g,c}$ are the* positive *and* negative halfspaces *defined by g and c.*

Let $\mathbf{P} \subseteq \mathbf{R}^h$ *be a polytope, let* $g \in \mathbf{R}^h$ *and* $c \in \mathbf{R}$*, and suppose that* $\mathbf{P} \subset H^+_{g,c}$ *or* $\mathbf{P} \subset H^-_{g,c}$*. Then* $\mathbf{F} := H_{g,c} \cap \mathbf{P}$ *is called a* face *of* $\mathbf{P}$*. If* $g \neq 0$*, then* $H_{g,c}$ *is called a* supporting hyperplane *of* $\mathbf{P}$*. If* $\mathbf{F} \neq \mathbf{P}$ *and* $\mathbf{F} \neq \emptyset$*, then* $\mathbf{F}$ *is a* proper face *of* $\mathbf{P}$*.*

The dimension of a face $\mathbf{F}$ *is the dimension of the smallest affine subspace of* $\mathbf{R}^h$ *that contains it. Its co-dimension is* $\dim(\mathbf{P}) - \dim(\mathbf{F})$*. A* facet *of a polytope* $\mathbf{P}$ *is a proper face that is maximal with respect to inclusion and is thus of co-dimension 1. A minimal proper face of a polytope is a singleton* $\{p\} \subseteq \mathbf{P}$*; in this case,* p *is a* vertex*.*

Intersections of faces are again faces: If $g_1, g_2 \in \mathbf{R}^h$ and $c_1, c_2 \in \mathbf{R}$ define faces $\mathbf{F}_1, \mathbf{F}_2$ of $\mathbf{P}$ and if $\mathbf{P} \subset H^+_{g_1,c_1} \cap H^+_{g_2,c_2}$, then $\mathbf{P} \subset H^+_{g_1+g_2,c_1+c_2}$, and $\mathbf{F}_1 \cap \mathbf{F}_2 = \mathbf{P} \cap H_{g_1+g_2,c_1+c_2}$. Any face is an intersection of facets.

By definition, every face $\mathbf{F}$ of a polytope $\mathbf{P} \subset \mathbf{R}^h$ is characterized by a linear inequality $\langle g, x \rangle \geq c$ that is valid on $\mathbf{P}$ and that holds as an equality on $\mathbf{F}$. This linear inequality is unique only if $\mathbf{F}$ is a facet. Sometimes it is convenient to give all linear equations that hold on a face $\mathbf{F}$. These linear equations determine the smallest affine subspace of $\mathbf{R}^h$ containing $\mathbf{F}$.

When a polytope is defined as the convex hull of a finite number of points f_i, $i \in I$, then it is of interest to know which subsets of $\{f_i\}_{i \in I}$ lie on a common face. Indeed, it is the purpose of this thesis to compute the smallest face of the marginal polytope containing the data vector t, and we determine this face by identifying which vectors f_i belong to it.

Definition 2.3.4. *For a finite set* I *let* $\{f_i\}_{i \in I} \subset \mathbf{R}^h$*, and let* $\mathbf{P}$ *be the convex hull of* $\{f_i\}_{i \in I}$*. A subset* $F \subseteq I$ *is called* facial *(with respect to* $\mathbf{P}$*), if there exists a face* $\mathbf{F}$ *of* $\mathbf{P}$ *with* $F = \{i : f_i \in \mathbf{F}\}$*. For any subset* $S \subseteq I$*, denote by* $F_{\mathbf{P}}(S)$ *the smallest facial set that contains* S*.*

Since the intersection of facial sets is again facial, $F_{\mathbf{P}}(S)$ is well-defined.

Lemma 2.3.5. *Let $\{f_i\}_{i\in I} \subset \mathbf{R}^h$, let $\phi : \mathbf{R}^h \to \mathbf{R}^{h'}, x \mapsto Bx + d$ be an affine map, and let $f'_i = \phi(f_i)$. If $\mathbf{P}$ is the convex hull of the f_i, then $\mathbf{P}' := \phi(\mathbf{P})$ is the convex hull of the f'_i. The faces and facial sets of $\mathbf{P}$ and $\mathbf{P}'$ are related as follows:*

1. *Any inequality $\langle g', x'\rangle \geq c'$ that is valid on $\mathbf{P}'$ corresponds to an inequality $\langle g, x\rangle \geq c$ that is valid on $\mathbf{P}$, where $g = B^t g'$ and $c = c' - \langle g', d\rangle$. Thus, if $\mathbf{F}'$ is a face of $\mathbf{P}'$, then $\phi^{-1}(\mathbf{F}')$ is a face of $\mathbf{P}$.*

2. *A subset of I that is facial with respect to $\mathbf{P}'$ is also facial with respect to $\mathbf{P}$. Thus, $F_{\mathbf{P}}(S) \subseteq F_{\mathbf{P}'}(S)$ for any $S \subseteq I$.*

Proof. The first statement follows from

$$c \leq \langle g', \phi(f_i)\rangle = \langle g', Bf_i + d\rangle = \langle B^t g', f_i\rangle + \langle g', d\rangle,$$

which holds for any $i \in I$. The second statement follows immediately from the equation above and the fact that $F_{\mathbf{P}}(S)$ is the smallest facial set containing S. □

We note that in Lemma 2.3.5, the dimension of $\phi(\mathbf{P})$ is at most equal to h. We only apply Lemma 2.3.5 to coordinate projections ϕ with $h' < h$.

Remark 2.3.1. *Sometimes it is convenient to embed the polytope in a vector space that has one additional dimension using a map $\mathbf{R}^h \to \mathbf{R}^{h+1}, x \mapsto \tilde{x} := (1, x)$. This has the advantage that all defining inequalities can be brought into a homogeneous form with vanishing constant c: Note that $\langle g, f_i\rangle - c = \langle \tilde{g}_c, \tilde{f}_i\rangle$, where $\tilde{g}_c := (c, g)$.*

When a defining inequality of a face $\mathbf{F}$ is given, its facial set F can be obtained by checking whether $f_i \in \mathbf{F}$ for each $i \in I$. In the other direction, when a facial set F is given, it is much more difficult to compute a defining inequality of the corresponding face $\mathbf{F}$. However, it is straightforward

to compute the linear equations defining $\mathbf{F}$: The set of such equations $0 = \langle g, x\rangle - c = \langle \tilde{g}, \tilde{x}\rangle$ corresponds to the set of vectors $\tilde{g} \in \ker \tilde{A}_F^t$, where $\tilde{A}_F$ is the matrix obtained from A by adding a row of ones and dropping the columns not in F.

To sum up, we recall the two binary random variables hierarchical log-linear model example to illustrate the basic concepts we covered in this section.

Example 2.3.6 (Two binary variables example)**.** *Consider two binary random variables, X_a, X_b, under the saturated hierarchical model. Let $\Delta = \{\{a\}, \{b\}, \{a, b\}\}$. That is, it contains all possible probability distributions with full support.*

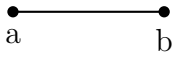


Figure 2.5: The simplicial complex Δ

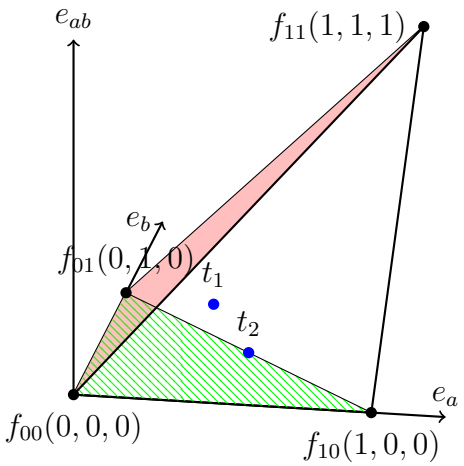


Figure 2.6: The marginal polytope $\mathbf{P}_\Delta$

The design matrix of this model is

$$\tilde{A} = \begin{array}{c} \overbrace{}^{f_{00}} \overbrace{}^{f_{01}} \overbrace{}^{f_{10}} \overbrace{}^{f_{11}} \\ \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{l} \theta_{00} \\ \theta_{01} \\ \theta_{10} \\ \theta_{11} \end{array} \quad \textbf{\textit{or}} \quad A = \begin{array}{c} \overbrace{}^{f_{00}} \overbrace{}^{f_{01}} \overbrace{}^{f_{10}} \overbrace{}^{f_{11}} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{l} \theta_{01} \\ \theta_{10} \\ \theta_{11} \end{array}$$

In the following, we give two contingency tables and their corresponding sufficient statistics. The first one belongs to the relative interior of the marginal polytope $\mathbf{P}_\Delta$*, and the second one belongs to a proper face of* $\mathbf{P}_\Delta$*.*

- *sample 1:* $\{n(00) = 2, n(10) = n(01) = n(11) = 1\}$*;* $\frac{t_1}{N} = [0.4, 0.4, 0.2]$*, not on the face;*
- *sample 2:*$\{n(00) = n(11) = 0, n(10) = n(01) = 1\}$*;* $\frac{t_2}{N} = [0.5, 0.5, 0]$*, on face.*

3 Review of literature

3.1 Contingency tables, log-linear models: early developments

The history of the log-linear model and contingency tables is given in Fienberg and Rinaldo (2007), from which we extract some important features that are related to our research.

The term "contingency tables" refers to tables of cross-classified categorical data. Computing the MLE for contingency tables started with Bartlett (1935) who showed that you can get the MLE of a $2 \times 2 \times 2$ table under the model with no three-way interaction and fixed two-way marginal totals by solving a cubic equation. Here we give Bartlett's example, but use our notation as follows:

Example 3.1.1. *Consider data of three binary variables, which is classified into a* $2 \times 2 \times 2$ *contingency table. Bartlett's model is based on the following two assumptions: no three-way interactions, and fixed two-way marginal totals. Let* $I = \{000, 100, 010, 110, 001, 101, 011, 111\}$ *be the set of cells, and* $n = \{n(i);\ i \in I\}$ *as the observed cell counts.*

The cell probabilities of Bartlett's model should fit the following equation:

$$p(000)p(110)p(101)p(011) = p(010)p(100)p(001)p(111) \tag{3.1.1}$$

Since Bartlett assumes that the two-way marginal totals are fixed, whenever we adjust the count of one cell, all other cell counts will make the same or the opposite adjustment. For example, if we add a value c to $n(000)$, we need to minus c from $n(100)$ due to fixed total $n(+00)$. Therefore

the deviations from expectation in all cells are the same, which we denote as x here. The MLE of the cell counts should fit Equation (3.1.1), and therefore can be solved with the following cubic equation:

$$(n(000)+x)(n(110)+x)(n(101)+x)(n(011)+x) = (n(010)-x)(n(100)-x)(n(001)-x)(n(111)-x) \tag{3.1.2}$$

Bartlett was first to study the MLE computation of contingency tables, but he didn't consider two fundamental problems:

1. The systematic computation of the MLE;

2. The existence of this MLE.

As can readily been seen in Example 3.1.1, if cell counts $n(000) = 0$, $n(111) = 0$, then solving the cubic Equation (3.1.2) will always end up with a negative cell count, i.e. the MLE of this contingency table doesn't exist.

Deming and Stephan (1940) proposed the practical Iterative Proportional Fitting (IPF) algorithm to solve Equation (3.1.1): . To compute the MLE of the expected cell counts, the IPF updates the cell counts iteratively using fixed marginal counts. The IPF is still used nowadays and we will use it later in this thesis.

Roy and Kastenbaum (1956) studied three dimensional contingency tables with no three-way interaction, and without fixing the marginal totals. They offered a new functional representation of cell probabilities in any three-way interactions three dimensional contingency table(not limited to $2 \times 2 \times 2$ tables):

$$p(ijk) = \frac{p(ij+)p(i+k)p(+jk)}{p(i++)p(+j+)p(++k)} \tag{3.1.3}$$

To compute the MLE of the cell counts, they use Lagrangian multipliers to make the likelihood function subject to constraint (3.1.3). Both Bartlett (1935) and Roy and Kastenbaum (1956) didn't concern themselves with the existence of the MLE, maybe due to the fact that the contingency tables they considered were of small dimension and the cell counts were all positive.

Birch (1963) first introduced the log-linear model of three dimensional contingency tables, and this brought the research of contingency tables into a new era. Birch took the logarithm of (3.1.3):

$$\log p(ijk) = -\log p(i++) - \log p(+j+) - \log p(++k) + \log p(ij+) + \log p(i+k) + \log p(+jk),$$

which in general can be written as,

$$\log p(ijk) = u + u_{1i} + u_{2j} + u_{3k} + u_{12ij} + u_{13ik} + u_{23jk} + u_{123ijk}, \tag{3.1.4}$$

where in this case $u_{123ijk} = 0$ since there is no three-way interaction. Birch (1963) derived the likelihood function with respect to the log-linear parameters and computed the MLE. He also showed that the MLE exists if all the cell counts are positive. Since then, the study of log-linear models has drawn a lot of attention from the research community. Some of the first books on this subject are Haberman (1974a) and Bishop et al. (1975b).

3.2 Existence of the MLE

The study of the existence of the MLE started at almost the same time as the study of log-linear models. Fienberg (1970) gave sufficient conditions for the existence of the MLE under the assumption that the model they consider cannot be written as the product of several independent models. Fienberg's sufficient conditions are: (1) the observed data cannot be split into several disjoint subtables; (2) the observed marginal totals are positive.

Haberman (1974b) gave a necessary and sufficient condition for the existence of the MLE, which we state as a lemma here:

Lemma 3.2.1 (Haberman (1974b)). *Let A, a $|J| \times |I|$ matrix, be the design matrix of the log-linear model, and let $n = (n(i), i \in I)$ be the vector of the observed cell counts. A necessary and sufficient condition for the existence of the MLE is that there exists $z \in ker(A)$ such that $n + z > 0$.*

Since $Az = 0$, we have $An = A(n + z) = t$, i.e. the two cell counts n and $n + z$ have the same sufficient statistic, and again since $n + z > 0$, the MLE exists. For discrete log-linear models, the distribution of cell counts is an exponential family. Barndorff-Nielsen (1978) (Theorem 9.13 and Corollary 9.6) gave necessary and sufficient conditions for the existence of the MLE of the canonical parameters in the exponential family. Barndorff-Nielsen showed that the MLE exists if and only if the data belongs to the relative interior of the convex support of the distribution. Neither Haberman (1974b) nor Barndorff-Nielsen (1978)'s conditions are constructive. Eriksson et al. (2006) gave a practical algorithm to detect the existence of the MLE. First they developed a geometric interpretation of Lemma 3.2.1 as follows.

Lemma 3.2.2 (Eriksson et al. (2006)). *The MLE of the log-linear models exists if and only if the marginal totals(sufficient statistics) $t = A * n$ belong to the relative interior of the marginal polytope C_A. In other words the MLE doesn't exist if and only if t belongs to a face of C_A.*

The term "marginal polytope" was introduced by Wainwright and Jordan (2003), and denotes the convex hull spanned by the f_i's as defined earlier in this thesis. Eriksson et al. (2006) gave an algorithm for determining if the sufficient statistic t lies on a facet of the marginal polytope. This was further developed by Fienberg and Rinaldo (2012), who proposed to check if the sufficient statistic belongs to a face of the marginal polytope using a linear programming method as well

as other methods. While their methodology can handle low dimensional data, it cannot be used for more than 16 binary variables. In Chapter 7, we extend their approach to high dimensional settings. This is done by finding good inner and outer approximations to the smallest face F of the marginal polytope containing the data, i.e. by finding a face containing or contained by F as close as possible in a sense that will be made clear in Chapter 7.

3.3 Computation of the MLE

When the dimension of the data becomes very large, neither the IPF algorithm nor regular convex optimization methods are feasible for the MLE computation. The likelihood function is intractable. In machine learning literature, a lot of effort has been devoted to the approximation of this likelihood function. Peterson (1987) defined and applied a mean field learning algorithm for neural networks. The basic idea is to approximate the complex CDF function (also called the partition function in machine learning literature) by its mean. Saul et al. (1996) developed a mean field theory for sigmoid belief networks, where they used a completely factorized distribution Q to approximate the intractable distribution P by minimizing the Kullback-Leibler divergence between P and Q: $KL(Q|P) = \sum Q \log \frac{Q}{P}$. For more variational methods, readers can refer to the following review papers: Jordan et al. (1999) and Wainwright and Jordan (2008).

Recently another line of research on composite likelihood has become active, for instance, Dillon and Lebanon (2010), Sutton and McCallum (2007), Asuncion et al. (2010), Wiesel and Hero (2012) and Liu and Ihler (2012). The history of composite likelihood methods can be traced back to the 1970s. Besag (1974) first studied the conditional probability models for finite system of lattice data. The conditional probability models approach was extended to non-lattice data in Besag (1975). Besag proposed one special conditional composite likelihood technique, the product of local

conditional densities of a single variable given its neighbours, which he named "pseudo-likelihood". Lindsay (1988) proposed a more general version of pseudo-likelihood, which he named "composite likelihood". Following the definition proposed by Dillon and Lebanon (2010), we now give the definition of composite likelihood,

Definition 3.3.1. *Let $X = (X_1, X_2, \cdots, X_p)$ be a random variable with a given probability density function $p(x|\theta)$ parameterized by θ. Let $(X_{A_i}, X_{B_i}), i = 1, 2, \cdots, k$ be k pairs of subsets of the random variables, where $A_i \neq \emptyset$ and $A_i \cap B_i = \emptyset$. The composite likelihood for θ corresponding to the pairs $(X_{A_i}, X_{B_i}), i = 1, 2, \cdots, k$ is the product of the local likelihoods associated to the conditional probabilities of X_{A_i} given X_{B_i}, $p(x_{A_i}^{(n)}|x_{B_i}^{(n)}; \theta)$. For a given sample $\{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$, this composite log-likelihood $cl(\theta)$ is therefore equal to*

$$cl(\theta) = \sum_{n=1}^{N} \sum_{i=1}^{k} \log p(x_{A_i}^{(n)}|x_{B_i}^{(n)}; \theta). \tag{3.3.1}$$

This is a very general definition of the composite likelihood. By choosing different A_i and B_i, one can get various types of composite likelihood (see in Varin et al. (2011)). We note that for $B_i = \emptyset,\ i = 1, 2, \cdots, k$, $cl(\theta)$ is the sum of the logarithm of the likelihoods associated to the local marginal probabilities. When working with graphical models, the most commonly used composite likelihoods are those associated with the pairs $A_v = \{v\}$ and $B_v = \{N_v\}$, $v = 1, 2, \cdots, p$. The maximum composite likelihood estimate of θ(abbreviated MCLE) is the value of θ that maximizes the composite likelihood as given in (3.3.1). Lindsay (1988) showed that the MCLE is asymptotically normally distributed with a covariance which is larger, in the positive definite matrix sense, than that of the regular MLE.

Maximizing the composite likelihood is still a difficult task in a high dimensional setting. For discrete graphical log-linear models, Liu and Ihler (2012) first proposed to compute the MCLE

by maximizing separately each of the local components $l^v(\theta) = \sum_{i=1}^n \log p(x_v^{(i)}|x_{N_v}^{(i)})$, $v \in V$ by distributed computing, and subsequently combining the local estimates through linear consensus or maximum consensus to achieve a global estimate for θ. They showed that this global estimate is consistent. For Gaussian graphical models, Wiesel and Hero (2012) also proposed the marginal composite likelihood method as well as the pseudo-likelihood method. The local component of their marginal likelihood is $l^v(\theta) = \sum_{i=1}^n \log p(x_v^{(i)}, x_{N_v}^{(i)})$. Like Liu and Ihler (2012), to find the MCLE, Wiesel and Hero (2012) also used distributed computing and combined local results by an averaging scheme or by ADMM. They also proved that the local marginal likelihood estimator is equal to the local conditional estimator in each component. Meng et al. (2013) named the one-hop MCLE the MCLE from the composite likelihood built from the local marginal model

$$l^{M,1}(\theta) = \sum_{i=1}^n \sum_{v \in V} \log p(x_v^{(i)}, x_{N_v}^{(i)}; \theta).$$

They then proposed the two-hop MCLE obtained by maximizing

$$l^{M,2}(\theta) = \sum_{i=1}^n \sum_{v \in V} p(x_v^{(i)}, x_{N_v}^{(i)}, x_{N_{2,v}}^{(i)}; \theta).$$

As in the one-hop case, the two-hop MCLE is obtained by combining local maxima. They showed numerically that the two-hop estimate was more accurate than the one-hop estimate under increased computational cost. However, they stated that the two-hop estimate obtained from the local marginal and conditional likelihoods are different. In our arXiv paper Massam and Wang (2013), we showed that for the discrete model, the asymptotic variance of the two-hop estimate is smaller than the asymptotic variance of the one-hop estimate. Following our paper, Meng et al. (2014) proved a parallel theorem for the Gaussian graphical model and studied the asymptotic properties of their estimates. In this thesis, parallel to their method on Gaussian graphical models, we study the marginal likelihood and conditional likelihood in discrete log-linear models both in

the one-hop case and two-hop case. First we prove that the conditional and marginal estimates, one-hop and two-hop are equal. We then proceed to studying the asymptotic properties of these estimates.

4 Approximating the maximum likelihood estimate

In this section, we are going to study the first topic: the systematic computation of the MLE in hierarchical log-linear models. To get the MLE, we need to solve the following optimization problem:

$$\hat{\theta}^g = \arg\max \quad l(\theta) = \arg\max \langle \theta, \frac{t}{N} \rangle - \log \sum_{i \in I} \exp\langle \theta, f_i \rangle \tag{4.0.1}$$

As we mentioned before, the log-partition function $k(\theta) = \log \sum_{i \in I} \exp\langle \theta, f_i \rangle$ is untraceable in high-dimensional log-linear models. To avoid this problem, we can use the composite likelihood methods in Definition 3.3.1. There are various types of composite likelihoods described in the literature, the most popular one being defined as the product over all vertices $v \in V$ of the local conditional likelihood for X_v given $X_{\mathcal{N}_v}$, where $\mathcal{N}_v$ denotes the set of neighbours of v in graph G. This type of composite likelihood method breaks down equation (4.0.1) into the sum of p local composite likelihood functions:

$$cl(\theta) = \sum_{v \in V} l_v(\theta^v) = \sum_{v \in V} \sum_{i=1}^{n} \log p(x_v^{(i)} | x_{\mathcal{N}_v}^{(i)}; \theta^v),$$

where θ^v is a subset of θ which contains the parameters involved in $p(X_v | X_{\mathcal{N}_v})$

In recent work on high-dimensional Gaussian graphical models, Wiesel and Hero (2012) and Meng et al. (2013) take another approach. They use a different composite likelihood which is the product, over all vertices $v \in V$, of local marginal likelihoods. In this section, we first recall the definition of the conditional composite likelihood estimate, then extend the marginal composite

likelihood in Meng et al. (2013) to discrete graphical models, and finally show that the maximum likelihood estimates obtained from these two types, conditional and marginal, of local models are in fact identical and thus the composite likelihood obtained by any type of consensus from these two types of likelihood are equal. The computational complexity of the marginal computation is exponential in the number of vertices in the neighborhood of v, whereas the conditional computation is linear in this number, so there is no advantage in working with marginal composite likelihoods.

4.1 Conditional composite likelihood methods

We first define the standard conditional composite likelihood function. For $i = (i_v, v \in V)$, let $X^{(1)}, \ldots, X^{(N)}$ be a sample of size N from the distribution of X, which belongs to a hierarchical log-linear model $\mathcal{M}_\Delta$. We recall that the global log-likelihood function is

$$l(\theta) \propto \sum_{i=1}^{N} \log p(X^{(i)}) = \langle \theta, t \rangle - Nk(\theta) \tag{4.1.1}$$

For a given vertex $v \in V$, let $\mathcal{N}_v$ be the set of neighbours of v in the given graph G. The composite likelihood function based on the local conditional distribution of X_v given $X_{V\setminus\{v\}}$ or equivalently, due to the Markov property, the conditional distribution of X_v given its neighbours $X_{\mathcal{N}_v}$ is $L^{PS}(\theta) = \prod_{v\in V} L^{v,PS}(\theta)$ where

$$L^{v,PS}(\theta) = \prod_{i=1}^{N} p(X_v^{(i)} | X_{\mathcal{N}_v}^{(i)}; \theta) \tag{4.1.2}$$

and the superscript "PS" stands for "pseudo-likelihood", the name often given to the conditional composite likelihood (Besag (1974)). As given by (2.1.4), for a given cell i, we have

$$\begin{aligned} \log p(i) &= \log p(X_v = i_v, v \in V) = \theta_0 + \sum_{j \lhd i} \theta_j \\ &= \theta_0 + \sum_{j \lhd i,\ S(j) \subseteq v \cup \mathcal{N}_v, S(j) \not\subseteq \mathcal{N}_v} \theta_j + \sum_{j \lhd i,\ S(j) \subseteq \mathcal{N}_v} \theta_j + \sum_{j \lhd i,\ S(j) \not\subseteq v \cup \mathcal{N}_v} \theta_j \end{aligned}$$

Let

$$J^{PS_v} = \{j \in J \mid S(j) \subseteq v \cup \mathcal{N}_v, S(j) \not\subseteq \mathcal{N}_v\} = \{j \in J \mid v \in S(j)\},$$

next we show that elements of set J^{PS_v} index the parameters in the v-th component in the conditional likelihood function, i.e. $p(X_v^{(i)}|X_{\mathcal{N}_v}^{(i)})$. For $i_v \neq 0$, we have

$$\begin{aligned} p(X_v = i_v|\ X_{\mathcal{N}_v} = i_{\mathcal{N}_v}) = p(X_v = i_v|\ X_{V\setminus\{v\}} = i_{V\setminus\{v\}}) &= \frac{p(X_V = i_V)}{p(X_{V\setminus\{v\}} = i_{V\setminus\{v\}})} \\ = \frac{e^{\theta_0+\sum_{j\lhd i,\ j\in J^{PS_v}}\theta_j+\sum_{j\lhd i,\ S(j)\subseteq\mathcal{N}_v}\theta_j+\sum_{j\lhd i,\ S(j)\not\subseteq v\cup\mathcal{N}_v}\theta_j}}{\sum_{k\in I|\ k_{V\setminus\{v\}}=i_{V\setminus\{v\}}}\left(e^{\theta_0+\sum_{j\lhd k,\ j\in J^{PS_v}}\theta_j+\sum_{j\lhd k,\ S(j)\subseteq\mathcal{N}_v}\theta_j+\sum_{j\lhd k,\ S(j)\not\subseteq v\cup\mathcal{N}_v}\theta_j}\right)} & \\ = \frac{e^{\sum_{j\lhd i,\ j\in J^{PS_v}}\theta_j}}{1+\sum_{k\in I|\ k_{V\setminus\{v\}}=i_{V\setminus\{v\}},\ k_v\neq 0} e^{\sum_{j\lhd k,\ j\in J^{PS_v}}\theta_j}} & \qquad (4.1.3) \end{aligned}$$

and

$$p(X_v = 0|\ X_{V\setminus\{v\}} \ = \ i_{V\setminus\{v\}}) = \frac{1}{1+\sum_{k\in I|\ k_{V\setminus\{v\}}=i_{V\setminus\{v\}},\ k_v\neq 0} e^{\sum_{j\lhd k,\ j\in J^{PS_v}}\theta_j}} \qquad (4.1.4)$$

Equality (4.1.3) is due to the fact that the set of $j \in J$ such that $j \lhd k,\ S(j) \not\subseteq v \cup \mathcal{N}_v$, is the same whether $k_v = i_v$ or $k_v \neq i_v$, and therefore the term $e^{\theta_0+\sum_{j\lhd k,\ S(j)\not\subseteq k_v\cup\mathcal{N}_v}\theta_j}$ cancels out at the numerator and the denominator. The same goes for the set of $j \in J$ such that $j \lhd k,\ S(j) \subseteq \mathcal{N}_v$.

Remark 4.1.1. *In the equation above, we worked with $p(X_v|X_{V\setminus\{v\}})$ rather than with $P(X_v|X_{\mathcal{N}_v})$, though the two are equal; we did this to emphasize that*

$$\theta^{v,PS} = (\theta_j,\ j \in J^{PS_v}),\ \ v \in V \qquad (4.1.5)$$

of the v-th component $L^{v,PS}$ of conditional composite distribution is a sub vector of θ, the parameter of the global likelihood function.

Except for the pseudolikelihood, there are also some other types of conditional composite likelihood methods. Asuncion et al. (2010) proposed their version of composite likelihood which is the

conditional likelihood of a subset of random variables conditional on another subset. By increasing the size of the local components, the composite likelihood estimation can be made more accurate, but computational complexity is sacrificed. In our research, we modified the pseudo-likelihood based on this idea and proposed the two-hop conditional composite likelihood.

The two-hop conditional composite likelihood function is $L^{PS_2}(\theta) = \prod_{v\in V} L^{v,PS_2}(\theta)$ where

$$L^{v,PS_2}(\theta) = \prod_{i=1}^{N} p(X_v^{(i)}, X_{\mathcal{N}_v}^{(i)} | X_{\mathcal{N}_{2v}}^{(i)}). \tag{4.1.6}$$

The expression of $p(X_v^{(i)}, X_{\mathcal{N}_v}^{(k)} | X_{\mathcal{N}_{2v}}^{(k)})$ is the same as (4.1.3) and (4.1.4) but with $J^{v,PS}$ replaced by J^{v,PS_2} where

$$J^{v,PS_2} = \{j \in J \mid S(j) \subseteq \mathcal{M}_v, S(j) \not\subseteq \mathcal{N}_{2v}\}.$$

In a parallel way to Remark 4.1.1, we note that

$$\theta^{v,PS_2} = \{\theta_j, j \in J^{v,PS_2}\}$$

is a sub vector of $\theta = (\theta_j, j \in J)$, the argument of the global likelihood function.

Let $\mathcal{M}_v$ be the one-hop or two-hop neighborhood of v. The marginal composite likelihood is the product

$$L^{\mathcal{M}}(\theta) = \prod_{v\in V} \prod_{k=1}^{N} p(X_{\mathcal{M}_v}^{(k)}) = \prod_{v\in V} L^{\mathcal{M}_v}(\theta). \tag{4.1.7}$$

where $L^{\mathcal{M}_v}(\theta) = \prod_{k=1}^{N} p(X_{\mathcal{M}_v}^{(k)})$. The $\mathcal{M}_v$-marginal model is clearly multinomial and the corresponding data can be read in the $\mathcal{M}_v$-marginal contingency table obtained from the full table. The density of the $\mathcal{M}_v$-marginal multinomial distribution is of the general exponential form

$$f(t^{\mathcal{M}_v};\ \theta^{\mathcal{M}_v}) = \exp\{\langle t^{\mathcal{M}_v}, \theta^{\mathcal{M}_v}\rangle - N k^{\mathcal{M}_v}(\theta^{\mathcal{M}_v})\} \tag{4.1.8}$$

where $t^{\mathcal{M}_v}$, $\theta^{\mathcal{M}_v}$ and $k^{\mathcal{M}_v}$ are respectively the $\mathcal{M}_v$-marginal canonical statistic, canonical parameter and cumulate generating function.

In order to identify the $\mathcal{M}_v$-marginal model, we first establish the relationship between θ and $\theta^{\mathcal{M}_v}$. For the remainder of this thesis, the symbol j is to be understood as an element of $I_{\mathcal{M}_v}$ whenever used in the notation $\theta_j^{\mathcal{M}_v}$, and it is to be understood as the element of J obtained by padding it with entries $j_{V\setminus\mathcal{M}_v} = 0$ whenever used in the notation θ_j. We now give the general relationship between the parameters of the overall model, and those of the $\mathcal{M}_v$-marginal model. The proof is given in Appendix B.1.

Lemma 4.1.1. *Let $\mathcal{M}_v$ be the one-hop or two-hop neighborhood of $v \in V$. For $j \in J, S(j) \subset \mathcal{M}_v$, the parameter θ_j of the overall model, and the parameter $\theta_j^{\mathcal{M}_v}$ of the marginal model are linked by the following:*

$$\theta_j^{\mathcal{M}_v} = \theta_j + \sum_{j' | j' \lhd_0 j} (-1)^{|S(j) - S(j')|} \log\big(1 + \sum_{i \in I, i_{\mathcal{M}_v} = j'} \exp \sum_{\substack{k | k \lhd i \\ k \not\lhd j'}} \theta_k\big) \tag{4.1.9}$$

We want to identify which of the marginal parameters are equal to the corresponding overall parameter, and in particular which marginal parameters are equal to zero when the global parameter is equal to zero. Let $\mathcal{M}_v^c$ denote the complement of $\mathcal{M}_v$ in V. We define the buffer set at v as follows:

$$\mathcal{B}_v = \{w \in \mathcal{M}_v \mid \exists w' \in \mathcal{M}_v^c \text{ with } (w, w') \in E\}. \tag{4.1.10}$$

We have the following result.

Lemma 4.1.2. *Let $\mathcal{M}_v$ be the one-hop or two-hop neighborhood of $v \in V$. For $j \in J, S(j) \subset \mathcal{M}_v$ the following holds:*

(1.) if $S(j) \not\subset \mathcal{B}_v$, then $\theta_j^{\mathcal{M}_v} = \theta_j$,

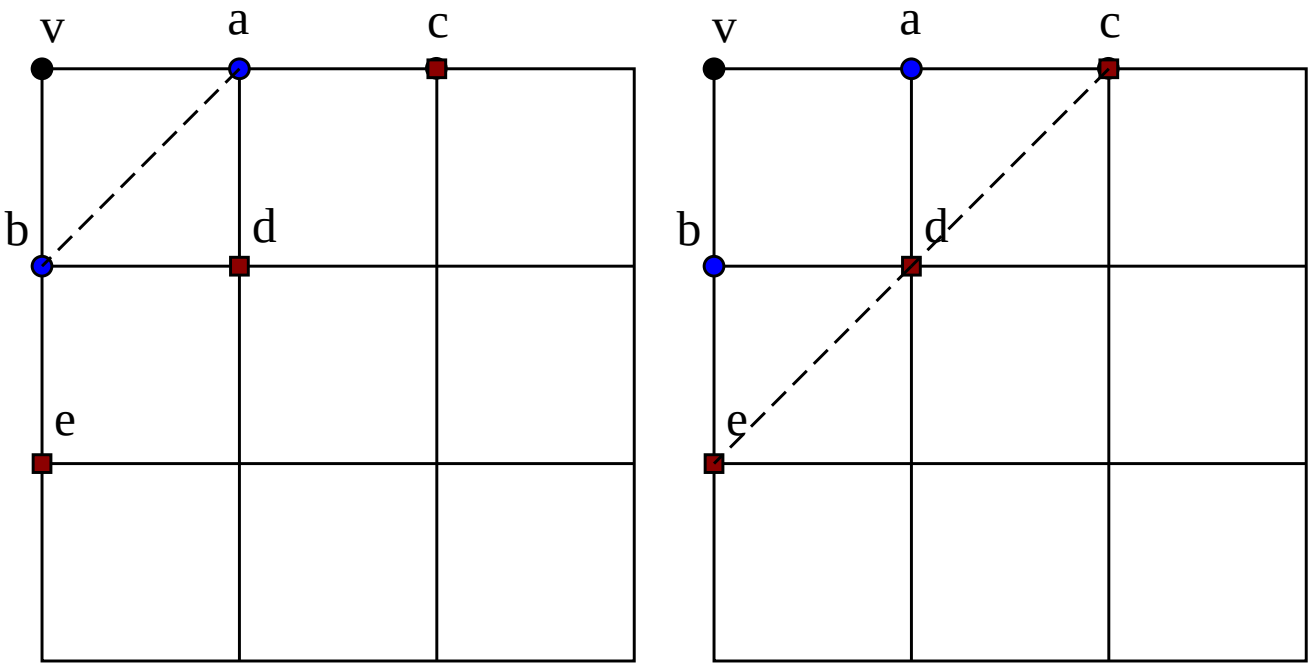
(2.) if $S(j) \subset \mathcal{B}_v$, then in general $\theta_j^{\mathcal{M}_v} \neq \theta_j$, and (4.1.9) holds.

Moreover, for $i \in I, S(i) \subset \mathcal{M}_v$,

(3.) If $S(i) \not\subset \mathcal{B}_v$, *then* $\theta_i^{\mathcal{M}_v} = 0$ *whenever* $\theta_i = 0$.

The proof is given in Appendix B.2. From the lemma above, we see that, for $j \in J$ such that $S(j) \subset \mathcal{M}_v, S(j) \not\subset \mathcal{B}_v$, the corresponding global and $\mathcal{M}_v$-marginal log-linear parameters are equal. We see also that for $i \in I$ such that $S(i) \in \mathcal{M}_v, S(i) \not\subset \mathcal{B}_v$, if the log-linear parameter is zero in the global model, it remains zero in the $\mathcal{M}_v$-marginal model.

4.2 A convex relaxation of the local marginal models



(a) convex relaxation of one-hop marginal model

(b) convex relaxation of two-hop marginal model

Figure 4.1: The convex relaxation of the one-hop and two-hop marginal models of vertex "v" in the 4×4 grid graph

It is clear from (4.1.9) that even though maximizing the marginal likelihood from (4.1.8) is convex in $\theta^{\mathcal{M}_v}$, it is not convex in θ. We would therefore like to replace the problem of maximizing

the likelihood function (4.1.8) non convex in θ by a convex relaxation problem. We know from *(1.)* of Lemma 4.1.2 that $\theta_j^{\mathcal{M}_v} = \theta_j$ for j in the set $\{j \in J : \ S(j) \subset \mathcal{M}_v, \ S(j) \not\subset \mathcal{B}_v\}$.

We also know from *(3.)* of Lemma 4.1.2 that if the global model parameter $\theta_i, S(i) \subset \mathcal{M}_v, S(i) \not\subset \mathcal{B}_v$ is equal to zero, then $\theta_i^{\mathcal{M}_v}$ is also equal to zero. Following the work on Gaussian graphical models by Meng et al. (2014), it is natural to consider the following graphical model relaxation of the $\mathcal{M}_v$-marginal model.

Let $\mathcal{M}_{l,v}$ denote the relaxed hierarchical log-linear model obtained from the $\mathcal{M}_v$-marginal model by keeping interactions given by edges with at least one endpoint in $\mathcal{M}_v \setminus \mathcal{B}_v$ and all interactions in the power set $2^{\mathcal{B}_v}$. The convex relaxation of the marginal model is illustrated with a 4×4 grid graph in Figure 4.1. The parameter set of the one-hop marginal model for variables $X_{\mathcal{M}_{1,v}}$ is $\theta^{\mathcal{M}_{1,v}} = \{\theta_v, \theta_{va}, \theta_{vb}, \theta_{ab}\}$, and the parameter set of the two-hop marginal model is $\theta = \{\theta_v, \theta_a, \theta_b, \theta_{va}, \theta_{vb}, \theta_{ac}, \theta_{ad}, \theta_{bd}, \theta_{be}, \theta_c, \theta_d, \theta_e, \theta_{cd}, \theta_{de}, \theta_{ce}, \theta_{cde}\}$. The index l takes values $l = 1$ or $l = 2$ when $\mathcal{M}_v$ is respectively the one-hop or two-hop neighborhood of v.

The J-set of this local model is

$$J^{\mathcal{M}_{l,v}} = \{j \in J \mid S(j) \subset \mathcal{M}_v, S(j) \not\subset \mathcal{B}_v\} \cup \{i \in I \mid S(i) \subset \mathcal{B}_v\} \ . \tag{4.2.1}$$

Let $p^{\mathcal{M}_{l,v}}(X_{\mathcal{M}_v})$ denote the marginal probability of $X_{\mathcal{M}_v}$ in the $\mathcal{M}_{l,v}$-marginal model. The local estimates of $\theta_j, j \in \{j \in J | \ S(j) \subset \mathcal{M}_v, \ S(j) \not\subset \mathcal{B}_v\}$ are obtained by maximizing the $\mathcal{M}_{l,v}$-marginal log likelihood

$$L^{\mathcal{M}_{l,v}}(\theta) = \prod_{k=1}^{N} p^{\mathcal{M}_{l,v}}(X_{\mathcal{M}_v} = i_{\mathcal{M}_v}^{(k)}) = \exp\{\langle \theta^{\mathcal{M}_{l,v}}, t^{\mathcal{M}_{l,v}} \rangle - N k^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}})\} \tag{4.2.2}$$

which is a convex maximization problem in

$$\theta^{\mathcal{M}_{l,v}} = (\theta_j, j \in J^{\mathcal{M}_{l,v}}).$$

At this point, we need to make two important remarks.

Remark 4.2.1. *The vector $\theta^{v,PS}$ defined in* (4.1.5) *is a sub vector of $\theta^{\mathcal{M}_{l,v}}$. Therefore maximizing* (4.2.2) *for either $l = 1$ or $l = 2$ will yield an estimate of $\theta^{v,PS}$.*

Remark 4.2.2. *The $\mathcal{M}_{l,v}$-marginal model, $l = 1, 2$, is a hierarchical log-linear model but not necessarily a graphical model. For example, if we consider a four-neighbour lattice and a given vertex v_0 and its four neighbours that we will call $1, 2, 3, 4$ for now, then the generating set of the relaxed $\mathcal{M}_{1,v_0}$-marginal model is*

$$\Delta^{\mathcal{M}_{1,v_0}} = \{(v_0, 1), (v_0, 2), (v_0, 3), (v_0, 4), (1, 2, 3, 4)\}.$$

This is not a discrete graphical model since a graphical model would also include the interactions $(v_0, 1, 2), (v_0, 2, 3), (v_0, 3, 4), (v_0, 1, 4), (v_0, 1, 2, 3, 4)$. It was therefore crucial to set up our problem as we did it in Section 2, within the framework of hierarchical log-linear models rather than the more restrictive class of discrete graphical models.

4.3 Equality of the maximal conditional and marginal composite likelihood estimate

Let $\hat{\theta}^{\mathcal{M}_{l,v}}, l = 1, 2$ denote the maximum likelihood estimate of $\theta^{\mathcal{M}_{l,v}}$ obtained from the local marginal likelihood (4.2.2).

Theorem 4.3.1. *The "PS" component of $\hat{\theta}^{\mathcal{M}_{1,v}}$,i.e. $(\hat{\theta}_j^{\mathcal{M}_{1,v}}, j \in J^{v,PS})$ is equal to the maximum likelihood estimate of $\theta^{v,PS}$ obtained from the local conditional likelihood* (4.1.2).
Similarly, The PS_2 component of $\hat{\theta}^{\mathcal{M}_{2,v}}$,i.e. $(\hat{\theta}_j^{\mathcal{M}_{2,v}}, j \in J^{v,PS_2})$ is equal to the maximum likelihood estimate of θ^{v,PS_2} obtained from the local conditional likelihood (4.1.6).

The proof is given in Appendix B.3. At this point, we ought to make an important observation. In the case of the two-hop marginal likelihood, it can happen that the buffer $\mathcal{B}_v$ may not be equal

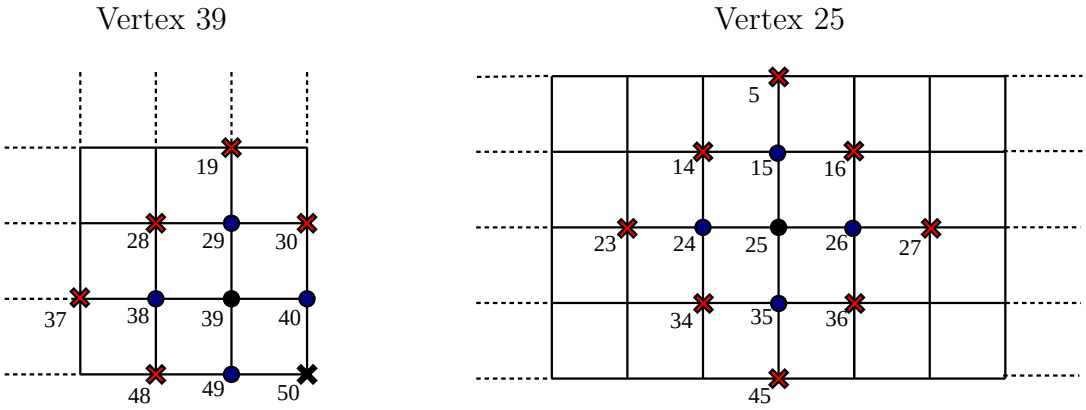


Figure 4.2: Two vertices in a 5×10 lattice: Theorem 4.3.1 applies to vertex 25 but not vertex 39

to $\mathcal{N}_{2v}$. For example, if we consider a four-neighbour 5×10 lattice and number the vertices by rows starting from the left, vertex 39 is such that $\mathcal{N}_{2v} = \{19, 28, 30, 37, 48, 50\}$ while $\mathcal{B}_v = \mathcal{N}_{2v} \setminus \{50\}$. The argument in the proof of Theorem 4.3.1 for j such that $S(j) \not\subset \mathcal{N}_{2v}$ then breaks down since in the $\mathcal{M}_{2,v}$-marginal model, some cells such as $i_{\mathcal{M}_v} = (i_{30} = 1, i_{50} = 1, 0_{\mathcal{M}_v \setminus \{30,50\}})$, with support in $\mathcal{N}_{2v}$ no longer have a complete support. This situation is illustrated in Figure 4.2 where for the sake of comparison, we also look at vertex 25 for which $\mathcal{N}_{2v} = \mathcal{B}_v$ and so Theorem 4.3.1 applies.

In Tables 4.1 and 4.2 we give the numerical values of the maximum likelihood estimate $\theta_j, j \in J^{\mathcal{M}_{2,v}}$ obtained from the four local models $PS, PS_2, \mathcal{M}_{1,v}$ and $\mathcal{M}_{2,v}$ for j such that $j \in J^{PS_{25}}$ and for j such that $j \in J^{PS_{39}}$, respectively. We see that in the first case, the values of $\hat{\theta}_j$ obtained from the local likelihoods $l^{PS_{25}}$ and $l^{\mathcal{M}_{1,25}}$ are identical and similarly for those obtained from $l^{PS_{2,25}}$ and $l^{\mathcal{M}_{2,25}}$, while in the second case, the values obtained from the PS_2 and $\mathcal{M}_{2,v}$ models are slightly different. The values obtained from the PS and $\mathcal{M}_{1,v}$ models are identical since then $\mathcal{B}_v = \mathcal{N}_v$ and the proof of Theorem 4.3.1 does not break down.

Remark 4.3.1. *The equality of the estimates holds also for the marginal estimates obtained by*

Models	$\hat{\theta}_{25}$	$\hat{\theta}_{15,25}$	$\hat{\theta}_{24,25}$	$\hat{\theta}_{25,26}$	$\hat{\theta}_{25,35}$
$\mathcal{M}_{1,v}$	-0.0536	0.5914	-0.4808	-0.8314	-0.8461
$\mathcal{M}_{2,v}$	-0.0779	0.5221	-0.5310	-0.7274	-0.7459
(v, PS)	-0.0536	0.5914	-0.4808	-0.8314	-0.8461
$(v, 2PS)$	-0.0779	0.5221	-0.5310	-0.7274	-0.7459

Table 4.1: The local MLE of some $\theta_j, j \in J^{25,PS}$ in the 5×10 lattice

Models	$\hat{\theta}_{39}$	$\hat{\theta}_{29,39}$	$\hat{\theta}_{38,39}$	$\hat{\theta}_{39,40}$	$\hat{\theta}_{39,49}$
$\mathcal{M}_{1,v}$	-1.0799	-0.3306	-0.3647	-0.5791	1.1749
$\mathcal{M}_{2,v}$	-1.0386	-0.3519	-0.5020	-0.5445	1.1946
(v, PS)	-1.0799	-0.3306	-0.3647	-0.5791	1.1749
$(v, 2PS)$	-1.0381	-0.3531	-0.5019	-0.5448	1.1947

Table 4.2: The local MLE of some $\theta_j, j \in J^{39,PS}$ in the 5×10 lattice

Mizrahi et al. (2014) if, for q a clique of G and $v \in q \subset \mathcal{A}_q$, satisfying the strong LAP condition with respect to $\mathcal{A}_q$, we retain only the parameters $\theta_j, j \in J^{PS_v} \cap q$. We also note that Theorem 9 in that paper may not be true in some cases. For example, take vertex 7 in a 3×3 lattice numbered from left to right starting with the top row, take $q = \{7, 8\}$ as the clique of interest. Then $\mathcal{A}_q = \{4, 7, 8\}$ satisfies the strong LAP condition, but θ_8 in the $\mathcal{A}_q$-marginal model cannot be equal to θ_8 in the joint model as our Lemma 4.1.2 shows.

4.4 Computational complexity of the local marginal and conditional methods

In order to illustrate the algorithms and computational complexity of MLE computation of our local marginal models and local conditional models, we use the Ising model with binary data as an example.

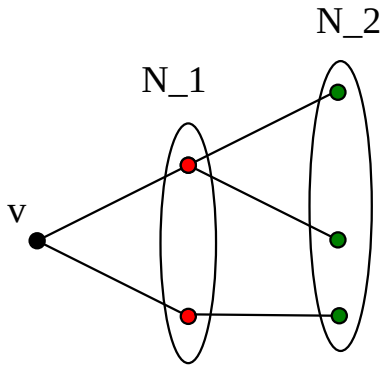


Figure 4.3: A small example for one-hop and two-hop local models

The graph above illustrates the one-hop and two-hop local models of node v. We assume each node takes binary values $\{0, 1\}$. Here we use N_1 to denote the one-hop neighbours of v, N_2 - the neighbours of neighbours of node v, and $N_1 \cup N_2$ - the two-hop neighbours of v. Let $p = |N_1|$ and $q = |N_2|$, so $p + q = |N_1 \cup N_2|$.

4.4.1 One-hop Local Conditional Model

In the one-hop conditional models, the probability density function of X_v given its 1-hop neighbours X_{N_1} is

$$f(x_v|x_{N_1}, \theta) = \frac{\exp(x_v\theta_v + x_v x_{N_1}\theta_{v,N_1})}{1 + \exp(\theta_v + x_{N_1}\theta_{v,N_1})},$$

where θ_v is a scale value, $\theta_{v,\mathcal{N}_1} \in R^p$, so the number of parameters in the function is $p+1$. Given N sample points, we can write the negative pseudo log-likelihood function as follows:

$$l(\theta) = \sum_{i=1}^{N} [\log(1 + \exp(\theta_v + x^i_{N_1}\theta_{v,N_1})) - x^i_v\theta_v - x^i_v x^i_{N_1}\theta_{v,N_1}],$$

We use the limited-memory BFGS algorithm found in the Matlab package "minFunc" of Schmidt (2005) to compute the pseudo-likelihood estimates for each local conditional model. One can refer to Nocedal (1980) and Schmidt et al. (2009) for the details about the algorithm. The BFGS algorithm approximates Newton's method. We don't need to evaluate the Hessian matrix, but the gradient of the log-likelihood is necessary. The gradient can be computed as follows:

$$\begin{aligned} \frac{dl(\theta)}{d\theta_v} &= \sum_{i=1}^{N} [\frac{\exp(\theta_v + x^i_{N_v}\theta_{v,N_1})}{1+\exp(\theta_v + x^i_{N_1}\theta_{v,N_v})} - x^i_v] \\ \frac{dl(\theta)}{d\theta_{v,N_1}} &= \sum_{i=1}^{N} [\frac{\exp(\theta_v + x^i_{N_1}\theta_{v,N_1})}{1+\exp(\theta_v + x^i_{N_1}\theta_{v,N_1})} - x^i_v x^i_{N_1}] \end{aligned}$$

The cost for evaluating the negative log-likelihood function and its gradient is linear to the number of parameters times sample size: $\mathcal{O}(N(p+1))$. As shown in Nocedal (1980) and Schmidt et al. (2009), the cost per iteration of L-BFGS method is $\mathcal{O}(m(p+1))$, where m is a small constant chosen by user, and $p+1$ is the number of parameters in the log-likelihood function. In order to reach an accuracy of ϵ under standard assumptions, one needs $\mathcal{O}(\log(1/\epsilon))$ iterations. Therefore, the total cost for computing the MLE of a 1-hop local conditional model is $\mathcal{O}(\log(1/\epsilon[(m+N)(p+1)])$, which is linear to the number of parameters.

4.4.2 Two-hop Local Conditional Model

In the 2-hop local conditional model as shown in the previous example, there are some node parameters: $\theta_v \in R$, $\theta_{N_1} \in R^p$ and some edge parameters $\theta_{(v,N_1)} = \{\theta_{ij} | i = v, j \in \mathcal{N}_1\} \in R^p$, $\theta_{(N_1,N_2)} = \{\theta_{ij} | i \in \mathcal{N}_1, j \in \mathcal{N}_2\} \in R^q$. The parameter set is therefore $\Theta = \{\theta_v, \theta_{N_1}, \theta_{(v,N_1)}, \theta_{(N_1,N_2)}\} \in$

$R^{(1+2p+q)}$. The probability density function of $X_{v\cup N_1}$ given X_{N_2} is

$$f(x_{v\cup N_1}|x_{N_2},\theta) = \frac{\exp(x_v\theta_v + x_{N_1}\theta_{N_1} + x_v x_{N_1}\theta_{(v,N_1)} + x_{N_1}x_{N_2}\theta_{(N_1,N_2)})}{\sum_{x_{v\cup\mathcal{N}_1}\in I_{v\cup\mathcal{N}_1}} \exp(x_v\theta_v + x_{N_1}\theta_{N_1} + x_v x_{N_1}\theta_{(v,N_1)} + x_{N_1}x_{N_2}\theta_{(N_1,N_2)})},$$

Given N sample points, we can write the negative log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} \big[\log(\sum_{x_{v\cup\mathcal{N}_1}\in I_{v\cup\mathcal{N}_1}} \exp(x_v\theta_v + x_{N_1}\theta_{N_1} + x_v x_{N_1}\theta_{(v,N_1)} + x_{N_1}x^i_{N_2}\theta_{(N_1,N_2)}))$$
$$- (x^i_v\theta_v + x^i_{N_1}\theta_{N_1} + x^i_v x^i_{N_1}\theta_{(v,N_1)} + x^i_{N_1}x^i_{N_2}\theta_{(N_1,N_2)})\big] \quad (4.4.1)$$

We use the same algorithm to compute the MLE as we did in the 1-hop local conditional model. Evaluating the negative log-likelihood function is, however, much more complex. The cost for computing the logarithm in the log-likelihood function is exponential to the size of $v\cup N_1$: $\mathcal{O}(2^{p+1})$. Since we need to compute this logarithm in the negative log-likelihood function and the gradient function, the cost for one data point will be $\mathcal{O}((1+2p+q)2^{p+1})$, and $\mathcal{O}(N(1+2p+q)2^{p+1})$ for N sample points. Similar to the 1-hop case, the total cost for computing the MLE of a 2-hop local conditional model is $\mathcal{O}(\log(1/\epsilon[m(1+2p+q)+N(1+2p+q)2^{p+1}])$, which is exponential in the size of $v\cup\mathcal{N}_1$, or $\mathcal{M}_1$.

4.4.3 One-hop Local Marginal Model

Recall that when we complete the buffer set of each local marginal model, the number of parameters increases exponentially with the number of nodes in the buffer set, but we only increase one clique in each local marginal model. Therefore, using the IPF algorithm designed by Jirousek and Preucil (1995) to compute the MLE of the local marginal model turns out to be much more effective than maximizing the likelihood function. After we get the expected value of the marginal contingency table, we can apply formula (2.15) provided in Letac et al. (2012) to get the MLE of

nature parameters θ:

$$\theta_j = \sum_{j' \lhd j} (-1)^{|S(j)-S(j')|} \log \frac{p(j')}{p(0)}$$

We don't need to compute all the parameters in the local marginal model, since we just need the parameters $\{\theta_j, v \in S(j)\}$. In our example we just need to compute $\theta_v,\ \theta_{(v,N_1)}$, which costs $\mathcal{O}(p+1)$.

Recall that $\mathcal{M}_1 = v \cup \mathcal{N}_1$, and $I_{\mathcal{M}_v}$ denote the set of cells in the $\mathcal{M}_1$-marginal contingency table, in the one-hop local marginal model of node v, so we have

$$|I_{\mathcal{M}_1}| = 2^{p+1}.$$

We need to update all the cell counts in the $\mathcal{M}_1$-marginal contingency table. Therefore the total cost for the IPF algorithm is $\mathcal{O}(2^{p+1})$, which is exponential to $|\mathcal{M}_1|$.

4.4.4 Two-hop Local Marginal Model

The two-hop local marginal model is almost the same as the one-hop, except that the two-hop local marginal model has $1 + p + q$ nodes and more cliques. In our experiments, we choose to use the IPF algorithm to get the expected values of the two-hop marginal contingency table, and then computed canonical parameters $\theta_v,\ \theta_{(v,N_1)}$. The computational complexity is exponential to the number of nodes in the local marginal model: $|\mathcal{M}_2|$.

We took advantage of Matlab's matrix computation prowess to avoid multiple "for-loops". This allows us to update the contingency table m at a high speed, and the computational time to grow linearly with the number of cliques.

4.5 The maximum composite likelihood estimate

Since we have proved that the estimates of $\theta^{v,PS}$ obtained from local conditional and relaxed marginal likelihoods are identical, and computational complexity in the relaxed marginal model, we work only with the local estimates obtained from local conditional likelihoods. More precisely, for each local conditional likelihood $l^{v,PS}$ or l^{v,PS_2}, we consider the local maximum likelihood estimate $\hat{\theta}^{v,PS}$ or $\hat{\theta}^{v,PS_2}$. We define

$$\hat{\theta}^v = \begin{cases} \hat{\theta}^{v,PS} & \text{if we work with } l^{v,PS} \\ (\hat{\theta}_j^{v,PS_2},\ S(j) \subset \{v\} \cup \mathcal{N}_v) & \text{if we work with } l^{v,PS_2}\ . \end{cases} \tag{4.5.1}$$

In other words, from either $l^{v,PS}$ or l^{v,PS_2}, we retain $\hat{\theta}^v = (\hat{\theta}_j^v,\ S(j) \subset (\{v\} \cup \mathcal{N}_v) \setminus \mathcal{N}_v) = (\hat{\theta}_j^v,\ , v \in S(j))$ only. If we have m_j estimates $\hat{\theta}_j^{v_l}, l = 1, \ldots, m_j$, then we define the maximum composite likelihood estimate of θ to be

$$\bar{\theta} = (\bar{\theta}_j = \frac{\sum_{l=1}^{m_j} \hat{\theta}_j^{v_l}}{m_j},\ j \in J), \tag{4.5.2}$$

Let $\hat{\theta}^{PS}$ denote the vector obtained by stacking up the vectors $\hat{\theta}^v, v \in V$. We then have

$$\bar{\theta} = A\hat{\theta}^{PS}$$

where A is a $|J| \times \sum_{v \in V} |J^{v,PS}|$ where $J^{v,PS}$ is as defined in (4.1.5). If $S(j) = \{v\}$, then clearly, the row of A corresponding to $\bar{\theta}_j$ has all its entries equal to zero except for one entry equal to one in the column block $J^{v,PS}$. If $j \in J^{v_l,PS},\ l = 1, \ldots, m_j$, and $S(j) \subset (\{v_l\} \cup \mathcal{N}_{v_l}) \setminus \mathcal{N}_{v_l}$ the row corresponding to $\bar{\theta}_j$ has all its entries equal to zero except for one entry equal to $\frac{1}{m_j}$ in each of the column blocks $J^{v_l,PS},\ l = 1, \ldots, m_j$. For example, if the model considered is the discrete graphical model Markov

with respect to the four-cycle with vertex set $V = \{a, b, c, d\}$ and $\mathcal{D} = \{ab, ac, bd, cd\}$, we have

$$\bar{\theta} = \begin{pmatrix} \bar{\theta}_a \\ \bar{\theta}_{ab} \\ \bar{\theta}_b \\ \bar{\theta}_{bd} \\ \bar{\theta}_c \\ \bar{\theta}_{cd} \\ \bar{\theta}_d \\ \bar{\theta}_{db} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 \end{pmatrix} \begin{pmatrix} \hat{\theta}^a_a \\ \hat{\theta}^a_{ab} \\ \hat{\theta}^a_{ac} \\ \hat{\theta}^b_b \\ \hat{\theta}^b_{ab} \\ \hat{\theta}^b_{bd} \\ \hat{\theta}^c_c \\ \hat{\theta}^c_{ca} \\ \hat{\theta}^c_{cd} \\ \hat{\theta}^d_d \\ \hat{\theta}^d_{bd} \\ \hat{\theta}^d_{cd} \end{pmatrix}.$$

In general, for $j \in J$ and $k \in J^{v,PS},\ v \in V$, the matrix A is defined by

$$A_{j,k} = \begin{cases} \frac{1}{m_j} & \text{if} \quad j_{v_l \cup \mathcal{N}_{v_l}} = k \in J^{v_l,PS},\ l = 1, \ldots, m_j \\ 0 & \text{otherwise.} \end{cases} \tag{4.5.3}$$

We have now defined our MCLE which we use to replace the global MLE maximizing (4.1.1). It is natural to ask whether the MCLE exists when the global MLE exists, and conversely, whether the global MLE exists when the MCLE exists. The existence of the global MLE is an important problem that has been considered in Fienberg and Rinaldo (2012) and more recently in Wang et al. (2016). We say that the MLE does not exist if we cannot find $\hat{\theta}$ such that the corresponding cell probabilities $p(i)$ and $p(0)$ as given by (2.1.4) and (2.1.5) are strictly positive. The nonexistence of the global MLE has important consequences for inference. However, if we are only concerned with

estimation of the parameter θ or equivalently with $(p(i), i \in I)$, as the following lemma shows, the global MLE may not exist, but we may still accept the MCLE as an estimate of the parameter.

Lemma 4.5.1. *For a discrete log-linear model, if the global MLE exists, then the MCLE exists. but the converse is not necessarily true.*

Proof: If the global MLE exists, then $\hat{p}(X = i) > 0$ and $\hat{p}(X_{\mathcal{N}_v} = i_{\mathcal{N}_v}) > 0$,

$$\hat{p}(X_v = i_v | X_{\mathcal{N}_v} = i_{\mathcal{N}_v}) = \frac{\hat{p}(X = i)}{\hat{p}(X_{\mathcal{N}_v} = i_{\mathcal{N}_v})} > 0,$$

i.e. the composite MLE exists. We now give an example where the MCLE exists but the global MLE does not. Consider the four-cycle graphical model as described above, with binary variables.

Let the data be such that $n(i) = 1, i \in \{0000, 1000, 0100, 1010, 0101, 1011, 0111, 1111\}$ and $n(i) = 0$ otherwise so that the marginal counts are $t_c = t_d = 4, t_{ab} = 1, t_{bd} = t_{cd} = t_{ac} = 3$ where for $A \subset V$, t_A denotes t_j with $j_v = 1$ if $v \in A$ and $j_v = 0$ otherwise. Thus the data vector lies on the facet $t_c + t_d + t_{ab} - t_{bd} - t_{cd} - t_{ac} = 0$ of the marginal polytope of the four-cycle model. The reader is referred to Letac et al. (2012, Theorem 5.3) for the equations of the facets of the polytope corresponding to the four-cycle. From the theory on the existence of the global maximum likelihood estimate developed in Fienberg and Rinaldo (2012) and references therein, it can be concluded that the global MLE does not exist in this case. The facets corresponding to the local models built on $v = a$ have equations

$$\begin{aligned} t_{ab} &= 0; \\ t_a - t_{ab} &= 0; \\ t_b - t_{ab} &= 0; \\ 1 - t_a - t_b + t_{ab} &= 0; \end{aligned}$$

We can verify immediately that none of these equations are satisfied with the given data and

therefore the MLE of $\theta^{v,PS}$ in the a-local model. Similarly the MLE of $\theta^{v,PS}, v = b, c, d$ exists and thus the MCLE exists. $\square$

5 Asymptotic properties of the maximum composite likelihood estimate

In this chapter, we look at the asymptotic properties of the MCLE $\bar{\theta}$ when p is fixed and then when both p and N go to infinity. Though asymptotics when p is fixed have been given by Liu and Ihler (2012), we give our result here in Section 5.1 for completeness in our own notation.

5.1 The classical asymptotic regime

We consider here the behaviour of the MCLE $\bar{\theta}$ when $p = |V|$ is fixed and the sample size N goes to infinity. We have the following result.

Theorem 5.1.1. *The MCLE $\bar{\theta}$ as defined in* (4.5.2) *is asymptotically consistent and*

$$\sqrt{N}(\bar{\theta} - \theta^*) \to N(0, AGA^t) \tag{5.1.1}$$

where A is as defined in (4.5.3)*, G is the square $\sum_{v\in V} |J^{v,PS}|$-dimensional matrix with (v_l, v_m)-block entry*

$$G_{v_l,v_m} = I^{-1}(\theta^{*v_l,*})E(\frac{\partial l(\theta^{*v_l})}{\partial \theta^{*v_l}} \left(\frac{\partial l(\theta^{*v_m})}{\partial \theta^{*v_m}}\right)^t)I^{-1}(\theta^{*v_m}), \tag{5.1.2}$$

$l(\theta^{*v_l}) = l^{v_l,PS}((\theta^*)^{v_l,PS}|X)$ is the local conditional likelihood, given one sample point X, evaluated at the true local parameter $(\theta^*)^{v_l,PS}$ and $I(\theta^{*v_l}) = E(\frac{\partial l(\theta^{*v_l})}{\partial \theta^{*v_l}} \left(\frac{\partial l(\theta^{*v_l})}{\partial \theta^{*v_l}}\right)^t)$ is the v_l-local information matrix evaluated at the true value θ^{*v_l}, $v_l \in V$.

The mean square error therefore satisfies

$$NE(\|\bar{\theta}_j - \theta_j^*\|^2) \xrightarrow{N\to\infty} \sum_{l=1}^{m_j} \frac{1}{m_j^2}[I^{vl}(\theta^{vl,*})]^{-1}_{j,j} + \sum_{l_1=1}^{m_j}\sum_{l_2=l_1+1}^{m_j} \frac{2}{m_j^2}[G_{v_{l_1},v_{l_2}}]_{j,j} \tag{5.1.3}$$

The proof is given in Appendix B.4. In the expression of the mean square error (5.1.3) above, we note that to the diagonal elements of the inverse information matrix for each local model are added the cross-product terms $[G_{v_{l_1},v_{l_2}}]_{j,j}$, because the estimates of $\hat{\theta}_j^v$ coming from the v_{l_1} and v_{l_2} local conditional models with $j \in J^{v_{l_1},PS} \cap J^{v_{l_2},PS}$ are not independent. We also note here that our Theorem above coincides with Theorem 4.1 in Liu and Ihler (2012) with our matrix A being equal to their $(\sum_i W^i)^{-1}$.

To illustrate our result above, we simulate data from the 4-cycle graphical model. We simulate our data for the following values of the parameters

$$[\theta_a, \theta_b, \theta_c, \theta_d, \theta_{ab}, \theta_{ac}, \theta_{bd}, \theta_{cd}] = [0.53, 1.83, -2.25, 0.86, 0.31, -1.30, -0.43, 0.34].$$

The results are illustrated in Figure 5.1.

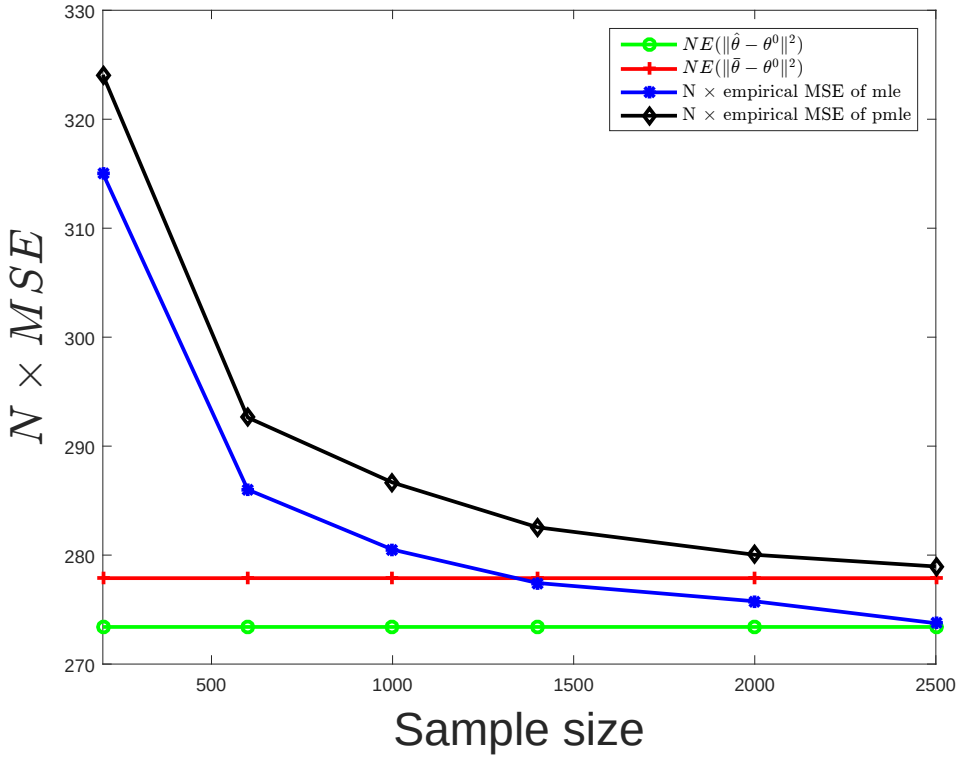


Figure 5.1: Empirical and theoretical mean square errors for the global MLE and the MCLE of the parameters for the four-cycle graphical model.

We now examine the asymptotic variance of the two estimates $\theta_j^{\mathcal{M}_{i,v}}$, $i = 1, 2, j \in J, S(j) \subset \mathcal{M}_v, S(j) \not\subset \mathcal{B}_v$. We distinguish between the buffer set of the relaxed $\mathcal{M}_{1,v}-$marginal model, and that of the $\mathcal{M}_{2,v}-$ marginal model, and denote them be $\mathcal{B}_{i,v}$, where $i = 1, 2$ respectively. We will

use the notation

$$\begin{aligned} J_{i,v} &= \{j \in I_{\mathcal{M}_{i,v}} : j \in J, S(j) \subset \mathcal{M}_{i,v}, S(j) \not\subset \mathcal{B}_{i,v}\} \subset J^{\mathcal{M}_{i,v}} \\ B_{i,v} &= \{j \in I_{\mathcal{M}_{i,v}} : S(j) \subset \mathcal{B}_{i,v}\} \\ \theta_{J_{i,v}} &= \{\theta_j, j \in J_{i,v}\} \\ \theta_{B_{i,v}} &= \{\theta_j, j \in B_{i,v}\} \end{aligned} \tag{5.1.4}$$

We consider the following four models that are defined by their J-sets $\mathcal{J}$:

1. the relaxed one-hop marginal model $\mathcal{M}_{1,v}$ with J-set equal to $\mathcal{J} = J_{1,v} \cup B_{1,v}$,

2. the relaxed two-hop marginal model $\mathcal{M}_{2,}$ with J-set equal to $\mathcal{J} = J_{2,v} \cup B_{2,v}$,

3. the overall model with J-set $\mathcal{J} = J$,

4. a new augmented marginal model, denoted $\bar{\mathcal{M}}_{2,v}$ that we will use in the argument below with J-set equal to $\mathcal{J} = J_{1,v} \cup B_{1,v} \cup J_{2\setminus 1,v} \cup B_{2,v}$, where $J_{2\setminus 1,v} = J_{2,v} \setminus J_{1,v}$.

We note that the density of the four models is of the general form (4.1.8) with $\theta = (\theta_j, j \in \mathcal{J})$ and with cumulative generating functions

$$\begin{aligned} k^{\mathcal{M}_{i,v}}(\theta^{\mathcal{M}_{i,v}}) &= \log(\textstyle\sum_{k \in I_{\mathcal{M}_{i,v}}} \exp \sum_{j \lhd k, j \in \mathcal{J}} \theta_j) \\ k^{J}(\theta) &= \log \textstyle\sum_{i \in I} \exp \sum_{j \lhd i, j \in \mathcal{J}} \theta_j \\ k^{\bar{\mathcal{M}}_{2,v}}(\theta^{\bar{\mathcal{M}}_{2,v}}) &= \log \textstyle\sum_{k \in I_{\mathcal{M}_v}} \exp \sum_{j \lhd k, j \in \mathcal{J}} \theta_j) \end{aligned}$$

for the models $\mathcal{M}_{i,v}, i = 1, 2$, the overall model and the augmented marginal model $\bar{\mathcal{M}}_{2,v}$ respectively and where the set $\mathcal{J}$ changes accordingly.

Whatever the model is, the symmetric matrix of the covariance of t is the $\mathcal{J} \times \mathcal{J}$ matrix

$$\frac{\partial^2 k(\theta)}{\partial \theta^2} = (\frac{\partial^2 k(\theta)}{\partial \theta_j \partial \theta_{j'}})_{j,j' \in \mathcal{J}} = (p_{j \cup j'} - p_j p_{j'})_{j,j' \in \mathcal{J}}$$

where we use the notation $j \cup j^{'}$ to denote the cell $i \in I_{\mathcal{M}_{i,v}}$ or $i \in I$ with support $j \cup j^{'}$ and

$$p_{j\cup j'} = p((j \cup j^{'})_{S(j\cup j')}),\ p_j = p(j_{S(j)})$$

denote the marginal probabilities. For $j, j^{'}$ given, since $p_{j\cup j'}, p_j, p_{j'}$ are marginal probabilities, the entries $p_{j\cup j'} - p_j p_{j'}$ are the same for all models with $j, j^{'} \in \mathcal{J}$. We will now give the following result concerning the variance of the estimates.

Theorem 5.1.2. *For any parameter $\theta_j,\ j \in J$, we can find a vertex $v \in V$ such that $v \in S(j)$. Let $\hat{\theta}_j^{\mathcal{M}_{1,v}}$, $\hat{\theta}_j^{\mathcal{M}_{2,v}}$ be the estimates obtained from maximizing* (4.2.2)*, the v-th component of the one-hop and two-hop marginal likelihoods respectively. Let $\hat{\theta}_j$ be the MLE obtained from maximizing the original likelihood function* (2.2.3)*, then we have*

$$var(\hat{\theta}_j^{\mathcal{M}_{1,v}}) \geq var(\hat{\theta}_j^{\mathcal{M}_{2,v}}) \geq var(\hat{\theta}_j). \tag{5.1.5}$$

The proof is given in Appendix B.5.

5.2 The double asymptotic regime

In this section, we consider the asymptotic properties of the MCLE when both p and N go to $+\infty$. In Theorem 5.2.1 below, we give its rate of convergence to the true value θ^*. In order to compare the behaviour of the MCLE with the global MLE, we also give, in Theorem 5.2.2, the rate of convergence of the global MLE under the same asymptotic regime.

It will be convenient to introduce the notation

$$f_j(x) = \prod_{l\in S(j)} \mathbb{1}(x_l = j_l) = \begin{cases} 1 & \text{if } j \lhd x \\ 0 & \text{otherwise} \end{cases},$$

and to write (4.1.3) as

$$p(x_v|x_{\mathcal{N}_v}) = \frac{\exp\{\sum_{j\in J^{v,PS}} \theta_j f_j(x_v, x_{\mathcal{N}_v})\}}{1 + \sum_{y_v\in I_v\setminus\{0\}} \exp\{\sum_{j\in J^{v,PS}} \theta_j f_j(y_v, x_{\mathcal{N}_v})\}}\ . \tag{5.2.1}$$

In this section, we work exclusively with $l^{v,PS}(\theta^{v,PS})$. Therefore for simplicity of notation we write θ for $\theta^{v,PS}$. Also, for convenience, we scale the log likelihood by the factor $\frac{1}{N}$. Then the v-local conditional log likelihood function is

$$\begin{aligned} l^{v,PS}(\theta) &= \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\log p(x_v^{(n)}|x_{\mathcal{N}_v}^{(n)}) \\ &= \textstyle\sum_{j\in J^{v,PS}}\theta_j\frac{1}{N}\sum_{n=1}^{N} f_j(x_v^{(n)}, x_{\mathcal{N}_v}^{(n)}) \\ &- \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\log\{1+\sum_{y_v\in I_v\setminus\{0\}}\exp\{\sum_{j\in J^{v,PS}}\theta_j f_j(y_v, x_{\mathcal{N}_v}^{(n)})\}\} \end{aligned}$$

The sufficient statistic is $t_j = \frac{1}{N}\sum_{n=1}^{N} f_j(x_v^{(n)}, x_{\mathcal{N}_v}^{(n)})$. We write

$$t_{J^{v,PS}} = [t_1, t_2, \cdots, t_{d_v}] \tag{5.2.2}$$

and

$$k^{v,PS}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\log\{1+\sum_{y_v\in I_v\setminus\{0\}}\exp\{\sum_{j\in J^{v,PS}}\theta_j f_j(y_v, x_{N_v}^{(n)})\}\} = \frac{1}{N}\sum_{n=1}^{N}\log Z^{n,v}(\theta),$$

where

$$Z^{n,v}(\theta) = 1+\sum_{y_v\in I_v\setminus\{0\}}\exp\{\sum_{j\in J^{v,PS}}\theta_j f_j(y_v, x_{\mathcal{N}_v}^{(n)})\}.$$

Then the log likelihood function is

$$l^{v,PS}(\theta) = \sum_{j\in J^{v,PS}}\theta_j t_j - k^{v,PS}(\theta)\ .$$

Its first derivative is

$$\begin{aligned} \frac{\partial l^{v,PS}(\theta)}{\partial\theta_k} &= t_k - \frac{\partial k^{v,PS}(\theta)}{\partial\theta_k}, \\ \frac{\partial k^{v,PS}(\theta)}{\partial\theta_k} &= \frac{1}{N}\sum_{n=1}^{N}\frac{\exp\{\sum_{j\in J^{v,PS}}\theta_j f_j(k_v, x_{\mathcal{N}_v}^{(n)})\}}{Z^{n,v}(\theta)} f_k(k_v, x_{\mathcal{N}_v}^{(n)}) \end{aligned}$$

with

$$\frac{\exp\{\sum_{j\in J^{v,PS}}\theta_j f_j(k_v, x_{\mathcal{N}_v}^{(n)})\}}{Z^{n,v}(\theta)} = p(X_v = k_v|x_{\mathcal{N}_v}^{(n)}) \tag{5.2.3}$$

We now want to compute $\frac{\partial^2 l^{v,PS}(\theta)}{\partial\theta_k\partial\theta_l} = -\frac{\partial^2 k^{v,PS}(\theta)}{\partial\theta_k\partial\theta_l}$, $k, l \in J^{v,PS}$. To simplify further our notation, we set

$$z_{y_v}(\theta) = \sum_{j\in J^{v,PS}} \theta_j f_j(y_v, x^{(n)}_{\mathcal{N}_v}). \tag{5.2.4}$$

For $k_v = l_v$, using (5.2.3), we obtain

$$\begin{aligned}\frac{\partial^2 k^{v,PS}(\theta)}{\partial\theta_k\partial\theta_l} &= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)} - (\frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)})^2\right) f_k(k_v, x^{(n)}_{\mathcal{N}_v}) f_l(l_v, x^{(n)}_{\mathcal{N}_v}) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(p(X_v = k_v|x^{(n)}_{\mathcal{N}_v}) - p(X_v = k_v|x^{(n)}_{\mathcal{N}_v})^2) f_k(k_v, x^{(n)}_{\mathcal{N}_v}) f_l(l_v, x^{(n)}_{\mathcal{N}_v}\right) .\end{aligned}$$

if $k_v \neq l_v$, then

$$\begin{aligned}\frac{\partial^2 k^{v,PS}(\theta)}{\partial\theta_k\partial\theta_l} &= \frac{1}{N}\sum_{n=1}^{N} -\frac{\exp z_{k_v}(\theta)\exp z_{l_v}(\theta)}{(Z^{n,v}(\theta))^2} f_k(k_v, x^{(n)}_{\mathcal{N}_v}) f_l(l_v, x^{(n)}_{\mathcal{N}_v}) \\ &= \frac{1}{N}\sum_{n=1}^{N}(-p(X_v = k_v|x^{(n)}_{\mathcal{N}_v})p(X_v = l_v|x^{(n)}_{\mathcal{N}_v})) f_k(k_v, x^{(n)}_{\mathcal{N}_v}) f_l(l_v, x^{(n)}_{\mathcal{N}_v}) .\end{aligned}$$

Let $W^{n,v} = (f_j(j_v, x^{(n)}_{\mathcal{N}_v}), j \in J^{v,PS})$ be the $d_v \times 1$ vector of indicators. We introduce the notation

$$\eta^{n,v}_{k,l}(\theta, x^{(n)}_{\mathcal{N}_v}) = \begin{cases} \frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)} - (\frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)})^2, & \text{if } k_v = l_v \\ -\frac{\exp z_{k_v}(\theta)\exp z_{l_v}(\theta)}{(Z^{n,v}(\theta))^2}, & \text{if } k_v \neq l_v . \end{cases} \tag{5.2.5}$$

Let $H^{n,v}(\theta, x^{(n)}_{\mathcal{N}_v})$ be the $d_v \times d_v$ matrix with (k, l) entry $\eta^{n,v}_{k,l}(\theta, x^{(n)}_{\mathcal{N}_v})$. Then the Fisher information matrix derived from $l^{v,PS}$ is

$$(k^{v,PS})''(\theta) = \frac{1}{N}\sum_{n=1}^{N} H^{n,v}(\theta, x^{(n)}_{\mathcal{N}_v}) \circ [W^{n,v}(W^{n,v})^t]$$

where $\circ$ denotes the Hadamard product of two matrices. We make two assumptions regarding the behaviour of the cumulative generating function $k^{v,PS}, v \in V$ at θ^*, similar to those made by Ravikumar et al. (2010) and Meng (2014).

(A) For the design matrix of the v-local conditional models, we assume that there exists $D_{max} > 0$ such that

$$\max_{v\in V} \lambda_{max}\Big(\frac{1}{N}\sum_{n=1}^{N} W^{n,v}(W^{n,v})^t\Big) \leq D_{max};$$

(B) We assume the minimum eigenvalue of the Fisher Information matrices $(k^{v,PS})''(\theta^*),\ v \in V$ is bounded, i.e., there exists $C_{min} > 0$ such that

$$C_{min} = \min_{v \in V} \lambda_{min} \frac{1}{N} \sum_{n=1}^{N} \left[H^{n,v}(\theta^*, x^{(n)}_{\mathcal{N}_v}) \circ [W^{n,v}(W^{n,v})^t] \right].$$

We are now ready to state our theorem on the asymptotic behaviour of $\bar{\theta}$.

Theorem 5.2.1. *Assume conditions (A) and (B) hold. If the sample size N and $|V| = p$ satisfy*

$$\frac{N}{\log p} \geq \max_{v \in V} \left(\frac{10CD_{max}d_v}{C^2_{min}}\right)^2,$$

where C is a positive constant such that $p^{\frac{C^2}{2}} \geq 2|J|$, then the MCLE $\bar{\theta} = (\bar{\theta}_j, j \in J)$ is such that

$$\|\bar{\theta} - \theta^*\|_F \leq \frac{5C}{C_{min}} \sqrt{\frac{\sum_{v \in V} d_v \log p}{N}} \tag{5.2.6}$$

with probability greater than $1 - \frac{2|J|}{p^{\frac{C^2}{2}}}$.

The proof is given in Appendix B.6. With a similar argument, we can derive the behaviour of the global MLE, which we will denote by $\hat{\theta}^G$. We need to make assumptions similar to (A) and (B). We assume that

$$(A')\ \text{ there exists }\ D_{max} > 0\ \text{ such that }\ \lambda_{max}\Big(\sum_{i \in I} f_i \otimes f_i\Big) \leq D_{max},$$

$$(B')\ \ 0 < \kappa^* = \lambda_{min}\Big[k''(\theta^*)\Big].$$

The asymptotic behaviour of $\hat{\theta}^G$ is given in the following theorem.

Theorem 5.2.2. *Assume conditions (A') and (B') hold. If N and p satisfy the condition*

$$\frac{N}{\log p} \geq \left(\frac{40C|J|D_{max}}{\kappa^{*2}}\right)^2,$$

where C is a positive constant such that $p^{2C^2} \geq 2|J|$, then the global MLE $\hat{\theta}^G = (\hat{\theta}^G_j, j \in J)$ is such that

$$\|\hat{\theta}^G - \theta^*\|_F \leq \frac{5C}{\kappa^*} \sqrt{\frac{|J|\ \log p}{N}} \tag{5.2.7}$$

with probability greater than $1 - \frac{2|J|}{p^{2C^2}}$.

The proof is given in Appendix B.7. Comparing Theorems 5.2.1 and 5.2.2, we see that for $\frac{N}{\log(p)} = \mathcal{O}(|J|^2)$, $\|\hat{\theta}^G - \theta^*\|_F = \mathcal{O}(\sqrt{\frac{|J|\ \log p}{N}})$ with high probability while for $\frac{N}{\log(p)} = \mathcal{O}(\max_{v\in V}(d_v^2))$, $\|\hat{\theta} - \theta^*\|_F = \mathcal{O}(\sqrt{\frac{\sum_{v\in V} d_v \log p}{N}})$. This implies that for the MCLE, the requirement on the sample size N is not as stringent as for the global MLE but of course, we lose some accuracy in the approximation of θ^*. The situation is, however, not bad since

$$\sqrt{\frac{\sum_{v\in V} d_v \log p}{N}} \Big/ \sqrt{\frac{|J|\ \log p}{N}} = \sqrt{\frac{\sum_{v\in V} d_v}{|J|}}$$

which is the square root of the ratio of the sum over $v \in V$ of the number of parameters in the v-local conditional models and the number of parameters in the global model. If the number of neighbours for each vertex is bounded by d, we see that this ratio is at most equal to $\frac{2^{d+1}}{|J|}$ and usually much smaller than that. For example, in an Ising model, $|J| = p + |E|$ and $\sum_{v\in V} d_v = p + 2|E|$ and therefore $\frac{\sum_{v\in V} d_v}{|J|} = 1 + \frac{|E|}{p+|E|} \leq 2$.

6 Existence of MLE in hierarchical log-linear models

In this section, we will study the second problem: the existence of the MLE in hierarchical log-linear models. We fix a discrete exponential family $\mathcal{E}_A$. While our main interest lies in hierarchical models, the results that we need are more naturally formulated in the language of discrete exponential families. We assume that a vector of observed counts $n = (n(i) : i \in I)$ is given. The log-likelihood function of parameters $\theta = \{\theta_j, j \in J\}$ is

$$l(\theta|n) = \langle \theta, t \rangle - Nk(\theta),$$

let $\hat{\theta}$ be the MLE of θ as defined in Definition 2.2.1. The function $l(\theta)$ is always bounded (clearly, it is never positive). As stated above, $l(\theta)$ is strictly concave (if the parameters are identifiable), and so the maximum is unique (up to identifiability), if it exists. However, a maximum need not exist, since the domain of the parameters θ is unbounded. To understand this, it is convenient to interpret the likelihood as a function of probabilities. Let l' be the function that assigns to any probability distribution p on I the value

$$l'(p) = \sum_{i \in I} n(i) \log p(i).$$

Then $l(\theta) = l'(p_\theta)$, and $\hat{\theta}$ is the MLE if and only if $p_{\hat{\theta}}$ maximizes the log-likelihood function $l'(p)$ subject to the constraint that p belongs to the hierarchical model, and thus that it is of the form p_θ for some θ. While the set of all probability distributions on I is compact, the hierarchical model itself

is not closed and therefore not compact, and so there is no guarantee that $\tilde{l}$ attains its maximum on the hierarchical model. However, things become better when we pass from the hierarchical model to its topological closure, where the topology comes from interpreting a probability distribution as a vector $p = (p(i))_{i \in I} \in \mathbf{R}^I$ of real numbers (this choice of the topology is canonical since we are dealing with a finite set I; for infinite sample spaces see Csiszár and Matúš (2005)). The closure is sometimes also called *completion* (Barndorff-Nielsen, 2014, p. 154). Since the closure of the hierarchical model is again compact, the continuous function l' always attains its maximum.

Theorem 6.0.1. *The closure of a discrete exponential family can be written as a union*

$$\overline{\mathcal{E}_A} = \bigcup_F \mathcal{E}_{F,A},$$

where F runs over all facial sets of the convex support polytope $\mathbf{P}_A$ and where $\mathcal{E}_{F,A}$ consists of all probability distributions of the form $p_{F,\theta}$, with

$$p_{F,\theta} = \begin{cases} \exp(\langle \theta, f_i \rangle - k_F(\theta)), & \text{if } i \in F, \\ 0, & \text{otherwise}, \end{cases}$$

where $k_F(\theta) = \log \sum_{i \in F} \exp(\langle \theta, f_i \rangle$.

Proof. See Barndorff-Nielsen (2014). For self-containedness we provide a proof in our notation in Appendix B.8. □

Theorem 6.0.1 shows that $\overline{\mathcal{E}_A}$ is a finite union of sets $\mathcal{E}_{F,A}$ that are exponential families themselves with a very similar parametrization, using the same number of parameters and the same design matrix A (or, rather, the submatrix A_F consisting of those columns of A indexed by F). However, for any proper facial set F, the parametrization $\theta \mapsto p_{F,\theta}$ is not injective, i.e. the parameters θ are not identifiable on $\mathcal{E}_{F,\Delta}$. The reason is that the matrix $\tilde{A}_F$ does not have full rank, even if $\tilde{A}$ has full rank, since all columns of $\tilde{A}_F$ lie on a supporting hyperplane defining F.

A second thing to note is that although the parameters θ on $\mathcal{E}_A$ and the parameters θ on $\mathcal{E}_{F,A}$ play similar roles, they are very different in the following sense: If $\theta^{(s)}$ is a sequence of parameters with $p_{\theta^{(s)}} \to p_{F,\theta}$ for some θ, then, in general, $\lim_{s\to\infty} \theta_j^{(s)} \neq \theta_j$ for all $j \in J$.

Theorem 6.0.2. *For any vector of observed counts n, there is a unique maximum p^* of $\tilde{l}$ in $\overline{\mathcal{E}_A}$. For t as defined in* (4.1.8)*, this maximum p^* satisfies:*

- $Ap^* = \frac{t}{N}$.
- $supp(p^*) = F_t$.

Proof. See Barndorff-Nielsen (2014). For self-containedness we provide a proof in our notation in Appendix B.9. □

Definition 6.0.3. *The maximum in Theorem 6.0.2 is called the* extended *maximum likelihood estimate (EMLE).*

Clearly, if the MLE θ^* exists, then $p^* = p_{\theta^*}$.

6.1 Faces of the marginal polytope $\mathbf{P}_\Delta$

As we showed in Lemma 3.2.2, the problem of determining the existence of MLE in hierarchical log-linear models is equivalent to finding the face of the marginal polytope $\mathbf{P}$ containing the sufficient statistics t. Recall that I_+ denotes the cells with positive cell counts in a contingency table, and I_0 denotes the empty cells, so we have the following lemma.

Lemma 6.1.1. *The sufficient statistics t belongs to a face $\mathbf{F}$ of marginal polytope $\mathbf{P}$, if and only if $f_i \in \mathbf{F}, \forall i \in I_+$.*

Proof.

$$t = \sum_{i \in I} \frac{n(i)}{N} f_i = \sum_{i \in I_+} \frac{n(i)}{N} f_i$$

$t \in \mathbf{F} \iff \langle t, g \rangle = 0 \iff \sum_{i \in I_+} \frac{n(i)}{N} \langle f_i, g \rangle = 0.$

$\langle f_i, g \rangle \geq 0 \ \forall i \in I$, so

$$\sum_{i \in I_+} \frac{n(i)}{N} \langle f_i, g \rangle = 0 \iff \langle f_i, g \rangle = 0, \ \forall i \in I_+$$

□

Let A, a $|I| \times |J|$ matrix be the design matrix of the hierarchical log-linear model generated by Δ, A_+ be the sub-matrix with rows indexed by the positive cells I_+ and A_0 as the sub-matrix indexed by the empty cells I_0. We give an algorithm to compute the smallest face or facial set containing sufficient statistics t in the following lemma.

Lemma 6.1.2. *Solution g^* of the non-linear problem*

$$\begin{aligned} \max \quad z = \quad & \|Ag\|_0 \\ s.t. \ & A_+ g = 0 \\ & A_0 g \geqslant 0 \end{aligned} \tag{6.1.1}$$

is a perpendicular vector to the smallest face containing t. The corresponding facial set is $F_t = I \setminus supp(Ag^*)$, where "supp" means the support of a vector.

Any vector g that belongs to the feasible set of problem (6.1.1) defines a face in the marginal polytope $\mathbf{P}_A$, we maximize the l_0 norm $\|Ag\|_0$ so that we get the smallest facial set F_t. The optimization problem (6.1.1) is highly non-linear and non-convex, but it can be solved by repeatedly

solving the associated ℓ_1-norm optimization problem:

$$\begin{aligned} \max \ & z = \|A_0 g\|_1 \\ s.t. \quad & A_+ g = 0 \\ & A_0 g \geq 0 \\ & A_0 g \leq 1 \end{aligned} \tag{6.1.2}$$

Problem (6.1.2) is a linear programming problem: we iterate until we get the smallest facial set F_t. The process is as follows:

Algorithm 1 Face computation using a linear programming method

Require: Design matrix A and positive cell index I_+

INITIALIZE $A_+ = A(I_+, :)$, $A_0 = A \setminus A_+$

Solve problem 6.1.2, get the solution g^* and the corresponding maximum z^*

while $A_0 \neq \emptyset$ and $z^* \neq 0$ **do**

Let matrix B be the submatrix of A_0, by taking columns of A_0 which satisfy $\langle f_i, g^* \rangle > 0$, update

$A_0 = A_0 \setminus B$,

Solve problem 6.1.2, get the solution g^* and the corresponding maximum z^*

end while

if $A_0 = \emptyset$ **then**

$F_t = I_+$

end if

if $Z^* = 0$ **then**

$F_t = I_+ \cup \{i | i \text{ is the index of } A_0\}$

end if

Now we are going to prove that we can solve the l_0 optimization problem (6.1.1), by implementing

the LP problem (6.1.2) repeatedly. The equivalent statement is as follows:

Theorem 6.1.3. *Assuming we get* max $z = 0$ *after repeatedly solving the linear programming problem* (6.1.2) K *times, let* $g_1, g_2, \cdots, g_K$ *be the corresponding optimization solutions. Denote* $A_0^{(1)}, A_0^{(2)}, \cdots, A_0^{(K)}$ *the new matrix in the cost function of each LP problem* (6.1.2)*, so we have* max $z = \|A_0^{(K)} g\|_1 = 0$. *Then* $g = \sum_{k=1}^{K} g_k$ *is the optimization solution of problem* (6.1.1).

Proof. Suppose g is not the optimization solution of (6.1.1), so there exists another vector g^* belonging to the feasible set in (6.1.1), such that there exists at least one row f_m in matrix A_0 satisfying

$$\langle f_m,\ g\rangle = 0; \quad \langle f_m,\ g^*\rangle > 0.$$

$\langle f_m,\ g\rangle = 0 \Longrightarrow \langle f_m,\ g_k\rangle = 0, \quad k = 1, 2, \cdots, K. \Longrightarrow f_m$ is still a row in matrix $A_0^{(K)}$.

Then $\langle f_m,\ g^*\rangle > 0$ is a contradiction to

$$\text{max } z = \|A_0^{(K)} g\|_1 = 0$$

□

Another similar way to compute F_t is to solve the $|I|$ linear programming problems:

$$\begin{aligned} max \quad & z_i =< f_i, c > \\ s.t. \quad & < t, c >= 0 \\ & Ac \geq 0 \\ & |c| \leq 1. \end{aligned} \tag{6.1.3}$$

Then

$$F_t = \{i | z_i = 0\}.$$

This algorithm is less efficient due to the large number of cells in I. This said, the fact that there is no communication among the $|I|$ linear programming problems allows us to solve the $|I|$

linear programming problems using the distributed computing. The algorithm is introduced in the supplementary material of Fienberg and Rinaldo (2012), where it is also proved that it outputs the correct result.

7 Approximations to the faces of the marginal polytope

The linear programming algorithm 1 works pretty well in low-dimensional contingency tables, but if the dimension p is very large, the number of rows of design matrix A is exponential in p, so we won't have enough memory or computing power to solve (6.1.2). Our simulations show that when $p > 16$ and each variable takes binary values, we cannot solve (6.1.2) anymore. We use local models to approximate facial sets in high-dimensional tables.

We consider a hierarchical model with simplicial complex Δ and marginal polytope $\mathbf{P}_\Delta$. In this section, we explain the details of our methodology for obtaining an inner and an outer approximation to the facial set F_t of the smallest face $\mathbf{F}_t$ of $\mathbf{P}_\Delta$ containing the data vector t. Our main tool is Lemma 7.0.1. For any $S \subseteq I$, we abbreviate the facial set $F_{\mathbf{P}_\Delta}(S)$ by $F_\Delta(S)$.

Lemma 7.0.1. *Let Δ and Δ' be simplicial complexes on the same vertex set with $\Delta' \subseteq \Delta$, and denote by f_i, f_i' $(i \in I)$ the rows of the design matrices of the corresponding hierarchical models. There exists a linear map $\phi : \mathbf{R}^h \to \mathbf{R}^{h'}$ such that $\phi(f_i) = f_i'$. In fact, ϕ is a coordinate projection. In particular, the marginal polytope $\mathbf{P}_{\Delta'}$ is a coordinate projection of $\mathbf{P}_\Delta$. Thus, for any $S \subseteq I$, we have $F_\Delta(S) \subseteq F_{\Delta'}(S)$*

Proof. The design matrix A_Δ has one column for each parameter θ_j, $j \in J_\Delta$. Removing sets from Δ leads to a smaller set $J_{\Delta'}$ and thus leads to a matrix $A_{\Delta'}$ with less rows. The definition of each row that remains does not change. The lemma now clearly follows from Lemma 2.3.5. □

Next we discuss marginal polytopes of decomposable (or reducible) models. Then, in Sections 7.2 and 7.3, we explain how to use Lemma 7.0.1 to obtain inner and outer approximations to $F_\Delta(S)$.

7.1 Decomposable models

Definition 7.1.1. *Let $V' \subset V$. The* restriction, *or* induced sub-complex *is $\Delta|_{V'} = \{S \in \Delta \mid S \subseteq V'\}$. The sub-complex $\Delta|_{V'}$ is* complete, *if $\Delta|_{V'}$ contains V' (and thus all subsets of V'). For brevity, in this case we say that V' is* complete *in Δ.*

Definition 7.1.2. *A subset $S \subset V$ is a* separator *of Δ if there exist $V_1, V_2 \subset V$ with $V_1 \cap V_2 = S$, $\Delta = \Delta|_{V_1} \cup \Delta|_{V_2}$ and $V_1 \neq S \neq V_2$. A simplicial complex that has a complete separator is called* reducible. *By extension, we also call the hierarchical model reducible.*

Definition 7.1.3. *A hierarchical model is* decomposable *if Δ can be written as a union $\Delta = \Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_r$ of induced sub-complexes $\Delta_i = \Delta|_{V_i}$ in such a way that*

1. *each Δ_i is a complete simplex: $\Delta_i = \{S \subseteq V_i\}$; and*
2. *$(\Delta_1 \cup \cdots \cup \Delta_i) \cap \Delta_{i+1}$ is a complete simplex.*

In other words, Δ arises by iteratively gluing simplices along complete sub-simplices.

The faces of a reducible hierarchical model are combinations of the faces of its two parts:

Proposition 7.1.4 (Eriksson et al. (2006))**.** *Suppose that Δ has a complete separator S that separates V into V_1 and V_2. Each face of $\mathbf{P}_{\Delta|_{V_1}}$ corresponds to an inequality*

$$\sum_{j \in J_{\Delta|_{V_1}}} g_j^{(1)} t_j \geq c_1.$$

The same inequality also defines a face of $\mathbf{P}_\Delta$*. Similarly, each face of* $\mathbf{P}_{\Delta|_{V_2}}$ *defines a face of* $\mathbf{P}_\Delta$*. Each face of* $\mathbf{P}_\Delta$ *either arises in this way, or it is the intersection of two such faces, one induced by* $\mathbf{P}_{\Delta|_{V_1}}$ *and one induced by* $\mathbf{P}_{\Delta|_{V_2}}$*.*

Proof. See Eriksson et al. (2006), Lemma 8. □

In the sequel, for any $V' \subseteq V$ and $i \in I = \prod_{v \in V} I_v$, it will be convenient to use the seemingly more complicated notation $\pi_{V'}(i) = (i_v, v \in V')$ for the marginal cell $i_{V'} \in I_{V'} := \prod_{v \in V'} I_v$. Similarly, for a set $S \subseteq I$, the restriction to V' is $\pi_{V'}(S) := \{\pi_{V'}(i) : i \in S\}$. For $T \subset I_{V'}$, the opposite action yields $\pi_{V'}^{-1}(T) = \{i \in I \mid i_{V'} \in T\}$.

We next translate Proposition 7.1.4 to the language of facial sets:

Lemma 7.1.5. *Suppose that* Δ *has a complete separator* S *that separates* V *into* V_1 *and* V_2*.*

1. *If* $F \subseteq I$ *is facial with respect to* Δ*, then* $\pi_{V_1}(F)$ *and* $\pi_{V_2}(F)$ *are facial with respect to* $\Delta|_{V_1}$ *and* $\Delta|_{V_2}$*.*

2. *Conversely, if* $F_1 \subseteq I_{V_1}$ *and* $F_2 \subseteq I_{V_2}$ *are facial with respect to* $\Delta|_{V_1}$ *and* $\Delta|_{V_2}$*, then* $\pi_{V_1}^{-1}(F_1) \cap \pi_{V_2}^{-1}(F_2)$ *is facial with respect to* Δ*.*

Thus, for any $T \subseteq I$*, let* $T_1 = \pi_{V_1}(T)$ *and* $T_2 = \pi_{V_2}(T)$*.*

$$F_\Delta(T) = \pi_{V_1}^{-1}(F_{\Delta|_{V_1}}(T_1)) \cap \pi_{V_2}^{-1}(F_{\Delta|_{V_2}}(T_2)).$$

Proof. Consider an inequality as in Proposition 7.1.4 that defines a face $\mathbf{F}$ of $\mathbf{P}_\Delta$ as well as a face $\mathbf{F}_1$ of $\mathbf{P}_{\Delta_1}$. Then the corresponding facial sets F and F_1 satisfy $F = \pi_{V_1}^{-1}(F_1)$; in order to check whether some f_i, $i \in I$, satisfies the inequality, we only need to look at the components involving V_1; that is, we only need to look at $\pi_{V_1}(i)$. □

Lemma 7.1.5 easily generalizes to more than one separator and thus to more than two components and it becomes particularly simple when these components are complete. Indeed, in that case, $F_{\Delta|_{V_1}}(T_1) = T_1$ and taking the preimage we obtain

$$\pi_{V_1}^{-1}(\pi_{V_1}(T)) = \{i \in I : \exists i' \in T \text{ such that } \pi_{V_1}(i) = \pi_{V_1}(i')\} \supseteq T.$$

The following lemma is an immediate consequence of Lemma 7.1.5.

Lemma 7.1.6. *Let Δ be a decomposable model with decomposition $\Delta = \Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_r$ where Δ_i is a complete simplex on V_i, and let $\pi_i = \pi_{V_i}$ be the corresponding marginalization map. Then, for any $T \subseteq I$,*

$$F_\Delta(T) = \pi_1^{-1}(\pi_1(T)) \cap \pi_2^{-1}(\pi_2(T)) \cap \cdots \cap \pi_r^{-1}(\pi_r(T)).$$

7.2 Inner approximations

To obtain an inner approximation, our strategy is to find a separator S of Δ and to complete it. More specifically, we augment Δ by adding all subsets of S. The result is a simplicial complex $\Delta_S = \Delta \cup \{M : M \subseteq S\}$ in which S is a complete separator. We can apply Lemma 7.1.5 to find the facial set $F_{\Delta_S}(I_+)$, and this will be our inner approximation of $F_\Delta(I)$.

An even simpler approximation is obtained by not only completing the separator itself, but also the two parts V_1, V_2 separated by S: The simplicial complex $\Delta_{V_1,V_2} := \{M : M \subseteq V_1\} \cup \{M : M \subseteq V_2\}$ is decomposable and contains Δ. Its facial sets can be computed from Lemma 7.1.6.

In general, the approximation obtained from a single separator (or, in general, a single supercomplex) is not good; that is, $F_t = F_\Delta(I_+)$ tends to be much larger than $F_{\Delta_S}(I_+)$ or $F_{\Delta_{V_1,V_2}}(I_+)$. Thus we need to combine information from several separators. For example, given two separa-

tors $S, S' \subseteq V$, we find a chain of approximations

$$G'_0 := I_+,$$

$$G_1 := F_{\Delta_S}(G'_0),\ G'_1 := F_{\Delta_{S'}}(G_1),$$

$$G_2 := F_{\Delta_S}(G'_1),\ G'_2 := F_{\Delta_{S'}}(G_2),$$

$$\vdots$$

that satisfy

$$I_+ \subseteq G_1 \subseteq G'_1 \subseteq G_2 \subseteq \cdots \subseteq F_t,$$

where all inclusions except the last one are due to the definition of $F_{\Delta_S}(T)$ or $F_{\Delta_{S'}}(T)$ as the smallest facial sets containing T in Δ_S or $\Delta_{S'}$. The last inclusion is a consequence of Lemma 7.0.1 since both Δ_S and $\Delta_{S'}$ contain Δ.

This chain of approximations has to stabilize at a certain point; that is, after a certain number of iterations, the approximations will not improve any more. The limit, which we denote by $F_{S,S'}(I^+) := \bigcup_i G_i = \bigcup_i G'_i$, can be characterized as the smallest subset of I that contains I^+ and is facial both with respect to Δ_S and $\Delta_{S'}$. The same iteration can be done replacing Δ_S and $\Delta_{S'}$ by Δ_{V_1,V_2} and $\Delta_{V'_1,V'_2}$. Applying in turn $F_{\Delta_{V_1,V_2}}$ and $F_{\Delta_{V'_1,V'_2}}$ gives another approximation $\tilde{F}_{S,S'}(I^+)$, namely the smallest subset of I that contains I^+ and is facial both with respect to Δ_{V_1,V_2} and $\Delta_{V'_1,V'_2}$. This latter approximation will be used in Section 10.1.1. Since $\tilde{F}_{S,S'}(I^+) \subseteq F_{S,S'}(I^+) \subseteq F_t$, $\tilde{F}_{S,S'}(I^+)$ is a worse approximation than $F_{S,S'}(I^+)$; it is, however, easier to compute.

We use the following strategies:

1 If possible, use all graph separators.

There are two problems with this strategy: First, if S is such that either V_1 or V_2 is large, then it is almost as difficult to compute $F_{\Delta|_{V_1}}$ and $F_{\Delta|_{V_2}}$, as $F_{\Delta|_V}$. Such "bad" separators always exist:

namely, each node $i \in V$ is separated by its neighbours from all other nodes. In this case, V_1 consists of i and its neighbours, and V_2 consists of $V \setminus \{i\}$. For such a "bad" separator we can only compute $F_{\Delta_{V_1,V_2}}$, but not F_{Δ_S}. Second, the number of separators may be large. Since we have to iterate over this set until the approximation converges, it may take a long time to compute the inner approximation. A faster alternative strategy is the following:

2 Look at separators such that both $V_1 \setminus S$ and $V_2 \setminus S$ are not too small (for example, $\min\{|V_1 \setminus S|, |V_1 \setminus S|\} \geq 3$).

We illustrate the first strategy in Section 10.1.2, using a graphical model associated with the NLTCS data set. In the case of the grids studied in Sections 10.1.1 and 10.2.2, which display a lot of regularity, we use an adapted strategy:

3 In a grid, use the horizontal, vertical and diagonal separators.

In the case of grids, the vertical separators form a family of pairwise disjoint separators. In Section 10.2 we show how we can make use of such a family to study faces of hierarchical models, even when the facial sets are so large that they become computationally intractable.

7.3 Outer approximations

According to Lemma 7.0.1, when we compute $F_{\Delta'}(S)$ for a simplicial complex $\Delta' \subseteq \Delta$, we obtain an outer approximation of $F_\Delta(S)$. Removing sets from Δ decreases the dimension of the marginal polytope, so it is often easier to compute $F_{\Delta'}(S)$ than to compute $F_\Delta(S)$. Our main strategy is to look at subcomplexes induced by subset $V' \subset V$.

Let $\Delta_{V'}$ be the simplicial complex induced by V'. Let $J \subset I$ be its set of interactions. When comparing Δ with $\Delta|_{V'}$, we have to be precise about whether we consider $\Delta|_{V'}$ as a simplex on V

or on V': when we consider it on V, Let A be the $I \times J$ design matrix with rows f_i, $i \in I$, when we consider it on V', the design matrix A' is an $I_{V'} \times J$ matrix with columns $f'_{i'}, i' \in I_{V'}$. Because we have the same set of interactions whether we are on V or V', for $i \in I$ and $i' \in I_{V'}$, we have:

$$f_i = f'_{i'} \Leftrightarrow i \in \pi_{V'}^{-1}(i'). \tag{7.3.1}$$

Therefore the marginal polytopes of the two models are the same since they are the convex hull of the same set of vectors $\{f_i, i \in I\} = \{f'_{i'}, i' \in I_{V'}\}$. The relationship between the facial sets on V and V' is as follows:

Lemma 7.3.1. *Let $V' \subseteq V$. For $K \subset I$, we have*

$$F_{\Delta|_{V'}}(K) = \pi_{V'}^{-1}(F'_{\Delta|_{V'}}(\pi_{V'}(K))).$$

Here, $F'_{\Delta|_{V'}}$ denotes the facial set when $\Delta_{V'}$ is considered as a simplicial complex on V', and $F_{\Delta|_{V'}}$ denotes the facial set when $\Delta_{V'}$ is considered as a simplicial complex on V.

Proof. For $K \subset I$, the two sets $\mathcal{A} = \{a_i, i \in K\}$ and $\mathcal{B} = \{b_{i'}, i' \in \pi_{V'}(K)\}$ are identical and therefore the smallest faces of the marginal polytopes for $\Delta_{V'}$ on V or V' containing $\mathcal{A}$ and $\mathcal{B}$ respectively are the same.

From the definition of $F'_{\Delta_{V'}}(\pi_{V'}(K))$, we know that the smallest face containing $\mathcal{B}$ is defined by $\{b_{i'}, i' \in F'_{\Delta_{V'}}(\pi_{V'}(K))\}$. From the definition of $F_{\Delta_{V'}}(K)$, the smallest face containing $\mathcal{A}$ is $\{a_i, i \in F_{\Delta_{V'}}(K)\}$. Also, from the equation (7.3.1), we have that $\{a_i,\ i \in \pi_{V'}^{-1}(F'_{\Delta_{V'}}(\pi_{V'}(K)))\} = \{b_{i'},\ i' \in F'_{\Delta_{V'}}(\pi_{V'}(K))\}$. Therefore, $F_{\Delta_{V'}}(K) = \pi_{V'}^{-1}(F'_{\Delta_{V'}}(\pi_{V'}(K)))$. □

In general, $F_{\Delta|_{V'}}(I_+)$ is not a good approximation of $F_\Delta(I_+)$. We can improve this approximation by considering several subsets of V. To be precise, if $V_1, \ldots, V_r \subseteq V$, then $F_\Delta(I_+) \subseteq F_{\Delta|_{V_i}}(I_+)$ for $i = 1, \ldots, r$, and thus $F_\Delta(I_+) \subseteq \bigcap_{i=1}^r F_{\Delta|_{V_i}}(I_+) =: F_{V_1,\ldots,V_r;\Delta}(I_+)$.

The question is now how to choose the subsets V_i. Clearly, the subsets V_i should cover V, and, more precisely, they should cover Δ, in the sense that for any $D \in \Delta$ there should be one V_i with $D \subseteq V_i$. The larger the sets V_i, the better the approximation becomes, but the more difficult it is to compute $F_{V_1,\dots,V_r;\Delta}(I_+)$.

One generic strategy is the following:

1. Use all subsets of V of fixed cardinality k plus all facets $D \in \Delta$ with $|D| \geq k$.

This choice of subsets indeed covers Δ. The parameter k should be chosen as large as possible such that computing $F_{V_1,\dots,V_r;\Delta}(I_+)$ is still feasible. Note that computing $F_{\Delta|_D}(I_+)$ for $D \in \Delta$ is trivial, since $\mathbf{P}_{\Delta|_D}$ is a simplex.

Another natural strategy first described in Massam and Wang (2015) is the following:

2. For fixed k, use balls $B_k(v) = \{w : d(v,w) \leq k\}$ around the nodes $v \in V$, where $d(\cdot,\cdot)$ denotes the edge distance in the graph.

In general, we choose subsets V_i to be large enough to preserve some of the structure of Δ. For example, for the grid graphs, we suggest the use of 3×3-subgrids. These graphs have two nice properties: first, they already have the appearance of a small grid, second, for any vertex $v \in V$, there is a 3×3 sub-grid that contains v and all neighbours of v. We will compare two different strategies:

3. For a grid, use all 3×3-subgrids.

4. Cover a grid by 3×3-subgrids.

In Section 10.2.2 we compare these two methods, and we observe that in the case of the 5×10 grid, it suffices to only look at a covering. In general, it is not enough to look at induced sub-

complexes, unless Δ has a complete separator (see Section 7.1). The approximation tends to be good nevertheless and gives the correct facial set in many cases.

7.4 Comparing the two approximations

Suppose that we have computed two approximations F_1, F_2 of F_t such that $F_1 \subseteq F_t \subseteq F_2$. If we are in the lucky case when $F_1 = F_2$, then we know that $F_t = F_1 = F_2$. In general, the cardinality of $F_2 \setminus F_1$ indicates the quality of our approximations.

F_1, F_2 and F_t can also be compared by the ranks of the matrices $\tilde{A}_{F_1}$, $\tilde{A}_{F_2}$ and $\tilde{A}_{F_t}$ obtained from $\tilde{A}$ by keeping only the columns indexed by F_1, F_2 and F_t, respectively. Clearly, $rank\tilde{A}_{F_1} \leq rank\tilde{A}_{F_t} \leq rank\tilde{A}_{F_2}$. Note that $rank\tilde{A}_{F_2}$ equals the dimension of the corresponding face $\mathbf{F}_2$ of $\mathbf{P}$, and $rank\tilde{A}_{F_t}$ equals the dimension of $\mathbf{F}_t$. But F_1 does not necessarily correspond to a face of $\mathbf{P}$. Nevertheless, we can bound the codimension of $\mathbf{F}_t$ in $\mathbf{F}_2$ by

$$\dim \mathbf{F}_2 - \dim \mathbf{F}_t \leq rankA_{F_2} - rankA_{F_1}.$$

In particular, if $rankA_{F_2} = rankA_{F_1}$, then we know that $F_t = F_2$. In this case, our approximations give us a precise answer, even if $F_1 \neq F_2$ and the lower approximation F_1 is not tight.

8 Statistical inference for the nonexistent MLE

Finding the smallest face containing the data vector t is one of the major accomplishments of our work. This is done, of course, to allow for correct statistical inference.

Given a contingency table, we would like to fit a log-linear model that generates this data. Such a log-linear model can help us understand the data and the relationship among variables. The first step in statistical inference of the hierarchical log-linear model is to estimate log-linear parameters, which will also give us the estimate of the cell probabilities. Next we provide the confidence interval. As a last step, we usually conduct the goodness-of-fit test or likelihood ratio test to see which model fits the given data set better. When the MLE exists, all these tasks can be achieved by traditional methods, which are explained in more detail in Agresti and Kateri (2011) and Bishop et al. (1975a). Whenever the MLE doesn't exist, a common occurence in discrete data analysis, we can't rely on any of the traditional methods, but alternative solutions are provided by Geyer et al. (2009) and Fienberg and Rinaldo (2012).

Now that we have identified the facial set of the smallest face containing t, we want to draw correct inference. We start by offering an identifiable and estimable parametrization in which the linear combinations of the original parameters can be estimated. Second, we use the dimension of the face defined by the facial set F_t to give the correct approximation to the chi-square or G^2 statistics. Confidence intervals in the correct model defined on F_t can then be obtained using

traditional methods.

8.1 Computing the extended MLE

If the MLE θ^* exists, then it can be computed by finding the unique maximum of the log-likelihood function $l(\theta)$ given in (2.2.3). As mentioned before, $l(\theta)$ is concave (or even strictly concave, if parameter θ is identifiable), and thus the maximum is, at least in principle, easy to find (in practice, for larger models, it may be difficult to evaluate the function $k(\theta)$, which involves a sum over I; but we will not discuss this problem here). In general, the maximum cannot be found analytically, but there are efficient numerical algorithms to maximize concave functions. Regular Newton's method or any modification of Newton's method can be used to find the MLE. An example of an algorithm commonly used is *iterative proportional fitting* (IPF), which can be thought of as an algorithm of Gauss-Seidel type.

When the MLE does not exist but the facial set $F = F_t$ of the data is known, then it is straight forward to compute the extended MLE p^*. In this case, we know that p^* lies in $\mathcal{E}_{F,A}$. To find p^*, we need to optimize the log-likelihood $\tilde{l}$ over $\mathcal{E}_{F,A} = \{p_{F,\theta} : \theta \in \mathbf{R}^{|J|}\}$, where J is the dimension of the original model. After plugging the parametrization $p_{F,\theta}$ into $\tilde{l}$, we need to optimize the restricted log-likelihood function

$$l_F(\theta) = \log(\prod_{i \in I_+} p_{F,\theta}(i)^{n(i)}) = \sum_{j \in J} \theta_j t_j - N k_F(\theta). \tag{8.1.1}$$

This problem is of a similar type as the problem to maximize l in the case when the MLE exists, and the same algorithms as discussed above can be used. The problem here is slightly easier, since F is smaller than I. The submatrix A_F from the original design matrix A by taking the rows indexed by cells in facial set F_t becomes the new design matrix of the distributions in the exponential family

$\mathcal{E}_{F,A}$. The original design matrix A is full rank, but A_F is not a full rank matrix as we remove rows that's not indexed by the facial set. As a result, the parametrization $\theta \mapsto p_{F,\theta}$ is not identifiable. Of course, this problem is easy to solve by selecting a set of independent parameters among the θ_j. Depending on the choice of the independent subset, the values of the parameters change, and in particular, it is meaningless to compare the values of the parameters θ_j with parameter values of any other distribution in $\mathcal{E}_A$ or in the closure $\overline{\mathcal{E}_A}$.

Before explaining how to find better parameters on $\mathcal{E}_{F,A}$, let us discuss what happens if the facial set F_t of the data is not known. As mentioned before, whether or not the MLE exists, the log-likelihood function $l(\theta)$ is always strictly concave (assuming that the parametrization is identifiable). When the MLE does not exist, then the maximum is not at a finite value θ^*, but lies "at infinity." Still, as noted in Geyer et al. (2009, Section 3.15), any reasonable version of Newton's method that tries to maximize the likelihood will send θ to infinity in the right direction. Such a numerical algorithm generates a sequence of parameter values $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$ with increasing log-likelihood values $l(\theta^{(1)}) \leq l(\theta^{(2)}) \leq \ldots$. Since $l(\theta)$ is concave, our optimization problem is numerically easy (at least in theory), and for any such reasonable algorithms, the limit $\lim_{s\to\infty} l(\theta^{(s)})$ will equal $\sup_\theta l(\theta) = \max_{p\in\overline{\mathcal{E}_A}} \tilde{l}(p)$. The algorithm stops when the difference $l(\theta^{(s+1)}) - l(\theta^{(s)})$ becomes negotiably small. The output, $\theta^{(s)}$, then gives a good approximation of the EMLE, in the sense that p^* and $p_{\theta^{(s)}}$ are close to each other.

For many applications, such as those found in machine learning, where it is more important to have good parameter values rather than modeling the "true underlying distribution," or when doing a likelihood test, where the value of the likelihood is more important than parameter values, this may be good enough.

However, in this numerical optimization, some of the parameters θ_j will tend to $\pm\infty$, which may

lead to numerical problems. For example, it may happen that one parameter goes to $+\infty$ and a second parameter to $-\infty$ in such a way that their sum remains finite. This implies that a difference between two large numbers has to be computed, however, this is numerically unstable. Also, it is not clear which parameters numerically tend to infinity. In fact, this may depend on the chosen algorithm; i.e. different algorithms may yield approximations of the EMLE that are qualitatively different in the sense that different parameters diverge. We give an example of this in Appendix C.

To avoid such problems, we propose a change of coordinates that allows us to control which parameters diverge, at least in the case where we know the facial set F_t. If we don't know F_t, but we know the approximations $F_1 \subseteq F_t \subseteq F_2$, we can use this knowledge to identify some of thoses parameters that definitely remain finite, and some of those parameters definitely diverge. Although we cannot control the behaviour of the remaining parameters, the more information we have about the facial set F_t, the better control we have of the above mentioned problems.

8.2 An identifiable parametrization

We have seen that when we use the parametrization $\theta \mapsto p_{F_t,\theta}$ of $\mathcal{E}_{A,F_t}$ in the case where $F_t \neq I$, we have to expect the following (interrelated) issues:

1. The parametrization is not identifiable, i.e. there are parameters θ, θ' with $p_{F_t,\theta} = p_{F_t,\theta'}$.

2. While the parametrization $\theta \mapsto p_{F_t,\theta}$ looks similar to the parametrization $\theta \mapsto p_\theta$ of $\mathcal{E}_A$, the values of the parameters in both parametrizations are not related to each other.

3. When $p_{\theta^{(s)}} \to p_{F_t,\theta}$ as $s \to \infty$ for some parameter values $\theta^{(s)}, \theta$, then some of the parameter values $\theta^{(s)}$ diverge to $\pm\infty$. When computing probabilities, there may be linear combinations of these diverging parameters that remain finite.

We can introduce an alternative parameterization of log-linear models as follows,

$$\mu_i(\theta) = \langle \theta, f_i \rangle = \log p(i)/p(0), i \in I. \tag{8.2.1}$$

The parameters μ_i can be interpreted as log odds ratios. Next we show that if F_t is known, then, with a convenient choice of L, the parameters μ_L solve 1 and 2 and improve 3. Afterwards, we discuss what can be done if F_t is not known. We briefly discuss the general solution of 3 in Appendix D. In any case, parameter choose depends on the facial set F_t; i.e. it is not possible to define a single parametrization that works for all facial sets simultaneously.

Suppose that F_t is known. We consider the parameters μ_i as in (8.2.1), and we make sure to choose the zero element 0 in I_+, since $p(0)$ is in the denominator in (8.2.1). The parameters μ_i are not independent, so we need to choose an independent subset L. We do this in two steps:

1. Choose a maximal subset L_t of F_t such that the parameters μ_i, $i \in L_t$ are independent.

2. Then extend L_t to a maximal subset $L \subseteq I$ such that the parameters μ_i, $i \in L$ are independent by adding elements $i \in I \setminus F_t$.

It follows from Theorem 6.0.2 that the following holds:

1. The subset μ_i, $i \in L_t$, of the parameters μ_L gives an identifiable parametrization of $\mathcal{E}_{F_t,A}$.

2. Let μ_i^*, $i \in L_t$, be the parameter values that maximize l_{F_t} (and thus give the EMLE). When the likelihood $l(\mu)$ is maximized numerically on I, then in successive iterations of the maximization, the estimates $\mu_i^{(s)}$ are such that

$$\mu_i^{(s)} \to \begin{cases} \mu_i^*, & i = 1, \ldots, h_t, \\ -\infty, & \text{otherwise.} \end{cases}$$

In particular, no parameter tends to $+\infty$.

The last property ensures a consistency of the parameters μ_i on $\mathcal{E}_A$ and on $\mathcal{E}_{F_t,A}$. This is important in those cases where the parameters have an interpretation, and where it is of interest to know the value of those parameters, which are well-defined. For example, in hierarchical models, the parameters correspond to "interactions" of the random variables, and it may be of interest to know which of these interactions are important, and the size of corresponding parameters. It is usually not parameter μ_i, but the original parameters θ_i that have an interpretation. When we understand parameters μ_i, we can also tell which of parameters θ_i or which combinations of parameters θ_i have finite well-defined values and can be computed, and which parameters diverge:

Lemma 8.2.1. *Suppose that $\theta^{(s)}$, $s \in \mathcal{N}$, are parameter values such that $p_{\theta^{(s)}} \to p^*$ as $s \to \infty$. For any $i \in L_t$, the linear combination*

$$\mu_i^{(s)} = \langle \theta^{(s)}, f_i \rangle$$

has a well-defined finite limit as $s \to \infty$. Any linear combination of the $\theta_i^{(s)}$ that has a well-defined finite limit (that is, a limit that is independent of the choice of the sequence $\theta^{(s)}$) is itself a linear-combination of the $\mu_i^{(s)}$ with $i \in L_t$.

Proof. The first statement follows from $\mu_i^{(s)} = \log p_{\theta^{(s)}}(i)/p_{\theta^{(s)}}(0) \to \log p^*(i)/p^*(0)$. For the second statement, note that any linear combination of the θ is also a linear combination of the μ, since the linear map $\theta \mapsto \mu(\theta)$ is invertible. We now show that if a linear combination $\sum_i a_i \mu_i$ involves some μ_j with $j \notin L_t$, then there exist sequences $\mu^{(s)}$, $\mu'^{(s)}$ of parameters with

$$\lim_{s\to\infty} p_{\mu^{(s)}} = \lim_{s\to\infty} p_{\mu'^{(s)}} \quad \text{and} \quad \lim_{s\to\infty} \sum_i a_i \mu_i^{(s)} \neq \lim_{s\to\infty} \sum_i a_i \mu_i'^{(s)}.$$

So suppose that $\mu^{(s)}$ is a sequence of parameters such that $\lim_{s\to\infty} p_{\mu^{(s)}}$ exists and such that

$\lim_{s\to\infty} \sum_i a_i \mu_i^{(s)}$ is finite. Define

$$\mu_i'^{(s)} = \begin{cases} \mu_j^{(s)} + 1, & \text{if i=j,} \\ \mu_i^{(s)}, & \text{otherwise.} \end{cases}$$

An easy computation shows that

$$\lim_{s\to\infty} p_{\mu'^{(s)}} = \lim_{s\to\infty} p_{\mu^{(s)}} \quad \text{and} \quad \lim_{s\to\infty} \sum_i a_i \mu_i'^{(s)} = \lim_{s\to\infty} \sum_i a_i \mu_i^{(s)} + a_j. \qquad \square$$

Suppose now that we do not know F_t, but that instead we have approximations F_1, F_2 that satisfy

$$I_+ \subseteq F_1 \subseteq F_t \subseteq F_2 \subseteq I.$$

In this case, we proceed as follows to obtain an independent subset L among the parameters μ_i:

1. Choose a maximal subset L_1 of F_1 such that parameters μ_i, $i \in L_1$ are independent.

2. Then extend L_1 to a maximal subset $L_2 \subseteq F_2$ by adding elements $i \in F_2 \setminus F_1$ such that parameters μ_i, $i \in L_2$ remain independent.

3. Finally, extend L_2 to a maximal subset $L \subseteq I$ by adding elements $i \in I \setminus F_2$ such that parameters μ_i, $i \in L$ remain independent

These parameters have the following properties that follow directly from Lemma 8.2.1:

Corollary 8.2.2. *Suppose that $\theta^{(s)}$, $s \in \mathcal{N}$, are parameter values such that $p_{\theta^{(s)}} \to p^*$ as $s \to \infty$, and let $\mu_i^{(s)} = \langle \theta^{(s)}, f_i \rangle$.*

1. *For any $i \in L_1$, the linear combination*

$$\mu_i^{(s)} = \langle \theta, f_i \rangle$$

has a well-defined finite limit as $s \to \infty$. Thus, any linear combination of the $\mu_i^{(s)}$ with $i \in L_1$ has a well-defined limit as $s \to \infty$.

2. *Any linear combination $\sum_i a_i \mu_i^{(s)}$ that has a well-defined limit as $s \to \infty$ is in fact a linear combination of the $\mu_i^{(s)}$ with $i \in L_2$. Thus, a linear combination that involves at least one $\mu_j^{(s)}$ with $j \in L \setminus L_2$ does not have a well-defined limit.*

Now let's have a look at the goodness-of-fit tests of log-linear models when the MLE doesn't exist. As we said in the introduction, the standard regularity conditions for the asymptotic distribution don't hold anymore. The Fisher information matrix of the original likelihood is singular, so the confidence interval of the MLE is not well defined. When the MLE exists, the asymptotic distribution of both the Pearson test and the likelihood ratio test a Chi-square distribution withs the degree of freedom(df) equal to the model's dimension, or the difference between the dimensions of the two compared models. In the non-existent MLE scenario, the asymptotic distribution is still a Chi-square distribution, but the value of the degrees of freedom is different. Suppose we want to compare the performance of two log-linear models $\mathcal{M}_0$ and $\mathcal{M}_1$, the likelihood ratio statistic can be written as

$$G^2 = -2(l_0(\theta_0) - l_1(\theta_1)),$$

where l_0 and l_1 are the log-likelihood functions of the two models respectively. Although we can't get the MLE, we can plugin the extended MLE to get the maximum value of the log-likelihood functions. When it comes to the degrees of freedom of G^2, we can get it from the difference of the dimensions of the two smallest faces containing the sufficient statistics of $\mathcal{M}_0$ and $\mathcal{M}_1$. Therefore, being able to get the smallest face F_t is crucial when conducting goodness-of-fit tests, whenever the MLE doesn't exist. We will illustrate the above statistical inference with the real data example in

Section 10.1.2.

9 Numerical experiments for the computation of the MLE

In this chapter, we compare the performance of parameter estimation on several moderate dimensional and high dimensional graphical models using:

- the local one-hop relaxed marginal likelihood method, denoted $\mathcal{M}_1$-**MLE** in legends,
- the local two-hop relaxed marginal likelihood method, denoted $\mathcal{M}_2$-**MLE** in legends,
- the local pseudo-likelihood method, denoted **PS-MLE** in legends,
- the local 2-hop composite likelihood method, denoted PS_2-**MLE** in legends,
- the global likelihood method of the overall model, denoted as **G-MLE** in the legends,

First, several graph structures are given,and the parameters are either randomly assigned ± 0.5, or generated from normal distributions, then we generate sample points from each given model using the Gibbs sampling scheme. We compute the relative mean square error(MSE) defined as:

$$\frac{\|\hat{\theta} - \theta\|^2}{\|\theta\|^2} = \frac{\sum_{j \in \mathcal{J}} (\hat{\theta_j} - \theta_j)^2}{\sum_{j \in \mathcal{J}} \theta_j^2}$$

on sample points of different size. We also compare the accuracy of our estimates by looking at their sample variance.

9.1 Models of moderate dimension

Three moderate-size graphs are considered: a 5×5 grid graph(Fig. 9.1), 3×10 grid graph(Fig. 9.2) and 5×10 grid graph(Fig.9.3). For the 5×5 gird graph, the node in the middle of the grid has the largest two-hop marginal model, which includes 13 variables out of 25, that is 52% of the global model in the case of the 3×10 grid graph, the largest two-hop marginal model includes 11 variables out of 30, that is 37% of the global model in the case of the 5×10 graph, the largest two hop marginal model has 13 variables out of 50, 26% of the global model. From the MSE curves below, we can see that our two-hop marginal estimate $\mathcal{M}_2$**-MLE** is extremely close to the global estimate **G-MLE**, and this is not because the two-hop marginal model almost covers the variables in the global model. That's why we choose these three moderate-size models to illustrate our methods.

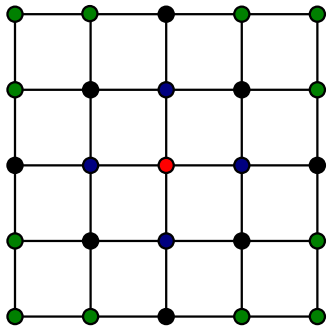


Figure 9.1: The 5×5 undirected grid graph. The one-hop neighbourhood of the red node is given by the blue nodes together with the red node. The two-hop neighbourhood is obtained from the one-hop neighbourhood by adding the black nodes.

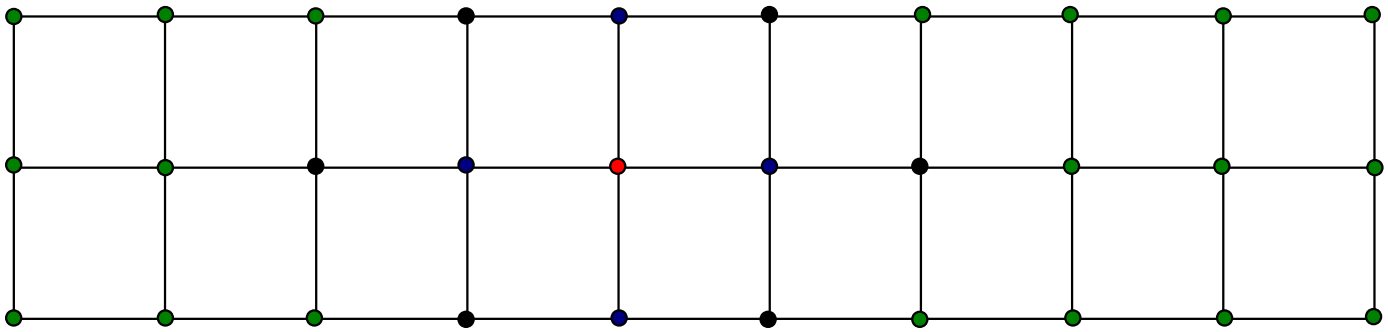


Figure 9.2: The 3×10 undirected grid graph. The one-hop neighbourhood of the red node is given by the blue nodes together with the red node. The two-hop neighbourhood is obtained from the one-hop neighbourhood by adding the black nodes.

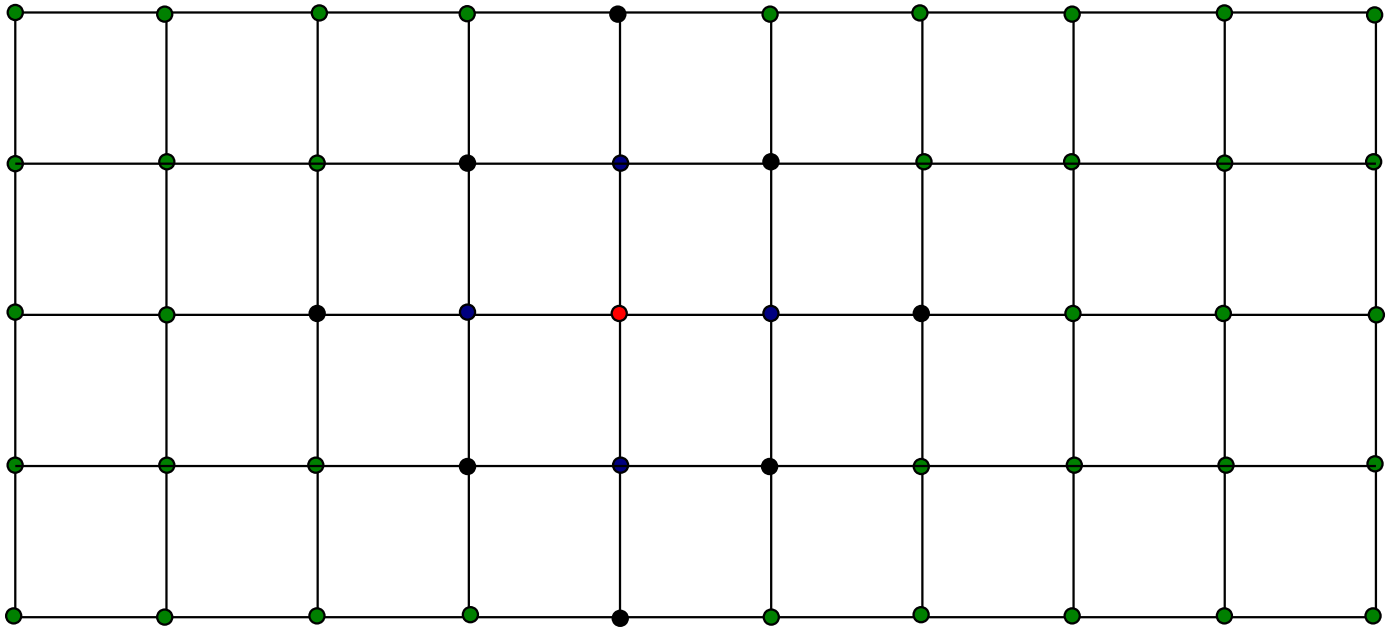
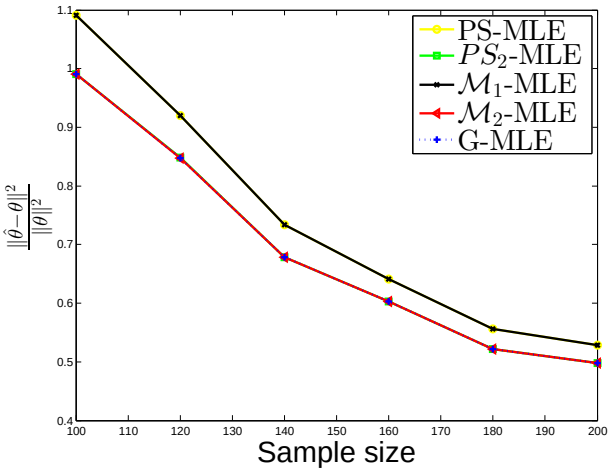


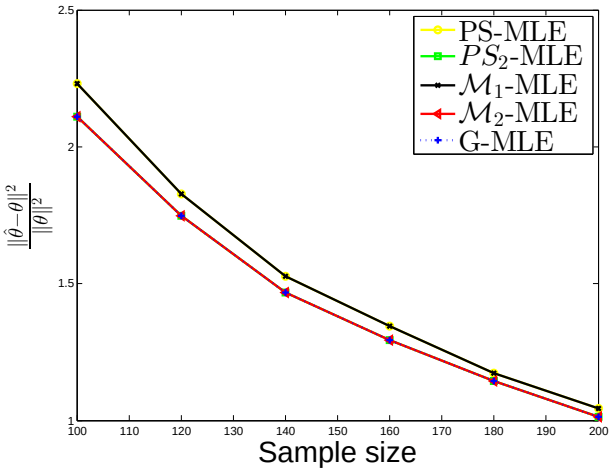
Figure 9.3: The 5×10 undirected grid graph. The one-hop neighbourhood of the red node is given by the blue nodes together with the red node. The two-hop neighbourhood is obtained from the one-hop neighbourhood by adding the black nodes.

We generate parameters from two different distributions: $\theta_j = \pm 0.5$ or $\theta_j \backsim \mathcal{N}(0, 0.1), \theta_{i,j} \backsim \mathcal{N}(0, 0.5)$ for 3×10 and 5×5 grid graphs. The relative MSE of different estimates are plotted versus

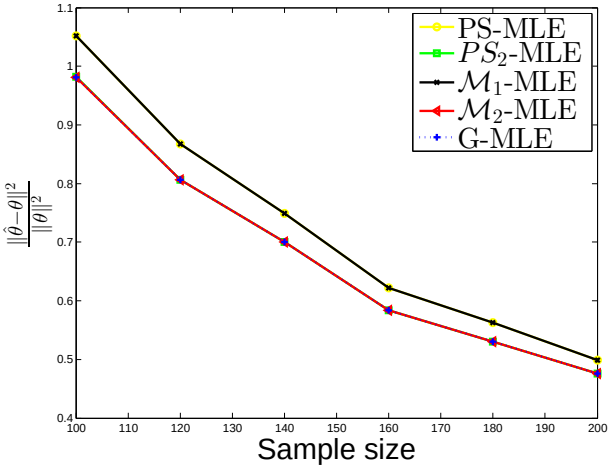
sample size as shown in Fig.9.4a, Fig.9.4b, Fig.9.4c and Fig. 9.4d . From these MSE curves, we can see that our one-hop marginal estimates($\mathcal{M}_1$-MLE) is extremely close to the pseudo-likelihood estimates(PS-MLE), and our two-hop marginal estimates($\mathcal{M}_2$-MLE) is extremely close to the global estimates(G-MLE). The MSE curves of the 3×10 graphical model and 5×5 graphical model are very similar, that's due to the fact that we compute the MLE from local marginal models, which share similar structures for these two models. Therefore the structure of the global model doesn't affect the estimates in a significant manner.



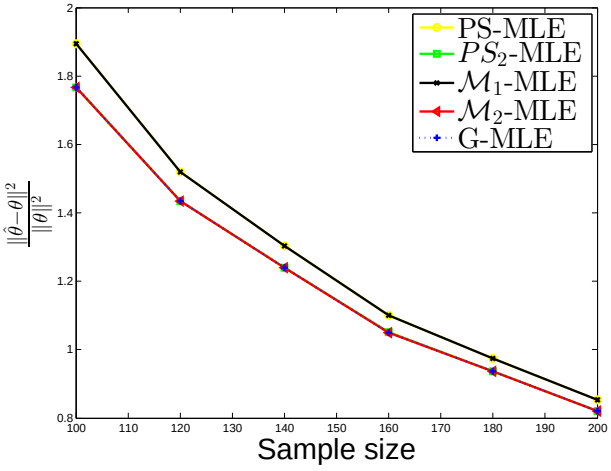
(a) 3×10 grid graph; $\theta_j = \pm 0.5$



(b) 3×10 grid graph $\theta_j \backsim \mathcal{N}(0, 0.1), \theta_{i,j} \backsim \mathcal{N}(0, 0.5)$



(c) 5×5 grid graph;$\theta_j = \pm 0.5$

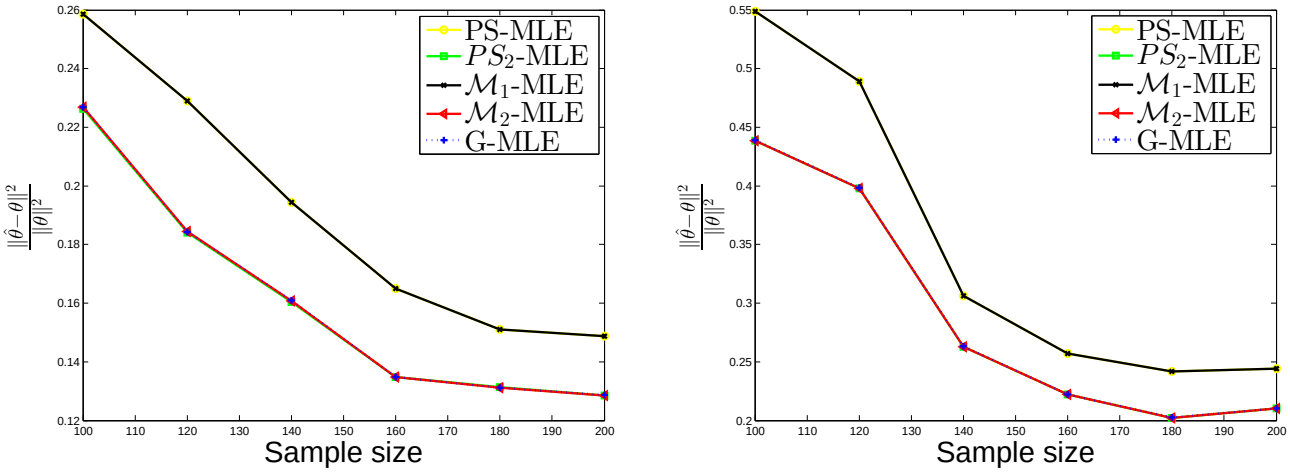


(d) 5×5 grid graph;$\theta_j \backsim \mathcal{N}(0, 0.1), \theta_{i,j} \backsim \mathcal{N}(0, 0.5)$

Figure 9.4: Relative MSE vs. sample size. The result is averaged over 100 experiments

.

We also compare the accuracy of different estimates by looking at their sample variance as shown in Fig 9.5a and Fig9.5b. The results in the plots are consistent with Theorem 2.



(a) sample variance of 5×5 grid graph

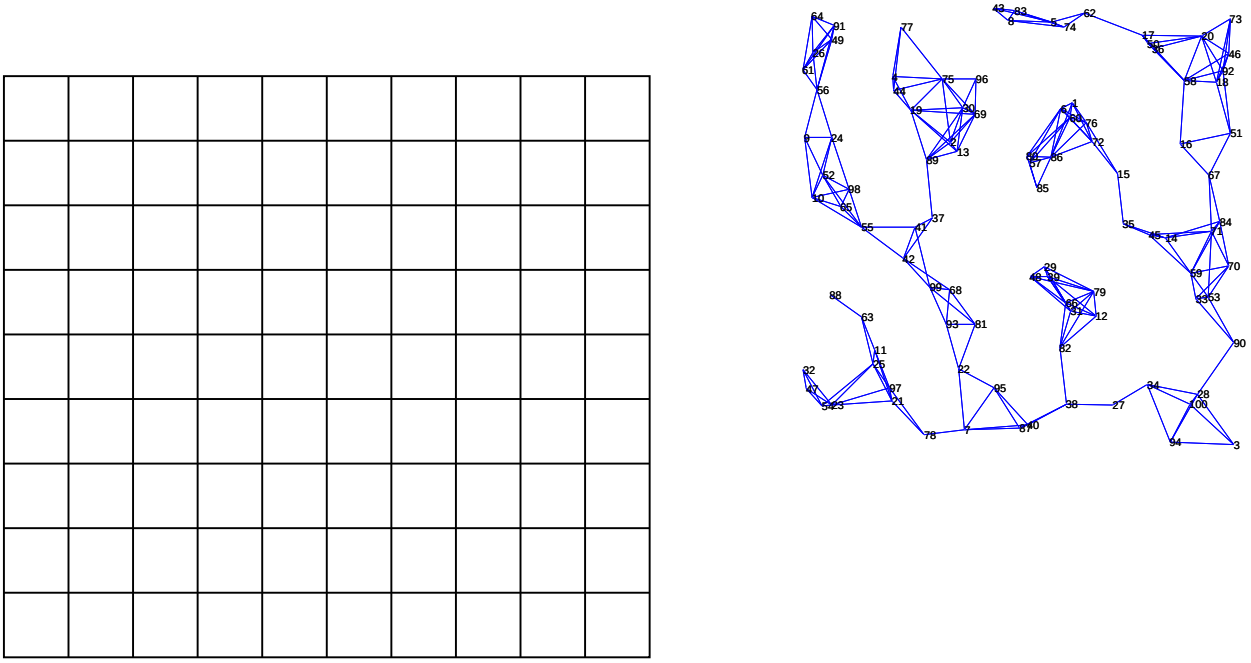
(b) sample variance 3×10 grid graph

Figure 9.5: Sample variance vs. sample size for (a) θ_9 in the 5×5 grid graph and (b) θ_8 in the 3×10 grid graph. The result is averaged over 100 experiments.

.

9.2 High-dimensional models

We first consider two high-dimensional discrete graphical models: 10×10 grid network(Fig. 9.6a) and 100-node random network (Fig. 9.6b). The 10×10 grid network describes the situation where every variable is affected only by its neighbours, or equivalently, it is independent of other nodes, given its neighbours. The random network is widely used in social science. Each vertex of a random network is connected to a limited number of members. Both of these two graphs are sparse



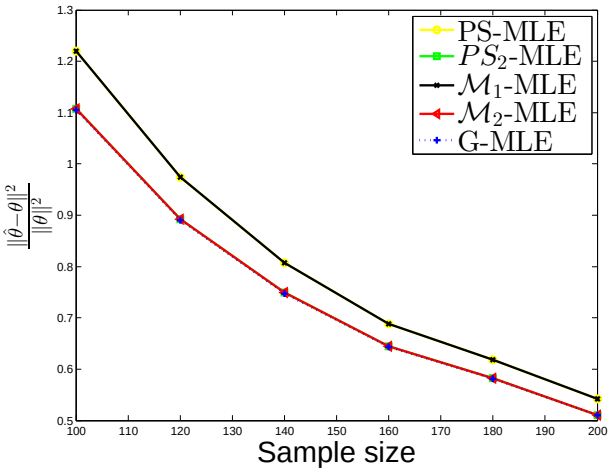
(a) the 10×10 grid graph

(b) the 100-node random network

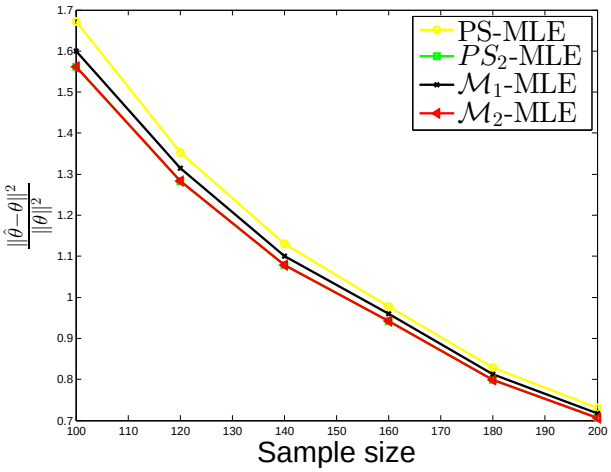
Figure 9.6: The two graphs underlying the two high-dimensional graphical models in section 9.2

.

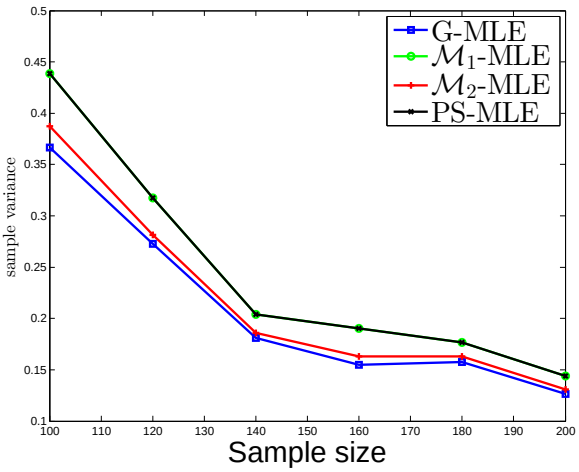
graphical models, as the graph degree is not too large, and thus the MLE computation is easier. In Fig. 9.7a and Fig. 9.7b, we can see that the relative mean square error of the conditional and marginal likelihood methods are the same, and that the two-hop cases are better than the one-hop cases. We also compute the MLE of the grid network graphical model, and we can't really see any difference between the MLE and the two-hop composite likelihood estimates. We also give the sample variance of some parameters in Fig. 9.7c and Fig. 9.7d.



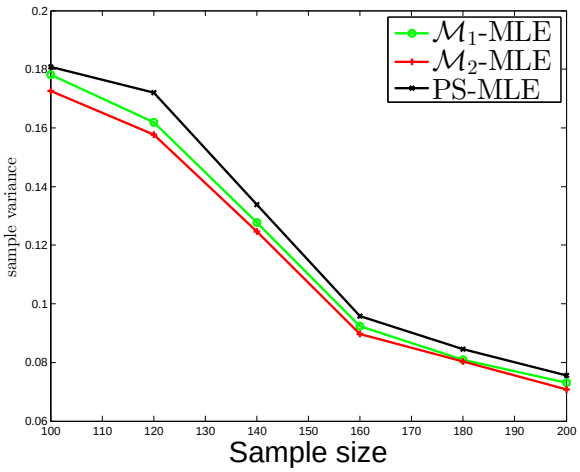
(a) MSE of 10×10 grid graph



(b) MSE of 100-node random graph

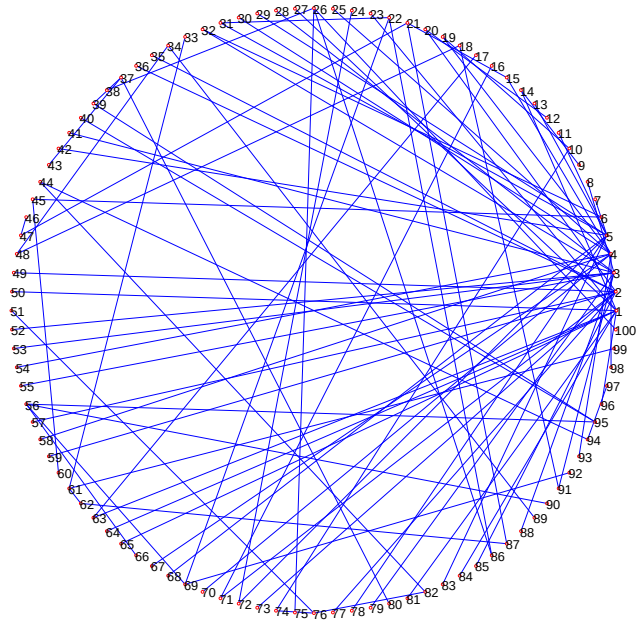


(c) sample variance of 10×10 grid graph

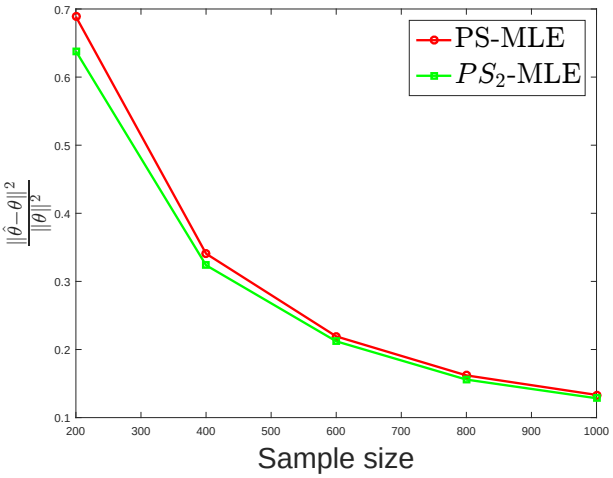


(d) sample variance of 100-node random graph

Figure 9.7: Relative MSE v.s. sample size for (a) the 10×10 grid graph and (b) the 100-node random graph. Sample variance vs. sample size for (c) θ_{43} in the 10×10 grid graph and (d) $\theta_{8,74}$ in the 100-node random graph. Parameters are assigned to ± 0.5 randomly, and the results are averaged over 100 experiments.



(a) The 100-node hub graph



(b) The relative MSE of parameters in the hub graph

The third example we look at is the hub network graph (Fig. 9.8a), which is also called the scale-free network. The biggest difference between the hub network and random network is the existence of hub nodes, whose degree increase as the number of variables increases: the hub graph is therefore not a sparse graphical model. In the 100-node hub network we generate, the degree of 5 vertices is 10, while the degree of other vertices is no larger than 5. For the vertices of large degree, the size of the local models is also large. We therefore only use conditional likelihood methods, as we have already shown that marginal and conditional methods are equivalent.

10 Numerical experiments on the existence of the MLE

10.1 Simulation study and application to real data

In this section, we illustrate our methodology. In 10.1.1, we simulate data for the graphical model of the 4×4 grid and show how to exploit the various types of separators in order to obtain good inner and outer approximations. We find that our methods give very accurate results in this model of modest size. In 10.1.2, we work with the NLTCS data set, a real-world data set. We compare different inner approximations F_1 and notice that most of the time, F_1 and F_2 are equal, and thus that they are both equal to F_t. We also compute the EMLE and compare the result to what happens when maximizing the likelihood functions l and l_{F_2}.

10.1.1 The 4×4 grid graph

We generated random samples of varying sizes for the graphical model of the 4×4 grid graph (Fig. 10.1). For each sample, we compute inner and outer approximations F_1 and F_2, and we compare them to the true facial set F_t, which we can obtain using linear programming. To obtain an inner approximation, we pick a separate set and complete it to create a reducible simplicial complex containing the 4×4 grid, we iterate the process over the 3 horizontal, 3 vertical and 8 diagonal separators. To compute the outer approximation, we cover the 4×4 grid by four 3×3-grids.

We first generate random samples from the uniform distribution, that is, from the probability

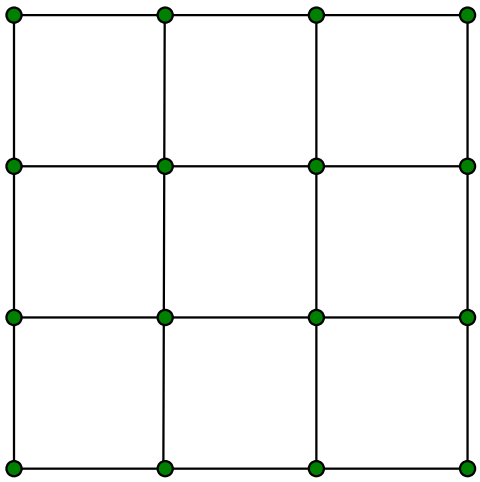


Figure 10.1: The 4×4 grid graph

distribution P_θ in the hierarchical model where all parameters $\theta_j, j \in J$ are set to zero. The results are given in Table 10.1. We repeat the experiment a thousand times for each sample size. As the table shows, for larger samples the probability that our random sample lies on a proper face becomes very small. If $F_t = I$, then clearly $F_t = F_2$. But we also found $F_t = F_2$ for all samples with t lying on a proper face, which shows that F_2 is an excellent approximation of F_t in this model. For the inner approximation, we observed some samples with $F_1 \neq F_t$, but they seem to be very rare.

Table 10.1: Facial set approximation for the 4×4 grid graph sampling from the uniform distribution

sample size	data on face	$F_1 = F_t$	$F_2 = F_t$
10	98.5%	96.3%	100.0%
15	68.9%	99.9%	100.0%
20	29.0%	100.0%	100.0%
50	0.0%	100.0%	100.0%

Second, to better understand what happens in the case of large samples, we change our sampling

scheme. Instead of sampling from the uniform distribution, we generate samples from the hierarchical model P_θ, where the vector of parameters θ is drawn from a multivariate standard normal distribution (for each sample, new parameters were drawn). The results are given in Table 10.2. Again, for each sample size, we run the experiment a thousand times. One can see that in this sampling scheme, we are much more likely to find that $F_t \neq I$. Observe that the squared length of the parameter vector θ is χ^2-distributed with 39 degrees of freedom (since the number of parameters is 40). Thus, the expected length of θ is 39, which is large enough to move the distribution p_θ close to the boundary of the model. Indeed, we observed that when the MLE does not exist, the length of the numerical estimate of the MLE vector is of the order of magnitude 40(see also the next example in Section 10.1.2). Again, in all the samples that we generated, $F_t = F_2$, and $F_1 = F_2$ in the vast majority of cases. Thus, for this graph of relatively modest size, our approximations are very good.

Table 10.2: Facial set approximation for the 4×4 grid graph with log-linear parameters from the standard normal distribution

sample size	data on face	$F_1 = F_t$	$F_2 = F_t$
10	100.0%	97.7%	100.0%
50	89.5%	100.0%	100.0%
100	71.0%	100.0%	100.0%
150	52.0%	100.0%	100.0%

10.1.2 The NLTCS data set

In order to illustrate how approximate knowledge of the facial set allows us to say which parameters can be estimated, and to conduct statistical inference (as explained in Section 8), we study

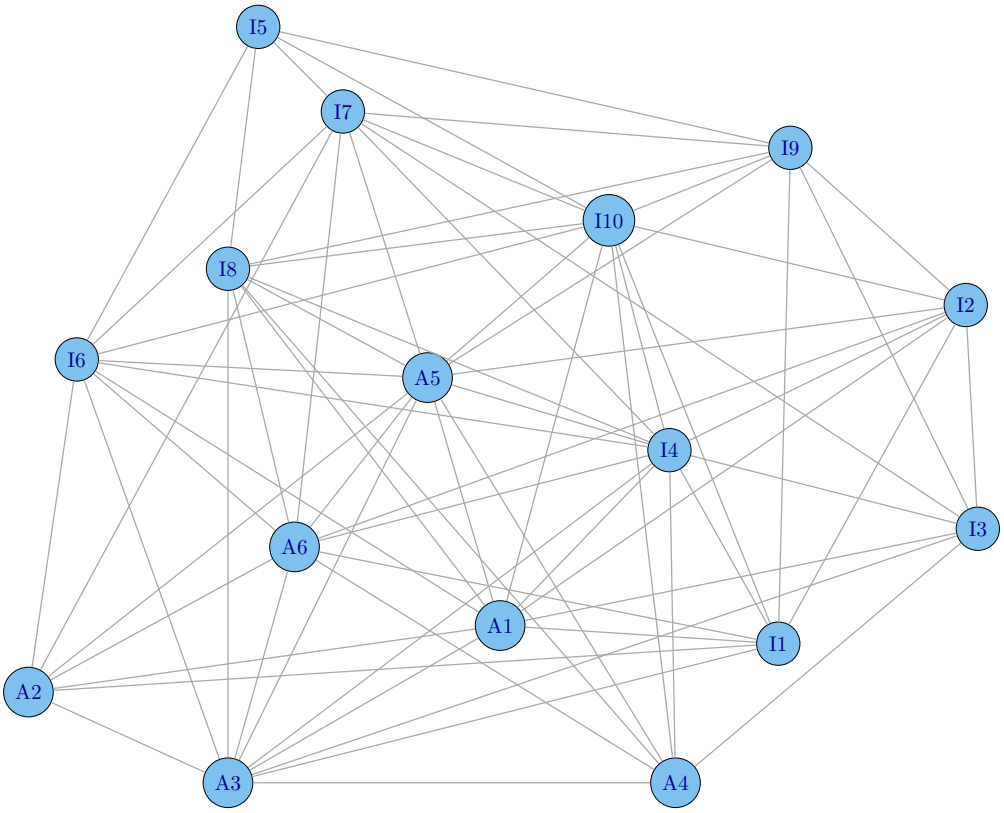


Figure 10.2: The graph for the NLTCS dataset

the NLTCS data set, which consists of 21574 observations on 16 binary variables, called ADL1, ..., ADL6, IADL1, ..., IADL10. In our code and the following equations, these variables are indexed by 16 integers from 1 to 16. The reader is referred to Dobra et al. (2011) for a detailed description of the data set. To associate a hierarchical log-linear model to this data, we rely on the results of Dobra et al. (2011) who use a Bayesian approach to estimate the posterior inclusion probabilities of edges. We construct a graph by saying that (x, y) is an edge if and only if the posterior inclusion probability of (x, y) is at least 0.40; see Figure 10.2. We then take the corresponding clique complex of this graph so that our hierarchical model is a graphical model. There are 314 parameters in this model, including up to 6-way interactions.

Using linear programming, we find the smallest facial set F_t containing the sufficient statistic. The face $\mathbf{F}_t$ is then the convex hull of the $f_i, i \in F_t$. The dimension of $\mathbf{F}_t$ is 302, and we can compute

the basis of the kernel space of $\mathbf{F}_t$, which gives us the following equations representation of $\mathbf{F}_t$:

$$\begin{cases} t_{10,12,13,16} - t_{12,13,16} - t_{10,13,16} + t_{13,16} = 0 \\ t_{8,9,10} - t_{8,9} - t_{9,10} + t_9 = 0 \\ t_{7,8,9,10} - t_{7,8,9} - t_{7,9,10} + t_{7,9} = 0 \\ t_{5,10,12,13,16} - t_{5,10,13,16} - t_{5,12,13,16} + t_{5,13,16} = 0 \\ t_{3,7,9,10} - t_{3,7,9} - t_{3,9,10} + t_{3,9} = 0 \\ t_{1,10,12,16} - t_{1,10,16} - t_{1,12,16} + t_{1,16} = 0 \\ t_{1,8,9,10} - t_{1,8,9} - t_{1,9,10} + t_{1,9} = 0 \\ t_{1,7,9,10} - t_{1,7,9} - t_{1,9,10} + t_{1,9} = 0 \\ t_{1,7,8,9,10} - t_{1,7,8,9} - t_{1,9,10} + t_{1,9} = 0 \\ t_{1,5,10,12,16} - t_{1,5,12,16} - t_{1,5,10,16} + t_{1,5,16} = 0 \\ t_{1,3,9,10} - t_{1,3,9} - t_{1,9,10} + t_{1,9} = 0 \\ t_{1,3,7,9,10} - t_{1,3,7,9} - t_{1,9,10} + t_{1,9} = 0 \end{cases} \tag{10.1.1}$$

Each equation represents a facet of some clique after we verified in the program. The intersection of these 12 facets gives us the smallest face containing sufficient statistic. Therefore our program can give both the facial set and the face equations. As we show in section 8.2, We can compute the extended MLE of the estimable parameters in a reduced exponential family supported on the facial set F_t, whose log-likelihood function is

$$l_{F_t}(\theta) = \log(\prod_{i \in I_+} p_{F_t,\theta}(i)^{n(i)}) = \sum_{j \in J} \theta_j t_j - N k_{F_t}(\theta). \tag{10.1.2}$$

The log-likehood function $l_{F_t}(\theta)$ with $\theta \in R^{|J|}$ is not identifiable, and the optimization algorithm

doesn't converge. We can find a linear transformation of θ such that the corresponding new parameterization is identifiable. In section 8.2, we introduced a new identifiable parameterization μ, which can be easily found if we know the facial set F_t. The log-likelihood function with respect to μ is

$$l_{F_t}(\mu) = \sum_{i \in I_+} \mu_i n(i) - N \log \sum_{i \in F_t} \exp(\mu_i). \tag{10.1.3}$$

In order to compare the maximum likelihood estimate obtained with or without worrying about its existence and with or without approximation to F_t, we maximize the log-likelihood function given in terms of μ (rather than θ) as in (8.2.1).

First we ignore the fact that the MLE might not exist and compute the MLE of μ using the standard "Minfunc" optimization software in Matlab: we call this estimate $\hat{\mu}^{\text{MLE}}$. Second, we find F_t and compute the EMLE with parameters denoted $\hat{\mu}^{\text{EMLE}}$. Third, we obtain an inner and outer approximation to F_t and consider the resulting information on the MLE of the parameters. We call the resulting estimate $\hat{\mu}^{F'_1/F'_2}$.

To compute $\hat{\mu}^{\text{EMLE}}$, we first compute the inner approximation F_1 that makes use of all the separators in the graph (Strategy 7.2 in Section 7.2). We also compute an outer approximation F_2 from all $\binom{16}{5} = 4368$ size five local models and the cliques of size six (Strategy 1 in Section 7.3). We obtain $F_1 = F_2$ and thus deduce that $F_t = F_1 = F_2$. We find $|F_t| = 49536$, and so $|F_t^c| = 2^{16} - 49536 = 16000$. Therefore, 16000 cell probabilities are zero in the EMLE. We can obtain the MLE by maximizing the log likelihood function l_{F_t} as in (8.1.1). Since $rank(A_{F_t}) = 302$, the dimension of $\mathbf{F}_t$ is 302, and there are only 302 parameters in l_F.

To show how to use the inner and outer approximations when F_t is not known, we choose to find coarser inner and outer approximations to F_t, respectively denoted F'_1 and F'_2, and use them to compute the other approximation $\hat{\mu}^{F'_1/F'_2}$ to the MLE. To compute F'_1, we just use 10

random separators. We find $|F_1'| = 36954$ and $\dim \mathbf{F}_1' = rank A_{F_1'} = 300$. To compute the outer approximation F_2', we consider the 4368 local size-five induced models and select from them those 1000 which have the facial sets of smallest cardinality, and then we glue them together. We find $|F_2'| = 50688$ and $\dim \mathbf{F}_2' = rank A_{F_2'} = 310$. Thus, we know that at least $|I \setminus F_2'| = 2^{16} - 50688 = 14848$ cell probabilities vanish in the extended MLE. Since we pretend not to know F_t, we replace l_{F_t} by

$$l_{F_2'}(\mu) = \sum_{i \in I_+} \mu_i n(i) - N \sum_{i \in F_2'} \exp(\mu_i). \tag{10.1.4}$$

For $i \in F_1'$, we know that μ_i is estimable, μ_i goes to negative infinity when $i \in F_2'^c$, and we cannot say anything for μ_i when $i \in F_2' \setminus F_1'$.

As explained in Section 8.2, the components of μ are not functionally independent. We choose $L_1 \subseteq F_1'$, $L_2 \subseteq F_2'$ and $L \subseteq I$ as in Section 8.2 (we note that the zero cell belongs to I_+). Then any μ_i, $i \in F_2'$, can be written as a linear combination of $\mu_{L_2} = (\mu_i, i \in L_2)$, and we can write $\mu_i = \langle b_i, \mu_L \rangle$ for an appropriate vector b_i. Thus, $l_{F_2'}(\mu)$ only depends on $\mu_{L_2} = (\mu_i, i \in L_2)$, and (10.1.4) can be rewritten as

$$l_{F_2'}(\mu_L) = \sum_{i \in I_+} \langle b_i, \mu_L \rangle n(i) - N \sum_{i \in F_2'} \exp\langle b_i, \mu_L \rangle. \tag{10.1.5}$$

Of course, the maximum of $l_{F_2'}$ does not exist, but, insofar as the maximization of l, the computer can still give us a numerical approximation, $\hat{\mu}_L$, and thus also a numerical estimate $\hat{\mu}_i = \langle b_i, \hat{\mu}_L \rangle, i \in F_2'$.

In total, there are $|L_2| = rank(A_{F_2'}) = 310$ independent parameters in the log likelihood function (10.1.5). Among them, we find $|L_2| = rank(A_{F_2'}) = 300$ estimable parameters $\mu_i, i \in L_2$. We cannot say anything about the 10 parameters indexed by $L_2 \setminus L_1$. If we know F_t, we can identify two more estimable parameters.

In Table 10.3, we give the three estimates of μ_i that we mentioned above, namely, $\hat{\mu}_i^{\mathrm{MLE}}, \hat{\mu}_i^{\mathrm{EMLE}}$

and $\hat{\mu}_i^{F'_1/F'_2}$. For convinence, in the parameter column, we write μ_i as $\mu_{k(i)}$ where $k(i) = \sum_{j=1}^{16} i_j 2^{j-1} \in \{0, \ldots, 2^{16} - 1\}$. We also list the naive estimator $\log \frac{n_i}{n_0}$. We list estimates for 19 of the 310 possible parameters. In the first column of the table, we indicate which category index i belongs to, that is, whether it belongs to F'_1, F_t or F'_2. In the second column, we list the particular parameters considered.

In Table 10.4, we list the estimates of the top five cell counts obtained using our method and compare them with those obtained by other methods in Dobra et al. (2011).

The graphical model of the NLTCS dataset we use above includes up to six-way interaction parameters. Let $\mathcal{M}_0$ denote this graphical model. Now let's consider another model with only two-way interaction parameters, and denote it by $\mathcal{M}_1$. We have already known that the MLE of $\mathcal{M}_0$ doesn't exist, and we observe that the MLE of M_1 exists from our program. Let M_0 denote the original six-way interaction model, M_1 denote the two-way interaction model, and l_0, l_1 be the log-likelihood functions of $\mathcal{M}_0$ and M_1 respectively. We can use the likelihood ratio test to see which model fits the data well. We define the test as follows,

$$H_0 : \text{The reduced model } \mathcal{M}_1 \text{ fits the data better}$$

$$H_a : \text{The current model } \mathcal{M}_0 \text{ fits the data better}$$

Although we don't have the MLE for M_0, the maximum value of l_0 using the extended MLE is still approximately correct. From the experiment, we get $l_0(\hat{\theta}_0) = -1.2954 \times 10^5$, $l_1(\hat{\theta}_1) = -1.2971 \times 10^5$, so the likelihood ratio statistic,

$$G^2 = -2(l_1(\hat{\theta}_1) - l_0(\hat{\theta}_0)) = 170.$$

The asymptotic distribution of G^2 is chi-quare distribution, and the adjust degree of freedom of this test is 214. The p-value $p(\chi^2(214) > G^2)$ is less then 0.001, so we reject the null hypothesis.

Table 10.3: The MLE from 3 methods and the naive estimate for the NLTCS dataset.

		naive estimate	maximum likelihood estimates		
	Parameter	$\log n_i/n_0$	$\hat{\mu}_i^{\text{MLE}}$	$\hat{\mu}_i^{\text{EMLE}}$	$\hat{\mu}_i^{F_1'/F_2'}$
$i \in F_1'$	μ_{512}	-1.2472	-1.2482	-1.2482	-1.2482
	μ_{65536}	-1.7644	-1.7976	-1.7975	-1.7975
	μ_{16}	-2.3958	-2.3844	-2.3846	-2.3846
	μ_{528}	-2.5429	-2.6504	-2.6504	-2.6504
	μ_{2048}	-2.8813	-2.7246	-2.7243	-2.7243
$i \in F_t \setminus F_1'$	μ_{32960}	$-\infty$	-13.8205	-13.8207	-13.8205
	μ_{34881}	$-\infty$	-14.3693	-14.3693	-14.3692
$i \in F_2' \setminus F_t$	μ_{36864}	$-\infty$	-30.8729	$-\infty$	-34.9805
	μ_{36880}	$-\infty$	-39.6536	$-\infty$	-45.2229
	μ_{388}	$-\infty$	-28.9090	$-\infty$	-29.4525
	μ_{32769}	$-\infty$	-32.3799	$-\infty$	-36.9537
	μ_{385}	$-\infty$	-37.1365	$-\infty$	-35.9399
	μ_{449}	$-\infty$	-38.9673	$-\infty$	-44.9405
	μ_{32785}	$-\infty$	-40.1221	$-\infty$	-45.8318
	μ_{389}	$-\infty$	-43.7297	$-\infty$	-40.0158
$i \in I \setminus F_2'$	μ_{256}	$-\infty$	-35.5482	$-\infty$	$-\infty$
	μ_{320}	$-\infty$	-42.5454	$-\infty$	$-\infty$
	μ_{257}	$-\infty$	-52.9224	$-\infty$	$-\infty$
	μ_{321}	$-\infty$	-60.2208	$-\infty$	$-\infty$

Table 10.4: Top six largest expected cell counts for the NLTCS data set according to the Grade of Membership model (GoM), Latent class model (LC), copula Gaussian graphical model (CGGM) and MLE.

Support of Cell	Observed	GoM	LC	CGGMs	MLE on facial set
$\emptyset$	3853	3269	3836.01	3767.76	3647.4
$\{10\}$	1107	1010	1111.51	1145.86	1046.9
$\{1:16\}$	660	612	646.39	574.76	604.4
$\{5\}$	351	331	360.52	452.75	336
$\{5,10\}$	303	273	285.27	350.24	257.59
$\{12\}$	216	202	220.47	202.12	239.24

10.2 Computing faces for large complexes

If our statistical model contains many variables and is not reducible, the problem of determining $\mathbf{F}_t$ quickly becomes infeasible. Not only does the marginal polytope become very complicated, but also the size of the objects that one has to store or compute grows exponentially. Consider for example a 10×10 grid of binary random variables. This hierarchical model has 280 parameters, and the total sample space has cardinality $|I| = 2^{100} \approx 1.27 \times 10^{30}$. If F_t is close to I, we cannot even list the elements of F_t, which consists of approximately 10^{30} elements. Therefore, we take a local approach and look for separators.

If Δ contains a complete separator separating V into V_1 and V_2, we can identify a facial set F implicitly without listing it explicitly. We only need the two projections $F_{V_1} = \pi_{V_1}(F)$ and $F_{V_2} = \pi_{V_2}(F)$. Since $F = \pi_{V_1}^{-1}(F_{V_1}) \cap \pi_{V_2}^{-1}(F_{V_2})$ (by Lemma 7.1.5), these two projections identify F,

and they allow us to do most of the operations that we would want to do with F. For example, for any $i \in I$, we can check whether $i \in F$ by checking whether $\pi_{V_1}(i) \in F_{V_1}$ and $\pi_{V_2}(i) \in F_{V_2}$, and we can check whether $F = I$ by checking whether $F_{V_1} = I_{V_1}$ and $F_{V_2} = I_{V_2}$. In particular, we can check whether the MLE exists by looking only at the two subsets V_1 and V_2.

If Δ contains a separator that is not complete, we can use similar ideas as those above, when computing inner and outer approximations to F_t, and also when comparing these two approximations. Suppose that S separates V_1 from V_2 in Δ. We want to use $F_2 := F_{\Delta|_{V_1}}(I_+) \cap F_{\Delta|_{V_2}}(I_+)$ as an outer approximation and $F_1 := F_{\Delta_S}(I_+)$ as an inner approximation to F_t. Due to the problems mentioned above, we do not directly compute F_1 and F_2, but we compute their projections on V_1 and V_2. Instead of F_2, we compute the facial set $F_{2,V_1} := F_{\Delta|_{V_1}}(\pi_{V_1}(I_+))$ of the V_1-marginal $\pi_{V_1}(I_+)$ with respect to $\Delta|_{V_1}$, and similarly we compute $F_{2,V_2} := F_{\Delta|_{V_2}}(\pi_{V_2}(I_+))$. Instead of F_1, we compute $F_{1,V_1} := F_{\Delta_S|_{V_1}}(\pi_{V_1}(I_+))$ and $F_{1,V_2} := F_{\Delta_S|_{V_2}}(\pi_{V_2}(I_+))$. Then we could recover F_1 and F_2 from the equations

$$F_2 = \pi_{V_1}^{-1}(F_{2,V_1}) \cap \pi_{V_2}^{-1}(F_{2,V_2}) \quad \text{and} \quad F_1 = \pi_{V_1}^{-1}(F_{1,V_1}) \cap \pi_{V_2}^{-1}(F_{1,V_2}).$$

For any $x \in I$, we can check whether $x \in F_1$ by checking whether $\pi_{V_1}(x) \in F_{1,V_1}$ and $\pi_{V_2}(x) \in F_{1,V_2}$. More importantly, we can check whether $F_1 = F_2$ by checking whether $F_{1,V_1} = F_{2,V_1}$ and $F_{1,V_2} = F_{2,V_2}$. This idea can be applied iteratively when $\Delta|_{V_1}$ or $\Delta|_{V_2}$ has a separator.

The next two subsections illustrate these ideas. In Section 10.2.1, we consider a graph with no particular regularity pattern on 100 nodes, and identify two convenient separators. In Section 10.2.2, we consider a grid graph and work with two families of "parallel" separators that can be used to iteratively improve the inner approximation.

10.2.1 US Senate Voting Records dataset

We consider the voting record of all 100 US Senators on 309 bills from January 1 to November 19 2015. Similar data for the years 2004–2006 was analyzed by Banerjee et al. (2008). The votes are recorded as "yea," "nay" or "not voting." We transformed the "not voting" into "nay" and consequently have a 100-dimensional binary data set. To fit a hierarchical model to this data set, we use the ℓ_1-regularized logistic regression method proposed by Ravikumar et al. (2011) to identify the neighbours of each variable and construct an Ising model. We set the penalty parameter to $\lambda = 32\sqrt{\log p/n} \approx 0.35$, resulting in the sparse graph in Figure 10.3. There are 277 parameters in this model (the number of vertices plus the number of edges). The graph consists of two large connected components and 14 independent nodes.

There are 309 sample points, and $|I_+| = 278$. We want to know whether the data lies on a proper face of the marginal polytope to see if the MLE of the parameters exists. From Lemma 7.1.5, we know that if we find complete separators, we need only work with each of the irreducible simplicial complexes defined by these separators. We easily "cut-off" a number of relatively small prime components and verify that the data does not lie on a proper face of their corresponding marginal polytopes. We are left with one irreducible prime component in each of the two connected subgraphs, i.e. one for each of the two parties as shown in Figure 10.4.

The democratic party simplicial complex Δ_d consists of 26 variables, and the model induced from Δ_d contains 77 parameters. The size of the design matrix A_{Δ_d} is $2^{26} \times 77$, which is too large to use linear programming to compute the facial set of the face $\mathbf{P}_{\Delta_d}$ containing the vector t_d. Therefore we look for separators that will help us obtain good inner and outer approximations. In Figure 10.4b, we indicate in yellow and pink two separators, which separate Δ_d into three simplicial complexes denoted (from top to bottom) by Δ_α, Δ_β and Δ_γ. The number of vertices of the three

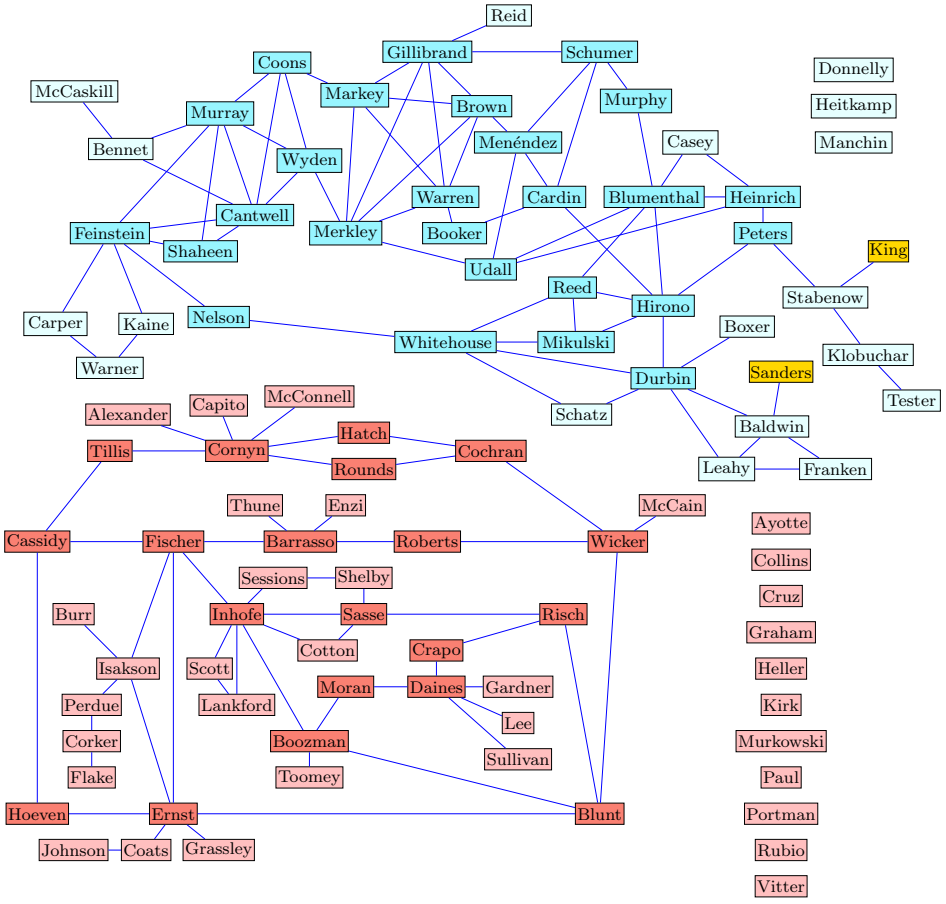
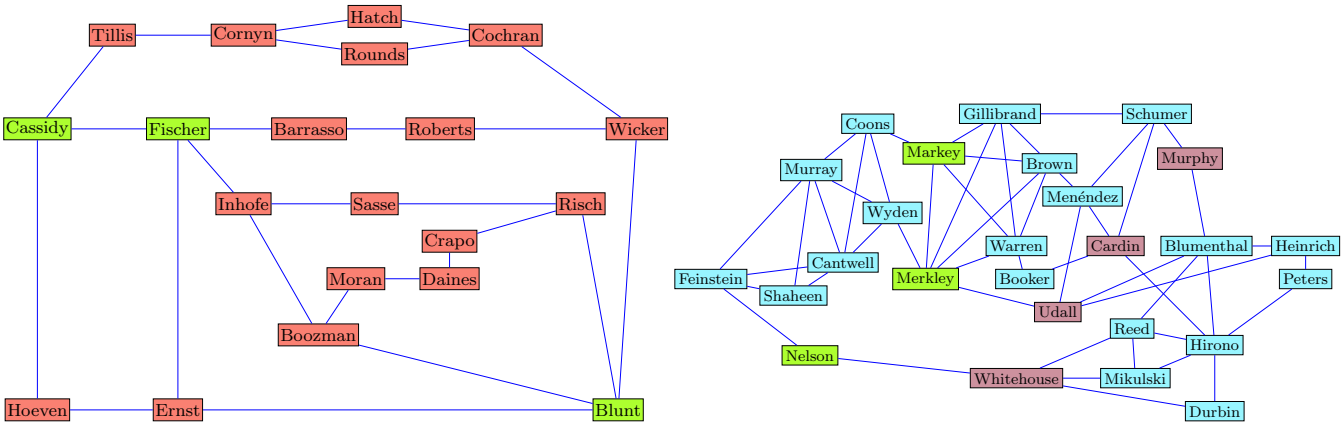


Figure 10.3: The graph for the US Senate Voting Records dataset. Golden nodes denote independent senators, blue nodes - democrats, and red nodes - republicans.



(a) (b)

Figure 10.4: The simplicial complexes after cutting off the small complete prime components: (a) the republican party prime component Δ_r (b) the democratic party prime component Δ_d. The light green and pink nodes are the two separator sets we selected to compute the facial sets.

simplicial complexes are 9, 13, and 11, respectively, so we can apply linear programming method first mentioned in the introduction to the three corresponding marginal polytopes.

The dimension of the model induced by Δ_α is 24. The corresponding data vector t_α lies in the relative interior of $\mathbf{P}_{\Delta_\alpha}$.

The dimension of the model induced by Δ_β is 34, and the data vector t_β lies on a facet $\mathbf{F}_{t_\beta}$ of $\mathbf{P}_{\Delta_\beta}$. To simplify our notation, we denote the 100 senators by an integer between 1 and 100, rather than

ID	Senator	ID	Senator	ID	Senator	ID	Senator
22	Nelson	37	Cardin	52	Murphy	61	Whitehouse
23	Reed	41	Markey	53	Hirono	87	Warren
26	Schumer	47	Udall	56	Gillibrand		

Table 10.5: Assigning numbers to the senators appearing in the equation of the faces

by their names. We only need to identify a few and their numbers are given in Table 10.5. The inequality of $\mathbf{F}_{t_\beta}$ is

$$t_{87} - t_{56,87} \geq 0, \tag{10.2.1}$$

where t_{87} denotes the marginal count of senator Warren voting "yea" and $t_{56,87}$ denotes the marginal counts of both senators Gillibrand and Warren voting "yea."

The dimension of the model induced by Δ_γ is 27. The data vector t_γ lies on the facet of $\mathbf{P}_{\Delta_\gamma}$ with inequality

$$t_{23} - t_{23,53} \geq 0. \tag{10.2.2}$$

The intersection of the two facets (10.2.1) and (10.2.2) gives the outer approximation $\mathbf{F}_2$ to F_t.

To get an inner approximation, we complete each separator, i.e. the yellow vertices are completed, and the pink vertices are completed, as shown in Figure 10.4b. Denote the three simplicial complexes with complete separators as $\Delta_{\tilde{\alpha}},\ \Delta_{\tilde{\beta}},\ \Delta_{\tilde{\gamma}}$ respectively. Then $\Delta_{\tilde{d}} = \Delta_{\tilde{\alpha}} \cup \Delta_{\tilde{\beta}} \cup \Delta_{\tilde{\gamma}}$ is a simplicial complex with two complete separators. The smallest face $\mathbf{F}_{t_{\tilde{d}}}$ of the marginal polytope $\mathbf{P}_{\Delta_{\tilde{d}}}$ containing the data vector $t_{\tilde{d}}$ is our inner approximation. Now the models of $\Delta_{\tilde{\alpha}}$, $\Delta_{\tilde{\beta}}$, $\Delta_{\tilde{\gamma}}$ and $\Delta_{\tilde{d}}$ are not models with main effects and two-way interactions only; they also include parameters for third and fourth order interactions. The dimension of the model induced by $\Delta_{\tilde{d}}$ is 91: we added 14 parameters to the original model by completing the two separators. Again, we apply the linear programming method to the three marginal polytopes $\mathbf{P}_{\Delta_{\tilde{\alpha}}}$, $\mathbf{P}_{\Delta_{\tilde{\beta}}}$ and $\mathbf{P}_{\Delta_{\tilde{\gamma}}}$.

The dimension of the model of $\Delta_{\tilde{\alpha}}$ is 27, and $\mathbf{F}_{t_{\tilde{\alpha}}}$ is a facet with equation

$$\langle g_1,\ t_{\tilde{\alpha}} \rangle = t_{41} - t_{22,41} - t_{41,70} + t_{22,41,70} = 0. \tag{10.2.3}$$

It follows that $\{g_1\}$ is a basis of the kernel of $A^t_{F_{\tilde{\alpha}}}$.

The dimension of the model for $\Delta_{\tilde{\beta}}$ is 48. The face $\mathbf{F}_{t_{\tilde{\beta}}}$ has codimension 5, with defining equations

$$\begin{cases} \langle g_2,\ t_{\tilde{\beta}} \rangle = t_{87} - t_{56,87} = 0 \\ \langle g_3,\ t_{\tilde{\beta}} \rangle = t_{47,52,61} + t_{37,52} - t_{37,52,61} - t_{37,47,52} = 0 \\ \langle g_4,\ t_{\tilde{\beta}} \rangle = t_{37,47,52,61} - t_{47,52,61} = 0 \\ \langle g_5,\ t_{\tilde{\beta}} \rangle = t_{37,52} + t_{26} - t_{26,52} - t_{26,37} = 0 \\ \langle g_6,\ t_{\tilde{\beta}} \rangle = t_{41} - t_{22,41} - t_{41,70} + t_{22,41,70} = 0 \end{cases} . \tag{10.2.4}$$

Again, $\{g_2, g_3, g_4, g_5, g_6\}$ is a basis of the kernel of $A_{F_{\tilde{\beta}}}$.

The dimension of the model for $\Delta_{\tilde{\gamma}}$ is 38. The face $\mathbf{F}_{t_{\tilde{\gamma}}}$ has codimension 3. It is defined by the equations

$$\begin{cases} \langle g_7,\ t_{\tilde{\gamma}} \rangle = t_{47,52,61} + t_{37,52} - t_{37,52,61} - t_{37,47,52} = 0 \\ \langle g_8,\ t_{\tilde{\gamma}} \rangle = t_{37,47,52,61} - t_{47,52,61} = 0 \\ \langle g_9,\ t_{\tilde{\gamma}} \rangle = t_{23} - t_{23,53} = 0 \end{cases} . \tag{10.2.5}$$

Again, $\{g_7, g_8, g_9\}$ is a basis of the kernel of $A_{F_{\tilde{\gamma}}}$.

From Lemma 7.1.5, we know that $\mathbf{F}_{t_{\tilde{d}}} = \mathbf{F}_{\tilde{\alpha}} \cap \mathbf{F}_{\tilde{\beta}} \cap \mathbf{F}_{\tilde{\gamma}}$, and the equations for $\mathbf{F}_{t_{\tilde{d}}}$ are

$$\begin{cases} \langle g'_1,\ t_{\tilde{d}} \rangle = t_{41} - t_{22,41} - t_{41,70} + t_{22,41,70} = 0 \\ \langle g'_2,\ t_{\tilde{d}} \rangle = t_{87} - t_{56,87} = 0 \\ \langle g'_3,\ t_{\tilde{d}} \rangle = t_{47,52,61} + t_{37,52} - t_{37,52,61} - t_{37,47,52} = 0 \\ \langle g'_4,\ t_{\tilde{d}} \rangle = t_{37,47,52,61} - t_{47,52,61} = 0 \\ \langle g'_5,\ t_{\tilde{d}} \rangle = t_{37,52} + t_{26} - t_{26,52} - t_{26,37} = 0 \\ \langle g'_9,\ t_{\tilde{d}} \rangle = t_{23} - t_{23,53} = 0 \end{cases} , \tag{10.2.6}$$

where the vectors $g'_1, \ldots, g'_9$ are the vectors $g_1, \ldots, g_9$ extended to R^{91} by adding zeros on the corresponding complementary coordinates. Note that since $g'_1 = g'_6$, $g'_3 = g'_7$, $g'_4 = g'_8$, we only need six of the nine equations. Thus, $\mathbf{F}_1 := \mathbf{F}_{t_{\tilde{d}}}$, defined by (10.2.6), is a strict subset of the face $\mathbf{F}_2$ defined by (10.2.1) and (10.2.2). Next, we refine our argument and show that indeed $\mathbf{F}_{t_d} = \mathbf{F}_2$.

From what we know, it follows that the orthogonal complement of the subspace generated by $\mathbf{F}_{t_{\tilde{d}}}$ is

$$G = \{g' \in R^{91} | g' = k_1 g'_1 + k_2 g'_2 + k_3 g'_3 + k_4 g'_4 + k_5 g'_5 + k_9 g'_9\}.$$

To describe $\mathbf{F}_{t_d}$, we want to describe the defining equations of $\mathbf{F}_{t_d}$. Each such equation is of the form $\langle g, t_d \rangle = 0$, where g is orthogonal to $\mathbf{F}_{t_d}$. For any such g, let g' be its extension to a vector in R^{91} by adding zero components. Then $g' \perp \mathbf{F}_{t_{\tilde{d}}}$, which implies that $g' \in G$. Therefore, we can find g by finding all vectors $g' \in G$ that vanish on all added components. This yields a system of linear equations in $k_1, \ldots, k_5, k_9$. We claim that all solution must satisfy $k_1 = k_3 = k_4 = k_5 = 0$. Indeed, the coefficient of any triple or quadruple interaction must vanish (since these don't belong to the original Ising model), which implies $k_1 = k_3 = k_4 = 0$, and also the coefficient of $t_{37,52}$ must vanish, which implies $k_5 = 0$. On the other hand, the vectors g'_2 and g'_9 only contain interactions that are already present in Δ, and so coefficients k_2 and k_9 are free. Thus the equations for $\mathbf{F}_{t_d}$ are

$$\begin{cases} \langle g_2,\ t_{\tilde{\beta}} \rangle = t_{87} - t_{56,87} & = 0, \\ \langle g_9,\ t_{\tilde{\gamma}} \rangle = t_{23} - t_{23,53} & = 0. \end{cases} \tag{10.2.7}$$

This is the same as the outer approximation $\mathbf{F}_2$.

The republican simplicial complex Δ_r consists of 20 variables, and the model induced from Δ_r contains 46 parameters. The size of the design matrix A_{Δ_r} is $2^{20} \times 46$, which is also too large to directly compute F_t. The yellow nodes in Figure 10.4a separate Δ_r into two simplicial complexes denoted (from left to right) by Δ_a and Δ_b. To compute the inner approximation, we complete the

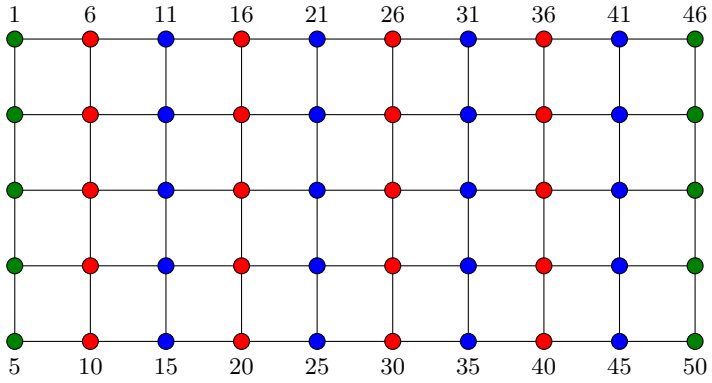


Figure 10.5: The two sets of separators used to get the inner approximation F_1 to F_t are represented by the red and blue nodes respectively

yellow separators and we get two new simplicial complexes $\Delta_{\tilde{a}}$ and $\Delta_{\tilde{b}}$. With the linear programming algorithm, we find that the corresponding data $t_{\tilde{a}}$ and $t_{\tilde{b}}$ lie in the relative interior of the polytopes $\mathbf{P}_{\Delta_{\tilde{a}}}$ and $\mathbf{P}_{\Delta_{\tilde{a}}}$, respectively. Therefore we have $\mathbf{F}_1 = \mathbf{P}_{\Delta_r}$. Since $\mathbf{F}_1 \subseteq \mathbf{F}_t \subseteq \mathbf{P}_{\Delta_r}$, we conclude that the corresponding data vector t_r lies in the relative interior of $\mathbf{P}_{\Delta_r}$.

10.2.2 The 5×10-grid graph

Let Δ be the simplicial complex of the 5×10 grid graph. We exploit the regularity of this graph and make use of the vertical separators in the grid to obtain inner and outer approximations of the facial sets. The graph has 50 nodes, which makes it too large to directly compute a facial set or even to store it. However, the 5×10 grid has 8 vertical separators marked in red and blue in Figure 10.5, and we can use these to approximate F_t. Since facial sets for 5×3-grids can be computed reasonably fast (3 to 4 seconds on a laptop with 2.50 GHz processor and 12 GB memory), we only use three of these vertical separators at a time, say the blue separators

$$S_2 = \{11, \ldots, 15\},\ S_4 = \{21, \ldots, 25\},\ S_6 = \{31, \ldots, 35\},\ S_8 = \{41, \ldots, 45\}.$$

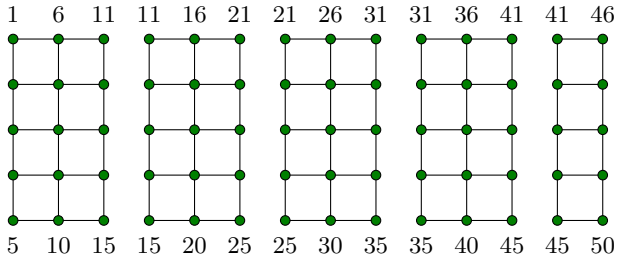


Figure 10.6: Five induced subgraphs

These separators separate the vertex sets

$$V_1 = \{1, \ldots, 15\},\ V_3 = \{11, \ldots, 25\},\ V_5 = \{21, \ldots, 35\},$$

$$V_7 = \{31, \ldots, 45\},\ V_9 = \{41, \ldots, 50\}.$$

Adding the blue separators to Δ gives a simplicial complex

$$\Delta_{S_2;S_4;S_6;S_8} := \Delta \bigcup_{j=2,4,6,8} \{F : F \subseteq S_j\}$$

with five irreducible components supported on the vertex sets V_1, V_3, V_5, V_7 and V_9 (Figure 10.7). To compute a facial set with respect to $\Delta_{S_2;S_4;S_6;S_8}$, according to Lemma 7.1.5, we need to compute

$$G_{1,V_1} := F_{\Delta_{S_2}|_{V_1}}(\pi_{V_1}(I_+)), \quad G_{1,V_3} := F_{\Delta_{S_2;S_4}|_{V_3}}(\pi_{V_3}(I_+)),$$

$$G_{1,V_5} := F_{\Delta_{S_4;S_6}|_{V_5}}(\pi_{V_5}(I_+)), \quad G_{1,V_7} := F_{\Delta_{S_6;S_8}|_{V_7}}(\pi_{V_7}(I_+)),$$

$$G_{1,V_9} := F_{\Delta_{S_8}|_{V_9}}(\pi_{V_9}(I_+)).$$

Then $G_1 := \bigcap_i \pi_{V_i}^{-1}(G_{1,V_i})$ is equal to $F_{\Delta_{S_2;S_4;S_6;S_8}}(I_+)$, and thus an inner approximation of F_t. As stated before, we do not need to compute G_1 explicitly, but we represent it by means of the G_{1,V_i}.

We can farther improve the approximations by also considering the red separators

$$S_1 = \{6, \ldots, 10\},\ S_3 = \{16, \ldots, 20\},\ S_5 = \{26, \ldots, 30\},\ S_7 = \{36, \ldots, 40\},$$

that separate

$$V_0 = \{1, \ldots, 10\},\ V_2 = \{6, \ldots, 20\},\ V_4 = \{16, \ldots, 30\},$$

$$V_6 = \{26, \ldots, 40\},\ V_8 = \{36, \ldots, 50\}.$$

As explained in Section 7.2, we want to compute $G_1^{(2)} := F_{\Delta_{S_1;S_3;S_5;S_7}}(G_1)$. Again, instead of computing $G_1^{(2)}$ directly, we need only compute the much smaller sets $G_{1,V_0}^{(2)} := \pi_{V_0}(G_1^{(2)}), G_{1,V_2}^{(2)} := \pi_{V_2}(G_1^{(2)}), \ldots, G_{1,V_8}^{(2)} := \pi_{V_8}(G_1^{(2)})$. So the question is: Is it possible to compute $G_{1,V_0}^{(2)}$, $G_{1,V_2}^{(2)}$, ..., $G_{1,V_8}^{(2)}$ from $G_{1,V_1}, G_{1,V_3}, \ldots, G_{1,V_9}$, without computing G_1 in between?

It turns out that this is indeed possible: By Lemma 7.1.5, all we need to compute $G_{1,V_i}^{(2)}$ is $G_{1,V_j} := \pi_{V_j}(G_1),\ j = i-1, i+1$. For $i = 0$, since $V_0 \subset V_1$, we can compute G_{1,V_0} from $\pi_{V_1}(G_1) = G_{1,V_1}$. For $i = 2, 4, 6, 8$, since $V_i \subset V_{i-1} \cup V_{i+1}$, we can compute G_{1,V_i} from $\pi_{V_{i-1} \cup V_{i+1}}(G_1)$, which itself can be obtained by "gluing" $\pi_{V_{i-1}}(G_1) = G_{1,V_{i-1}}$ and $\pi_{V_{i+1}}(G_1) = G_{1,V_{i+1}}$:

$$\pi_{V_{i-1} \cup V_{i+1}}(G_1) = \left(\pi_{V_{i-1}}^{V_{i-1} \cup V_{i+1}}\right)^{-1}(G_{1,V_{i-1}}) \cap \left(\pi_{V_{i+1}}^{V_{i-1} \cup V_{i+1}}\right)^{-1}(G_{1,V_{i+1}}),$$

where $\pi_{V''}^{V'}$ for $V'' \subseteq V'$ denotes the marginalization map from $I_{V'}$ to $I_{V''}$ and where $\left(\pi_{V''}^{V'}\right)^{-1}$ denotes the lifting from $I_{V''}$ to $I_{V'}$.

As explained in Section 7.2, we have to iterate this procedure: From $G_1^{(2)}$ we want to compute $G_1^{(3)} := F_{\Delta_{S_2;S_4;S_6;S_8}}(G_1')$ or, more precisely, we want to compute $G_{1,V_i}^{(3)} = \pi_{V_i}(G_1^{(3)})$ for $i = 1, 3, \ldots, 9$. Again, we do this without looking at $G_1^{(2)}$ directly just by using the information provided by the $G_{1,V_i}^{(3)}$. Iterating this procedure, we obtain a sequence of sets $G_{1,V_i}^{(k)}, G_{1,V_j}^{(k)}$ (with odd i and even j), which stabilizes after a finite number of steps. Let

$$F_{1,V_i} := \bigcup G_{1,V_i}^{(k)},$$

Our best inner approximation is then $F_1 = \bigcap_{i=0}^{9} \pi_{V_i}^{-1}(F_{1,V_i})$. Again, we do not compute F_1 explicitly, but we represent it in terms of the F_{1,V_i}.

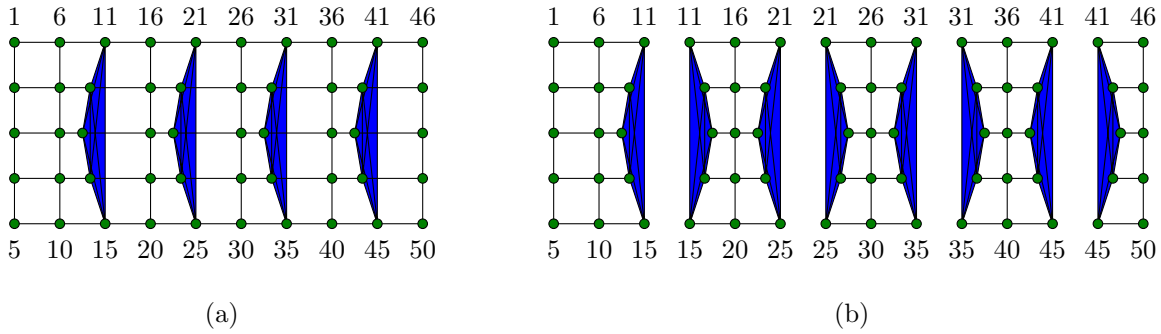


Figure 10.7: (a) The 5×10 grid graph with the blue separators completed. (b) The five irreducible subcomplexes after completion of the separators.

The process is visually represented in Figure 10.8.

Let us now consider the outer approximation F_2. We adapt Strategy 3 of Section 7.3 and cover the graph with 5×3 grid subgraphs, since the facial sets for such graphs can easily be computed. These subgrids are supported on the same vertex subsets $V_i, i = 1, \ldots, 8$ as used when computing F_1. This makes it possible to compare F_1 and F_2. For $i = 1, 3, \ldots, 8$ we compute $F_{2,V_i} = F_{\Delta|_{V_i}}(\pi_{V_i}(I_+))$. Our outer approximation is then $F_2 = \bigcap_i \pi_{V_i}^{-1}(F_{2,V_i})$. Again, we don't compute F_2 explicitly, but we only store F_{2,V_i} in a computer as a representation of F_2. To compare the two approximations F_1 and F_2, we need only compare their projections F_{1,V_i} and F_{2,V_i} pairwise, $i = 1, \ldots, 8$. We generated random data of varying sample size. For each fixed sample size, we generated 100 data samples. The simulation results are show in Table 10.6. For each simulated sample, we compute the sets F_{1,V_i} and F_{2,V_i} as described above. When computing F_{1,V_i}, we found that 2 iterations actually suffice. Then we checked whether F_2 is a proper subset of I (second column), and we checked whether $F_1 = F_2$ (third column). Both for small and large sample sizes, we found that $F_1 = F_2$ quite often.

Table 10.6: Facial set approximation for the 5×10 grid graph

sample size	$F_2 \neq I$	$F_1 = F_2$
50	100.0%	94.3%
100	100.0%	82.5%
150	99.9%	76.5%
200	99.6%	81.2%
300	96.4%	87.7%
400	92.9%	91.5%
500	84.8%	93.9%
1000	44.7%	99.9%

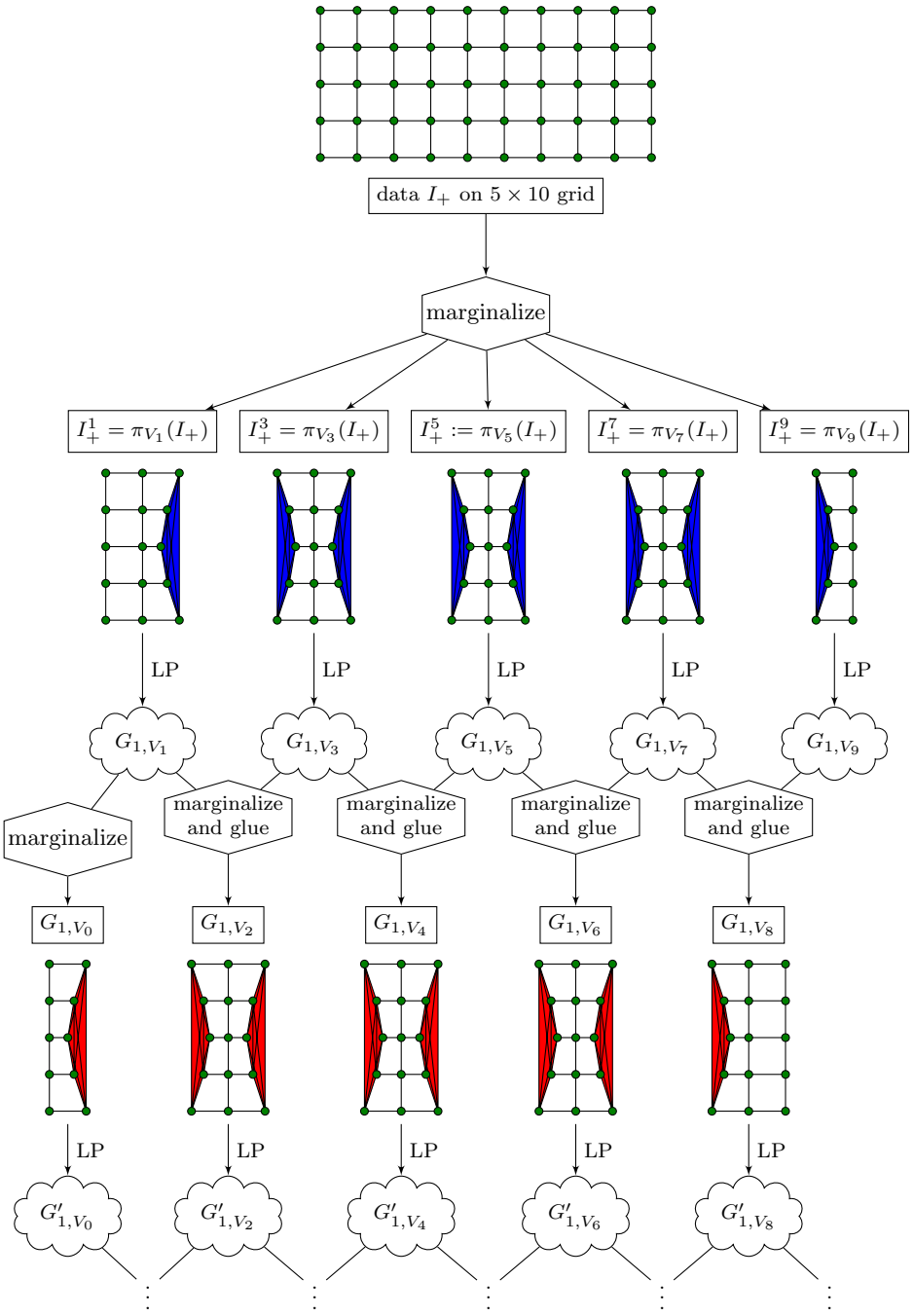


Figure 10.8: Flow chart describing the steps leading to the inner approximation

We also investigated what happens when the outer approximation is not computed using all 3×5-subgrids, but only a cover of four 3×5-subgrids and one 2×5-subgrid (as in Figure 10.6). In all our simulations, this easier approximation gave the same result. The same is not true for the

inner approximation: When using just one of the two families of parallel separators we obtain an inner approximation that is much too small.

11 Conclusion

In this thesis, we studied hierarchical log-linear models. We made two main contributions to this topic. First, we studied different types of composite likelihoods and succeeded in parameter estimation of high-dimensional log-linear models. We proved nice asymptotic properties of our estimates both when the dimension of data p is fixed and also when $p \to \infty$. As the dimension of statistical problems grows rapidly and sometimes the sample size is not sufficiently large, or even smaller than p, our asymptotic property when $p \to \infty$ is more relevant for big data analysis. Second, we studied the existence of the MLE by finding the smallest facial set of the marginal polytope of the hierarchical log-linear model. When the dimension of the marginal polytope is very large, we propose proper inner and outer approximations. Most of the time our approximations can capture the smallest face of the sufficient statistic, which is the real space the data fall into.

Throughout our research, we assume that the hierarchical log-linear model structure is known as a prior knowledge. For real data examples, we apply the $l1$–penalized logistic regression method proposed by Ravikumar et al. (2011) for finding the model structure. A problem of this method is that the logistic regression only gave the neighbours of a vertex, it didn't take 3-way or high-way interactions among variables into consideration. The model learning problem is still a difficult task to accomplish in the areas of hierarchical log-linear and graphical models. In Gaussian graphical model literature, researchers proposed various Bayesian structure learning algorithms, but we didn't

see much work in the discrete graphical model field. In terms of the prior distribution for the parameters in hierarchical log-linear models, we can use the conjugate prior distribution given by Massam et al. (2009), but we still need to think about the graph structure search algorithms. This will be the direction of our future work, and the research in Gaussian graphical models can give us a good point of departure.

Bibliography

Agresti, A. and Kateri, M. (2011). *Categorical data analysis*. Springer.

Asuncion, A. U., Liu, Q., Ihler, A. T., and Smyth, P. (2010). Learning with blocks: Composite likelihood and contrastive divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 33–40.

Banerjee, O., Ghaoui, L. E., and dAspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516.

Barndorff-Nielsen, O. (1978). Information and exponential families in statistical theory.

Barndorff-Nielsen, O. (2014). *Information and exponential families in statistical theory*. John Wiley & Sons.

Bartlett, M. S. (1935). Contingency table interactions. *Supplement to the Journal of the Royal Statistical Society*, 2(2):248–252.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The statistician*, pages 179–195.

Birch, M. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 220–233.

Bishop, Y., Fienberg, S., and Holland, P. (1975a). Discrete multivariate analysis: Theory and practice. 1975. *Cambridge: The Massachusetts Institute of Technology Press Google Scholar*.

Bishop, Y. M., Fienberg, S. E., and Holland, P. . (1975b). Discrete multivariate analysis: Theory and practice.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial intelligence*, 42(2-3):393–405.

Csiszár, I. and Matúš, F. (2005). Closures of exponential families. *Annals of probability*, pages 582–600.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.

Dillon, J. V. and Lebanon, G. (2010). Stochastic composite likelihood. *The Journal of Machine Learning Research*, 11:2597–2633.

Dobra, A., Lenkoski, A., et al. (2011). Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993.

Eriksson, N., Fienberg, S. E., Rinaldo, A., and Sullivant, S. (2006). Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models. *Journal of Symbolic Computation*, 41(2):222–233.

Fienberg, S. E. (1970). Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *Journal of the American Statistical Association*, 65(332):1610–1616.

Fienberg, S. E. and Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference*, 137(11):3430–3445.

Fienberg, S. E. and Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, 40(2):996–1023.

Geyer, C. J. (1991). Markov chain monte carlo maximum likelihood.

Geyer, C. J. et al. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, 3:259–289.

Haberman (1974a). The analysis of frequency data. *Illinois: University of Chicago Press*.

Haberman, S. J. (1974b). *The analysis of frequency data*, volume 194. University of Chicago Press Chicago.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30.

Jirousek, R. and Preucil, S. (1995). On the effective implementation of the iterative proportional fitting procedure. *Computational Statistics & Data Analysis*, 19(2):177–189.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Letac, G. and Massam, H. (2012). Bayes factors and the geometry of discrete hierarchical loglinear models. *Annals of Statistics*.

Letac, G., Massam, H., et al. (2012). Bayes factors and the geometry of discrete hierarchical loglinear models. *The Annals of Statistics*, 40(2):861–890.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary mathematics*, 80(1):221–39.

Liu, Q. and Ihler, A. (2012). Distributed parameter estimation via pseudo-likelihood. *arXiv preprint arXiv:1206.6420*.

Massam, H., Liu, J., Dobra, A., et al. (2009). A conjugate prior for discrete hierarchical log-linear models. *The Annals of Statistics*, 37(6A):3431–3467.

Massam, H. and Wang, N. (2013). Distributed parameter estimation of discrete hierarchical models via marginal likelihoods. *arXiv preprint arXiv:1310.5666*.

Massam, H. and Wang, N. (2015). A local approach to estimation in discrete loglinear models. *arXiv preprint arXiv:1504.05434*.

Meng, Z., Wei, D., Wiesel, A., et al. (2013). Distributed learning of gaussian graphical models via marginal likelihoods. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 39–47.

Meng, Z., Wei, D., Wiesel, A., and Hero, A. O. (2014). Marginal likelihoods for distributed parameter estimation of gaussian graphical models. *Signal Processing, IEEE Transactions on*, 62(20):5425–5438.

Nocedal, J. (1980). Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782.

Peterson, C. (1987). A mean field theory learning algorithm for neural networks. *Complex systems*, 1:995–1019.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B., et al. (2011). High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Roth, D. (1996). On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1-2):273–302.

Roy, S. and Kastenbaum, M. A. (1956). On the hypothesis of no" interaction" in a multi-way contingency table. *The Annals of Mathematical Statistics*, 27(3):749–757.

Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of artificial intelligence research*, 4(1):61–76.

Schmidt, M. (2005). minfunc: Unconstrained differentiable multivariate optimization in matlab (2012 version), http://www.di.ens.fr/ mschmidt/software/minfunc.html.

Schmidt, M. W., Berg, E., Friedlander, M. P., and Murphy, K. P. (2009). Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *International Conference on Artificial Intelligence and Statistics*, page None.

Sutton, C. and McCallum, A. (2007). Piecewise pseudolikelihood for efficient training of conditional random fields. In *Proceedings of the 24th international conference on Machine learning*, pages 863–870. ACM.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42.

Wainwright, M. and Jordan, M. (2003). Graphical models, exponential families, and variational inference. uc berkeley, dept. Technical report, of Statistics, Technical Report 649.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

Wiesel, A. and Hero, A. O. (2012). Distributed covariance estimation in gaussian graphical models. *IEEE Transactions on Signal Processing*, 60(1):211–220.

A Three properties of matrix eigenvalues

The following two lemmas about the eigenvalue of rank one matrices have trivial proofs.

Lemma A.0.1. *A matrix $u \otimes u$ where u is a vector of dimension $|J|$ has only one non-zero eigenvalue, which is equal to $\|u\|_F^2$.*

Lemma A.0.2. *Let a, b be two vectors of same dimension J. The matrix $a \otimes b$ has rank one, and therefore has only one nonzero eigenvalue whose value is $\langle a, b \rangle$.*

Lemma A.0.3. *If A, B, C are three square matrices such that $A = B + C$, then we have the classical inequality for minimum eigenvalue $\lambda_{min}(A) \geq \lambda_{min}(B) + \lambda_{min}(C)$. We also have the inequality:*

$$\lambda_{min}(A) \geq \lambda_{min}(B) - \|C\|_2,$$

where $\|C\|_2$ is the operator norm of C.

Proof. We need only prove the second inequality.

$$\lambda_{min}(B) = \min_{\|x\|_2=1} x^{'} B x = \min_{\|x\|_2=1} \{x^{'} A x + x^{'}(-C)x\} \leq y^{'} A y + y^{'}(-C)y, \ \forall y \text{ such that } ||y|| = 1.$$

Let y_0 be the unit-norm eigenvector of A corresponding to the minimum eigenvalue of A. Then since $y_0^{'}(-C)y_0 \leq \max_{||z||=1} z^{'}(-C)z$,

$$\begin{aligned} y_0^t A y_0 \quad &= \lambda_{min}(A) \geq \lambda_{min}(B) - y_0^{'}(-C)y_0 \geq \lambda_{min}(B) - \max_{\|z\|_2=1} z^{'}(-C)z \\ &= \lambda_{min}(B) - \lambda_{max}(-C) \\ \geq \quad &\lambda_{min}(B) - \|-C\|_2 = \lambda_{min}(B) - \|C\|_2, \end{aligned}$$

where the last inequality is due to the fact that $\lambda_{max}(-C) \leq \|-C\|_2$ and the lemma is proved. $\square$

B Some proofs

B.1 Proof of Lemma 4.1.1

We will use the notation $j \lhd_0 j'$ to mean that $j \lhd j'$ or $j = 0$, the zero cell. Let $p^{\mathcal{M}_v}(i)$ denote the marginal probability of $i \in I_{\mathcal{M}_v}$. We know that the $\mathcal{M}_v$-marginal distribution of $X_{\mathcal{M}_v}$ is multinomial. By the general parametrization of the multinomial model (2.1.7), for $j \in J, S(j) \subset \mathcal{M}_v$, since $S(j)$ is complete,

$$\theta_j^{\mathcal{M}_v} = \sum_{j' \in J,\, j' \lhd j} (-1)^{|S(j)|-|S(j')|} \log \frac{p^{\mathcal{M}_v}(j')}{p^{\mathcal{M}_v}(0)}, \tag{B.1.1}$$

where by abuse of notation, j such that $S(j) \subset \mathcal{M}_v$ is considered as an element of $I_{\mathcal{M}_v}$.

Moreover,

$$\begin{aligned} p^{\mathcal{M}_v}(j) &= \sum_{i \in I:\, i_{\mathcal{M}_v} = j} p(i) = \sum_{i \in I,\, i_{\mathcal{M}_v} = j} \exp\{ \sum_{j' \mid j' \lhd_0 j} \theta_{j'} + \sum_{\substack{j' \mid j' \lhd i \\ j' \not\lhd j \\ j'_{\mathcal{M}_v} \lhd_0 j}} \theta_{j'} \} \\ &= \Big(\exp \sum_{j' \mid j' \lhd_0 j} \theta_{j'} \Big) \Big(1 + \sum_{i \in I,\, i_{\mathcal{M}_v} = j} \exp \sum_{\substack{j' \mid j' \lhd i \\ j' \not\lhd j \\ j'_{\mathcal{M}_v} \lhd_0 j}} \theta_{j'} \Big) . \end{aligned}$$

Therefore $\log p^{\mathcal{M}_v}(j) = \sum_{j' \mid j' \lhd_0 j} \theta_{j'} + \log \Big(1 + \sum_{i \in I,\, i_{\mathcal{M}_v} = j} \exp \sum_{\substack{j' \mid j' \lhd i \\ j' \not\lhd j}} \theta_{j'} \Big)$, which we can write

$$\sum_{j' \mid j' \lhd_0 j} \theta_{j'} = \log p^{\mathcal{M}_v}(j) - \log \Big(1 + \sum_{i \in I,\, i_{\mathcal{M}_v} = j} \exp \sum_{\substack{k \mid k \lhd i \\ k \not\lhd j}} \theta_k \Big) . \tag{B.1.2}$$

Moebius inversion formula states that for $a \subseteq V$ an equality of the form $\sum_{b \subseteq a} \Phi(b) = \Psi(a)$ is equivalent to $\Phi(a) = \sum_{b \subseteq a} (-1)^{|a \setminus b|} \Psi(b)$. Here, using a generalization of the Moebius inversion

formula to the partially ordered set given by $\triangleleft$ on J, we derive from (B.1.2) that for $j \in J^{\mathcal{M}_v} \subset J$

$$\begin{aligned}\theta_j &= \sum_{j' \mid j' \triangleleft_0 j} (-1)^{|S(j)-S(j')|} \log p^{\mathcal{M}_v}(j') \\ &\quad - \sum_{j' \mid j' \triangleleft_0 j} (-1)^{|S(j)-S(j')|} \log \Big(1 + \sum_{i \in I,\, i_{\mathcal{M}_v} = j'} \exp \sum_{\substack{k \mid k \triangleleft i \\ k \not\triangleleft j'}} \theta_k\Big) \\ &= \theta_j^{\mathcal{M}_v} - \sum_{j' \mid j' \triangleleft_0 j} (-1)^{|S(j)-S(j')|} \log \Big(1 + \sum_{i \in I,\, i_{\mathcal{M}_v} = j'} \exp \sum_{\substack{k \mid k \triangleleft i \\ k \not\triangleleft j'}} \theta_k\Big) \qquad \text{(B.1.3)}\end{aligned}$$

which we prefer to write as (4.1.9).

B.2 Proof of lemma 4.1.2

Since (4.1.9) is already proved, statement *(2.)* holds. Let us prove that statement *(1.)* holds, i.e., that when $S(j) \not\subset \mathcal{B}_v$, the alternating sum on the right-hand side of (4.1.9) is equal to 0. Since $j \in J$, $S(j)$ is necessarily complete and $j' \triangleleft j$ is obtained by removing one or more vertices from $S(j)$.

If $S(j) \cap \mathcal{B}_v \neq \emptyset$ but $S(j) \not\subset \mathcal{B}_v$, there is at least one vertex $w \in S(j)$ which is not in $\mathcal{B}_v$. Let l_0 and l_w be the log terms in the alternating sum corresponding to $j' = 0$ and $j'_w \triangleleft j$ such that $S(j'_w) = \{w\}$ respectively. Since for any neighbours u of w in $\mathcal{M}_v$ and for any $i \in I$ such that $i_{\mathcal{M}_v} = j'$, the u-th coordinate i_u must be zero and since w cannot have a neighbour outside $\mathcal{M}_v$, the set $\{\theta_k, k \triangleleft i^{(1)}, k \not\triangleleft j'\}$ in l_0 for $i^{(1)}$ such that $i^{(1)}_{\mathcal{M}_v} = 0$ is the same as the set $\{\theta_k, k \triangleleft i^{(2)}, k \not\triangleleft j'\}$ in l_w for $i^{(2)}$ such that $i^{(2)}_{\mathcal{M}_v} = j'_w$ and $i^{(2)}_{V\setminus\mathcal{M}_v} = i^{(1)}_{V\setminus\mathcal{M}_v}$. The terms in l_0 and l_w in (4.1.9) are therefore exactly the same except for their sign, and these two terms cancel out. Similarly, for any given $j' \triangleleft j$ with $w \notin S(j')$, let $j'_w \in J$ be such that $S(j'_w) = S(j) \cup \{w\}$ and $j'_w \triangleleft j$, then, the set $\theta_k, k \triangleleft i^{(1)}, k \not\triangleleft j'$ in $l_{j'}$ and the set $\theta_k, k \triangleleft i^{(2)}, k \not\triangleleft j'_w$ in $l_{j'_w}$ are identical where, similarly to the argument above, $i^{(1)}$ is such that $i^{(1)}_{\mathcal{M}_v} = j'$ and $i^{(2)}$ is such that $i^{(1)}_{\mathcal{M}_v} = j'_w$ and $i^{(2)}_{V\setminus\mathcal{M}_v} = i^{(1)}_{V\setminus\mathcal{M}_v}$. Therefore the terms $l_{j'}$ and $l_{j'_w}$ cancel out and *(1.)* is proved.

To prove that *(3.)* holds, following (2.1.7), we have, for $S(i) = E \subset \mathcal{M}_v$

$$\begin{aligned}
\theta_i^{\mathcal{M}_v} &= \sum_{F\subset E}(-1)^{|E\setminus F|}\log p^{\mathcal{M}_v}(i_F, 0_{\mathcal{M}_v\setminus F}) \\
&= \sum_{F\subset E}(-1)^{|E\setminus F|}\log\Big(p(i_F,0_{V\setminus F}) + \sum_{L\subset V\setminus\mathcal{M}_v}\sum_{k_L\in I_L} p(i_F, 0_{\mathcal{M}_v\setminus F}, k_L, 0_{V\setminus(\mathcal{M}_v\cup L)})\Big) \\
&= \sum_{F\subset E}(-1)^{|E\setminus F|}\log\Big(\exp(\sum_{j\in J, j\lhd i_F}\theta_j) + \sum_{L\subset V\setminus F}\sum_{k_L\in I_L}\exp(\sum_{j\in J, j\lhd i_F}\theta_j + \sum_{j\not\lhd i_F, j\lhd(i_F,k_L)}\theta_j)\Big) \\
&= \sum_{F\subset E}(-1)^{|E\setminus F|}\log\Big(\exp(\sum_{j\in J, j\lhd i_F}\theta_j)\Big) \qquad \text{(B.2.1)} \\
&\quad + \sum_{F\subset E}(-1)^{|E\setminus F|}\log(1 + \sum_{L\subset V\setminus F}\sum_{k_L\in I_L}\exp(\sum_{j\not\lhd i_F, j\lhd(i_F,k_L)}\theta_j)\Big) \\
&= \theta_i + \sum_{F\subset E}(-1)^{|E\setminus F|}\log(1 + \sum_{L\subset V\setminus F}\sum_{k_L\in I_L}\exp(\sum_{j\not\lhd i_F, j\lhd(i_F,k_L)}\theta_j)\Big) \qquad \text{(B.2.2)}
\end{aligned}$$

Now, following an argument similar to that of *(1.)* above, we can show that the second component of the sum in (B.2.2) is equal to zero. It follows that when $\theta_i = 0$, we have $\theta_i^{\mathcal{M}_v} = 0$. This completes the proof of Lemma 4.1.2.

B.3 Proof of Theorem 4.3.1

The local relaxed marginal log likelihood is

$$\begin{aligned}
l^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}}) &= \sum_{k=1}^{N}\log p^{\mathcal{M}_{l,v}}(X_{\mathcal{M}_v} = i_{\mathcal{M}_v}^{(k)}) = \sum_{i_{\mathcal{M}_v}\in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v})\log p^{\mathcal{M}_{l,v}}(i_{\mathcal{M}_v}) \\
&= \langle\theta^{\mathcal{M}_{l,v}}, t^{\mathcal{M}_{l,v}}\rangle - N k^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}})
\end{aligned}$$

It is immediate to see that $\frac{\partial l^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}})}{\partial\theta_j} = t(j) - p^{\mathcal{M}_{l,v}}(j_{S(j)})$ where $p^{\mathcal{M}_{l,v}}(j_{S(j)})$ denotes the $j_{S(j)}$-marginal cell probability in the $\mathcal{M}_{l,v}$-marginal model. Therefore the likelihood equations $\frac{\partial l^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}})}{\partial\theta_j} = 0,\ j \in J^{\mathcal{M}_{l,v}}$ yield

$$t(j) - p^{\mathcal{M}_{l,v}}(j_{S(j)}) = 0, \qquad \text{(B.3.1)}$$

where $t(j) = n(j_{S(j)})$.

The following proof stands both in the case of one-hop and two-hop neighbourhood. We present it for the more general case of the two hop neighbourhood. The local conditional log likelihood is

$$\begin{aligned} l^{v,2PS}(\theta^{v,2PS}) &= \sum_{i_{\mathcal{M}_v} \in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v}) \log \frac{p(X_v = i_v, X_{\mathcal{N}_v} = i_{\mathcal{N}_v}, X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})}{p(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})} \\ &= \sum_{i_{\mathcal{M}_v} \in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v}) \log \frac{p^{\mathcal{M}^{2,v}}(X_{\mathcal{M}_v} = i_{\mathcal{M}_v})}{p^{\mathcal{M}^{2,v}}(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})} \\ &= \sum_{i_{\mathcal{M}_v} \in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v}) \log p^{\mathcal{M}^{2,v}}(X_{\mathcal{M}_v} = i_{\mathcal{N}_v}) - \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \log p^{\mathcal{M}^{2,v}}(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}) \\ &= l^{\mathcal{M}_{2,v}}(\theta^{\mathcal{M}_{2,v}}) - \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \log \sum_{x_{v \cup \mathcal{N}_v} \in I_{v \cup \mathcal{N}_v}} p^{\mathcal{M}^{2,v}}(X_{v \cup \mathcal{N}_v} = x_{v \cup \mathcal{N}_v}, X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}) \\ &= l^{\mathcal{M}_{2,v}}(\theta^{\mathcal{M}_{2,v}}) - Q \end{aligned} \tag{B.3.2}$$

where

$$Q = \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \log \sum_{x_{v \cup \mathcal{N}_v} \in I_{v \cup \mathcal{N}_v}} \exp\Big(\theta_0 + \sum_{\substack{k \lhd (x_{v \cup \mathcal{N}_v}, i_{\mathcal{N}_{2v}}) \\ k \in J^{\mathcal{M}_{2,v}}}} \theta_k\Big) \tag{B.3.3}$$

and $\theta_0 = -\log(\sum_{i_{\mathcal{M}_v} \in I_{\mathcal{M}_v}} \exp \sum_{k \lhd i_{\mathcal{M}_v}, k \in J^{\mathcal{M}_{2,v}}} \theta_k)$. The second equality above is due to the fact that in the expression (4.1.3) of $\frac{p(X_v = i_v, X_{\mathcal{N}_v} = i_{\mathcal{N}_v}, X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})}{p(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})}$, the θ_j such that $S(j) \notin \mathcal{M}_v$ and the θ_j such that $S(j) \subset \mathcal{N}_{2v}$ cancel out from the numerator and denominator, and it therefore does not matter, for the conditional distribution of $X_{v \cup \mathcal{N}_v}$ given $X_{\mathcal{N}_{2v}}$, what the relationship between the neighbours are. The only thing that matters is the relationship between the vertices in $v \cup \mathcal{N}_v$, and the vertices in $\mathcal{M}_v$, and according to Lemma 4.1.2, that remains unchanged when we change from the global model to the $\mathcal{M}_{2,v}$-marginal models.

We now differentiate the expression of $l^{v,2PS}$ in (B.3.3) with respect to $\theta_j, j \in J^{\mathcal{M}_{2,v}}$. We first note that

$$\frac{\partial \theta_0}{\partial \theta_j} = p^{\mathcal{M}^{2,v}}(j_{S(j)}).$$

If we use the notation

$$\mathbf{1}_{j\lhd(x_{v\cup\mathcal{N}_v},i_{\mathcal{N}_{2v}})} = \begin{cases} 1 & \text{if } j \lhd (x_{v\cup\mathcal{N}_v}, i_{\mathcal{N}_{2v}}) \\ 0 & \text{otherwise} \end{cases},$$

and the notation $p^{\mathcal{M}_{2,v}}(i_E),\ E \subset \mathcal{M}_v$ to denote the marginal probability of $X_E = i_E$ in the $\mathcal{M}_{2,v}$-marginal model, we have

$$\frac{\partial Q}{\partial \theta_j} = \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \frac{\sum_{x_{v\cup\mathcal{N}_v} \in I_{v\cup\mathcal{N}_v}} p^{\mathcal{M}^{2,v}}(x_{v\cup\mathcal{N}_v}, i_{\mathcal{N}_{2,v}}) \Big(\mathbf{1}_{j\lhd(x_{v\cup\mathcal{N}_v},i_{\mathcal{N}_{2v}})} - p^{\mathcal{M}^{2,v}}(j_{S(j)}) \Big)}{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2,v}})}.$$

If $j \in J^{\mathcal{M}_{2,v}}$ is such that $S(j) \subset \mathcal{N}_{2v}$, then $\mathbf{1}_{j\lhd(x_{v\cup\mathcal{N}_v},i_{\mathcal{N}_{2v}})} = \mathbf{1}_{j_{\mathcal{N}_{2v}}\lhd i_{\mathcal{N}_{2v}}}$ and

$$\begin{aligned} \frac{\partial Q}{\partial \theta_j} &= \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \frac{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}}) \Big(\mathbf{1}_{j_{\mathcal{N}_{2,v}}\lhd i_{\mathcal{N}_{2v}}} - p^{\mathcal{M}^{l,v}}(j_{S(j)}) \Big)}{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}})} \\ &= \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \Big(\mathbf{1}_{j_{\mathcal{N}_{2,v}}\lhd i_{\mathcal{N}_{2,v}}} - p^{\mathcal{M}^{2,v}}(j_{S(j)}) \Big) \\ &= n(j_{S(j)}) - N p^{\mathcal{M}^{2,v}}(j_{S(j)}) \end{aligned}$$

At the MLE of the local $\mathcal{M}_{l,v}$ model, from standard likelihood equations (see Lauritzen, 1996, Theorem 4.11), we have $\hat{p}^{\mathcal{M}^{l,v}}(j_{S(j)}) = \frac{n(j_{S(j)})}{N}$ and therefore

$$\frac{\partial Q}{\partial \theta_j} = 0, \quad j \in J^{\mathcal{M}_{2,v}},\ S(j) \subset \mathcal{N}_{2v}. \tag{B.3.4}$$

If $j \in J^{\mathcal{M}_{2,v}}$ is such that $S(j) \not\subset \mathcal{N}_{2v}$, i.e. if $j \in J^{v,2PS}$,

$$\begin{aligned} \frac{\partial Q}{\partial \theta_j} &= \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \frac{p^{\mathcal{M}^{2,v}}(j_{S(j)\cap(v\cup\mathcal{N}_v)}, i_{\mathcal{N}_{2v}}) \mathbf{1}_{j_{\mathcal{N}_{2v}}\lhd i_{\mathcal{N}_{2v}}} - p^{\mathcal{M}^{2,v}}(j_{S(j)}) p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}})}{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}})} \\ &= -p^{\mathcal{M}^{2,v}}(j_{S(j)}) \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) + \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} \frac{n(i_{\mathcal{N}_{2v}})}{p^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}})} p^{\mathcal{M}^{2,v}}(j_{S(j)\cap(v\cup\mathcal{N}_v)}, i_{\mathcal{N}_{2v}}) \mathbf{1}_{j_{\mathcal{N}_{2v}}\lhd i_{\mathcal{N}_{2v}}} \end{aligned}$$

Since in the $\mathcal{M}_{2,v}$-marginal model, all the vertices in $\mathcal{N}_{2,v}$ are connected by construction, at the MLE of the local $\mathcal{M}_{2,v}$ model, $\hat{p}^{\mathcal{M}^{2,v}}(i_{\mathcal{N}_{2v}}) = \frac{n(i_{\mathcal{N}_{2v}})}{N}$ and therefore

$$\begin{aligned} \frac{\partial Q}{\partial \theta_j} &= -N p^{\mathcal{M}^{2,v}}(j_{S(j)}) + N \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} p^{\mathcal{M}^{2,v}}(j_{S(j)\cap(v\cup\mathcal{N}_v)}, i_{\mathcal{N}_{2v}}) \mathbf{1}_{j_{\mathcal{N}_{2v}}\lhd i_{\mathcal{N}_{2v}}} \\ &= -N p^{\mathcal{M}^{2,v}}(j_{S(j)}) + N p^{\mathcal{M}^{2,v}}(j_{S(j)}) = 0 \end{aligned} \tag{B.3.5}$$

It follows from (B.3.4) and (B.3.5) that the $2PS$ component of $\hat{\theta}^{\mathcal{M}_{2,v}}$, i.e.

$$\hat{\theta}_j^{\mathcal{M}_{2,v}}, j \in J^{2,PS}$$

is the MLE of the local two-hop conditional likelihood. We therefore have

$$\hat{\theta}^{v,2PS} = (\hat{\theta}^{\mathcal{M}_{2,v}})_{2PS}.$$

B.4 Proof of Theorem 5.1.1

Given the definition of $\bar{\theta}$, to show (5.1.1), we only need to show that

$$\sqrt{N}(\hat{\theta} - \tilde{\theta}^*) \to N(0, G)$$

where $\tilde{\theta}^*$ is the column vector obtained by stacking up $\theta^{*v}, v \in V$ into one column vector. Through a classical expansion of the local conditional likelihood function $l(\theta^v) = \sum_{k=1}^{N} l^{v,PS}(\theta^{v,PS}|X^{(k)})$, we have that

$$\sqrt{N}(\hat{\theta}^v - \tilde{\theta}^{*v}) = \frac{1}{\sqrt{N}} I^{-1}(\theta^{*v}) \sum_{k=1}^{N} \frac{\partial l(\theta^{*v}|X^{(k)})}{\partial \theta^{*v}} + R_N$$

where R_n tends to 0 in probability as $n \to +\infty$. Let $U_{v,k} = I^{-1}(\theta^{*v}) \frac{\partial l(\theta^{*v}|X^{(k)})}{\partial \theta^{*v}}$ and let U_k be the vector obtained by stacking up the vectors $U_{v,k}, v \in V$ into a column vector. For $\bar{U}_n = \sum_{k=1}^{N} U_k$, we can then write

$$\sqrt{N}(\hat{\theta}^v - \tilde{\theta}^{*v}) = \sqrt{N}\bar{U}_N + R_N.$$

Each vector $U_k, k = 1, \ldots, N$ clearly have mean 0 and covariance G, as defined in (5.1.2). It is immediate to show that G is finite. By the central limit theorem we thus have that $\sqrt{N}(\hat{\theta} - \tilde{\theta}^*) \to N(0, G)$ and $\sqrt{N}(\hat{\theta} - \theta^*) \to N(0, AGA^t)$. The asymptotic expression for (5.1.3) is also an immediate consequence of this asymptotic distribution.

B.5 Proof of Theorem 5.1.2

From standard asymptotic theory, we know that the asymptotic variance of θ is equal to

$$\Big(\frac{\partial^2 k^{\bar{\mathcal{M}}_{2,v}}}{\partial(\theta^{\bar{\mathcal{M}}_{2,v}})^2}\Big)^{-1} \tag{B.5.1}$$

evaluated at the corresponding true value of the parameter. It will be convenient in the sequel to represent the symmetric matrix $K = \frac{\partial^2 k^{\bar{\mathcal{M}}_{2,v}}}{\partial(\theta^{\bar{\mathcal{M}}_{2,v}})^2}$ according to the different blocks determined by the subvectors of $\theta^{\bar{\mathcal{M}}_{2,v}}$ as follows

$$K = \begin{pmatrix} K_{J_{1,v},J_{1,v}} & K_{J_{1,v},B_{1,v}} & K_{J_{1,v},J_{2\setminus 1,v}} & K_{J_{1,v},B_{2,v}} \\ K_{B_{1,v},J_{1,v}} & K_{B_{1,v},B_{1,v}} & K_{B_{1,v},J_{2\setminus 1,v}} & K_{B_{1,v},B_{2,v}} \\ K_{J_{2\setminus 1,v},J_{1,v}} & K_{J_{2\setminus 1,v},B_{1,v}} & K_{J_{2\setminus 1,v},J_{2\setminus 1,v}} & K_{J_{2\setminus 1,v},B_{2,v}} \\ K_{B_{2,v},J_{1,v}} & K_{B_{2,v},B_{1,v}} & K_{B_{2,v},J_{2\setminus 1,v}} & K_{B_{2,v},B_{2,v}} \end{pmatrix}.$$

We observe that in the $\bar{\mathcal{M}}^{2,v}$ model, the subset $\mathcal{B}_{1,v} \subset V$ separates $\{v\}$ from $V \setminus \mathcal{M}_{1,v}$ and the set $\mathcal{B}_{1,v}$ is complete. Therefore using a standard formula in graphical models, we have that

$$K^{-1} = \begin{pmatrix} K_{J_{1,v},J_{1,v}} & K_{J_{1,v},B_{1,v}} \\ K_{B_{1,v},J_{1,v}} & K_{B_{1,v},B_{1,v}} \end{pmatrix}^{-1} + \begin{pmatrix} K_{B_{1,v},B_{1,v}} & K_{B_{1,v},J_{2\setminus 1,v}} & K_{B_{1,v},B_{2,v}} \\ K_{J_{2\setminus 1,v},B_{1,v}} & K_{J_{2\setminus 1,v},J_{2\setminus 1,v}} & K_{J_{2\setminus 1,v},B_{2,v}} \\ K_{B_{2,v},B_{1,v}} & K_{B_{2,v},J_{2\setminus 1,v}} & K_{B_{2,v},B_{2,v}} \end{pmatrix}^{-1} - K^{-1}_{B_{1,v},B_{1,v}}$$

where matrices on the right-hand-side of the equation are "padded" with zeros in the appropriate blocks.

Let $\theta_{J_{1,v}} = (\theta_j, j \in J_{1,v})$, then the covariance matrix of $(\hat{\theta}^{\bar{\mathcal{M}}^{2,v}})_{J_{1,v}}$ is $[K^{-1}]_{J_{1,v}}$. From the previous expression of K^{-1}, we have

$$[K^{-1}]_{J_{1,v}} = \Big[\begin{pmatrix} K_{J_{1,v},J_{1,v}} & K_{J_{1,v},B_{1,v}} \\ K_{B_{1,v},J_{1,v}} & K_{B_{1,v},B_{1,v}} \end{pmatrix}^{-1} \Big]_{J_{1,v}} \tag{B.5.2}$$

Since $(\theta)j, j \in J_{1,v} \cup B_{1,v}) = \theta^{\mathcal{M}_{1,V}}$, we have that

$$\begin{pmatrix} K_{J_{1,v},J_{1,v}} & K_{J_{1,v},B_{1,v}} \\ K_{B_{1,v},J_{1,v}} & K_{B_{1,v},B_{1,v}} \end{pmatrix} = \frac{\partial^2 k^{\bar{\mathcal{M}}_{1,v}}}{\partial(\theta^{\bar{\mathcal{M}}_{1,v}})^2} = [var(\theta^{\mathcal{M}_{1,v}})]^{-1}$$

and therefore

$$[K^{-1}]_{J_{1,v}} = [var(\theta^{\mathcal{M}_{1,v}})]_{J_{1,v}} = var([\theta^{\mathcal{M}_{1,v}}]_{J_{1,v}}). \tag{B.5.3}$$

Moreover, using standard linear algebra formulas, we have that

$$[K^{-1}]_{J_{1,v}} = \left(K_{J_{1,v}\circ(B_{1,v}\cup J_{2\setminus 1,v}\cup B_{2,v})}\right)^{-1} \geq \left(K_{J_{1,v}\circ(J_{2\setminus 1,v}\cup B_{2,v})}\right)^{-1} = \left[(K_{J_{1,v}\cup J_{2\setminus 1,v}\cup B_{2,v}})^{-1}\right]_{J_{1,v}},$$

$$(K_{J_{1,v}\cup J_{2\setminus 1,v}\cup B_{2,v}})^{-1} = var(\hat{\theta}^{\mathcal{M}_{2,v}}), \tag{B.5.4}$$

$$(K_{J_{1,v}\cup J_{2\setminus 1,v}\cup B_{2,v}})^{-1} \geq (K_{J_{1,v}\cup J_{2\setminus 1,v}\cup B_{2,v}})^{-1}. \tag{B.5.5}$$

Combing (B.5.2), (B.5.3) and (B.5.4), we obtain that

$$var([\hat{\theta}^{\mathcal{M}_{1,v}}]_{J_{1,v}}) \geq var([\hat{\theta}^{\mathcal{M}_{2,v}}]_{J_{1,v}}),$$

which is the first inequality in (5.1.5). Now, combining (B.5.4) and (B.5.5), we obtain that

$$var([\hat{\theta}^{\mathcal{M}_{2,v}}]_{J_{1,v}}) \geq var(\hat{\theta}_{J_{1,v}})$$

and taking the diagonal elements of those matrices yields (5.1.5). □

B.6 Proof of Theorem 5.2.1

To prove Theorem 5.2.1, we need two preliminary results.

Lemma B.6.1. *Let $\theta^{v,*} = (\theta^*)^{v,PS}$ be the true value of the parameter for the conditional model of X_v given $X_{\mathcal{N}_v}$, and let $\hat{\theta}^{v,PS}$ be the value of $\theta^{v,PS}$ that maximizes $l^{v,PS}(\theta^{v,PS})$. Then, for $t_{J^{v,PS}}$ as in (5.2.2), if there exists $\epsilon > 0$ such that*

$$\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^{v,*})\|_\infty \leq \epsilon \leq \frac{C_{min}^2}{10 D_{max} d_v} \tag{B.6.1}$$

then

$$\|\hat{\theta}^{v,PS} - \theta^{v,*}\|_F \leq \frac{5\sqrt{d_v}\epsilon}{C_{min}} \tag{B.6.2}$$

Proof. To simplify our notation in this proof, we drop any subscripts and superscripts containing v or PS, except when it is necessary to keep them to make the argument clear.

Let $Q(\Delta) = l(\theta^*) - l(\theta^* + \Delta)$. Clearly $Q(0) = 0$ and $Q(\hat{\Delta}) \leq Q(0) = 0$, where $\hat{\Delta} = \hat{\theta} - \theta^*$. Let $||\Delta||_F = \sqrt{\sum_{j \in J^{v,PS}} \Delta_j^2}$ denote the Frobenius norm of Δ. Define $C(\delta) = \{\Delta \mid s.t.\ \|\Delta\|_F = \delta\}$. Since $Q(\Delta)$ is a convex function of Δ, if we can prove

$$\inf_{\Delta \in C(\delta)} Q(\Delta) > 0, \tag{B.6.3}$$

then, by convexity of Q, it will follow that $\hat{\Delta}$ must lie within the sphere defined by $C(\delta)$, i.e. $\|\hat{\Delta}\|_F \leq \delta$. We are now going to prove that there exists $\delta > 0$ such that on $C(\delta)$, $Q(\Delta) > 0$. For $\Delta \in C(\delta)$, we have

$$\begin{aligned} Q(\Delta) &= l(\theta^*) - l(\theta^* + \Delta) = \theta^{*t}t - k(\theta^*) - ((\theta^* + \Delta)^t t - k(\theta^* + \Delta)) \\ &= k(\theta^* + \Delta) - k(\theta^*) - \Delta^t t = \Delta^t k^{'}(\theta^*) + \tfrac{1}{2}\Delta^t k^{''}(\theta^* + \alpha\Delta)\Delta - \Delta^t t, \quad \alpha \in [0,1] \\ &= \underbrace{\Delta^t[k^{'}(\theta^*) - t]}_{Q_1} + \underbrace{\frac{1}{2}\Delta^t k^{''}(\theta^* + \alpha\Delta)\Delta}_{Q_2} \end{aligned} \tag{B.6.4}$$

By Hölder's and Cauchy's inequality, we have the following bound for Q_1.

$$|Q_1| = |\Delta^t[k^{'}(\theta^*) - t]| \ \leq \ \|k^{'}(\theta^*) - t\|_\infty ||\Delta||_1 \leq \epsilon\sqrt{d}\|\Delta\|_F = \epsilon\sqrt{d}\delta \tag{B.6.5}$$

For Q_2, we have

$$Q_2 \geq \frac{1}{2}\|\Delta\|_F^2 \min_{\alpha \in [0,1]} \lambda_{min} k^{''}(\theta^* + \alpha\Delta) = \frac{1}{2}\delta^2 \min_{\alpha \in [0,1]} \lambda_{min} k^{''}(\theta^* + \alpha\Delta) \tag{B.6.6}$$

We now want to bound the term $q = \min_{\alpha \in [0,1]} \lambda_{min}[k^{''}(\theta^* + \alpha\Delta)]$ from below. We change the input of function $z_{y_v}(\theta)$ in equation (5.2.4) to be $z_{y_v}(\theta + \alpha\Delta) = \sum_{j \in J; v \in S(j)} (\theta_j + \alpha\Delta_j) f_j(y_v, x_{\mathcal{N}_v}^{(n)})$,

so that we can rewrite the entries of H in (5.2.5) as

$$\eta_{k,l}^{n,v}(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)}) = \begin{cases} \frac{\exp z_{k_v}(\theta^*+\alpha\Delta)}{1+\sum_{y_v\in I_v\setminus\{0\}}\exp z_{y_v}(\theta^*+\alpha\Delta)} - (\frac{\exp z_{k_v}(\theta^*+\alpha\Delta)}{1+\sum_{y_v\in I_v\setminus\{0\}}\exp z_{k_v}(\theta^*+\alpha\Delta)})^2, & \text{if } k_v = l_v \\ -\frac{\exp z_{k_v}(\theta^*+\alpha\Delta)\exp z_{l_v}(\theta^*+\alpha\Delta)}{(1+\sum_{y_v\in I_v\setminus\{0\}}\exp z_{y_v}(\theta^*+\alpha\Delta))^2}, & \text{if } k_v \neq l_v \end{cases}$$

then

$$\frac{\partial\eta_{k,l}^{n,v}(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)})}{\partial\alpha} = \sum_{y_v\in I_v\setminus\{0\}} (\eta_{k,l}^{n,v})'_{y_v}(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)})\frac{\partial z_{y_v}}{\partial\alpha},$$

where $(\eta_{k,l}^{n,v})'_{y_v}(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)}) = \frac{\partial\eta_{k,l}^{n,v}(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)})}{\partial z_{y_v}}$. It is easy to see that these derivatives can all be expressed in terms of probabilities of the type (5.2.3) and that they are always less than 1 in absolute value. Therefore, since $\frac{\partial z_{y_v}(\theta+\alpha\Delta)}{\partial\alpha} = \sum_{j\in J; v\in S(j)}\Delta_j f_j(y_v, x_{\mathcal{N}_v}^n)$, we have

$$\begin{aligned} |\tfrac{\partial\eta_{k,l}^{n,v}(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)})}{\partial\alpha}| &\leq \textstyle\sum_{y_v\in I_v\setminus\{0\}}\frac{\partial z_{y_v}}{\partial\alpha} = \sum_{y_v\in I_v\setminus\{0\}}\sum_{j\in J; v\in S(j)}\Delta_j f_j(y_v, x_{\mathcal{N}_v}^n) \\ &= \textstyle\sum_{j\in J; v\in S(j)}\Delta_j\sum_{y_v\in I_v\setminus\{0\}} f_j(y_v, x_{\mathcal{N}_v}^n) = \langle\Delta, W^n\rangle\ , \end{aligned} \tag{B.6.7}$$

since for each $j\in J^{v,PS}$, $\sum_{y_v\in I_v\setminus\{0\}} f_j(y_v, x_{\mathcal{N}_v}^n) = f_j(j_v, x_{\mathcal{N}_v}^n) = W_j^n$.

The Taylor series expansion of $\eta_{k,l}^{n,v}(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)})$ yields

$$\eta_{k,l}^{n,v}(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)}) = \eta_{k,l}^{n,v}(\theta^*, x_{\mathcal{N}_v}^{(n)}) + \alpha\frac{\partial\eta_{k,l}^{n,v}(\theta^*+\alpha'\Delta, x_{\mathcal{N}_v}^{(n)})}{\partial\alpha},\ \alpha'\in[0,\alpha]\ .$$

Let $K(\theta^*+\alpha'\Delta, x_{\mathcal{N}_v}^{(n)})$ denote the $d_v\times d_v$ matrix with entry $\frac{\partial\eta_{k,l}^{n,v}(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)})}{\partial\alpha}$. Coming back to (B.6.6), we have

$$\begin{aligned} k''(\theta^*+\alpha\Delta) &= \tfrac{1}{N}\textstyle\sum_{n=1}^N\left[H(\theta^*+\alpha\Delta, x_{\mathcal{N}_v}^{(n)})\circ[W^n(W^n)^t]\right] \\ &= \tfrac{1}{N}\textstyle\sum_{n=1}^N H(\theta^*, x_{\mathcal{N}_v}^{(n)})\circ[W^n(W^n)^t] + \alpha\tfrac{1}{N}\sum_{n=1}^N K(\theta^*+\alpha'\Delta, x_{\mathcal{N}_v}^{(n)})\circ[W^n(W^n)^t]\ . \end{aligned}$$

We write $||X||_2 = \lambda_{max}(X)$ for the operator norm of a matrix X. By Lemma A.0.3,

$$\lambda_{min}\Big(k''(\theta^*+\alpha\Delta)\Big) \geq \lambda_{min}\Big(\frac{1}{N}\sum_{n=1}^N H(\theta^*, x_{\mathcal{N}_v}^{(n)})\circ[W^n(W^n)^t]\Big) - \|\alpha\frac{1}{N}\sum_{n=1}^N K(\theta^*+\alpha'\Delta, x_{\mathcal{N}_v}^{(n)})\circ[W^n(W^n)^t]\|_2$$

and since $|\alpha| < 1$, we have

$$\begin{aligned} q &= \min_{\alpha\in[0,1]} \lambda_{min}[\tfrac{1}{N}\textstyle\sum_{n=1}^{N} H(\theta^* + \alpha\Delta, x_{\mathcal{N}_v}^{(n)}) W^n (W^n)^t] \\ &\geq \lambda_{min}(\tfrac{1}{N}\textstyle\sum_{n=1}^{N} \left[H(\theta^*, x_{\mathcal{N}_v}^{(n)}) \circ (W^n (W^n)^t)\right]) \\ &\qquad - \max_{\alpha\in[0,1]} \|\alpha \tfrac{1}{N}[\textstyle\sum_{n=1}^{N} K(\theta^* + \alpha\Delta, x_{\mathcal{N}_v}^{(n)}) \circ (W^n (W^n)^t)]\|_2 \\ &\geq C_{min} - \max_{\alpha\in[0,1]} \|\underbrace{\frac{1}{N}\sum_{n=1}^{N} \Delta^t W^n (W^n (W^n)^t)}_{A}\|_2 \\ &= C_{min} - \max_{\alpha\in[0,1]} ||A||_2 \ , \end{aligned} \tag{B.6.8}$$

where the last but one inequality is due to our Assumption (B). We now need to bound the spectral norm of $A = \frac{1}{N}\sum_{n=1}^{N} \Delta^t W^n (W^n (W^n)^t)$. For any $\alpha \in [0,1]$ and $y \in R^{d_v}$ with $||y||_F = 1$, we have

$$\begin{aligned} \langle y, Ay\rangle &= \frac{1}{N}\sum_{n=1}^{N} (\Delta^t W^n)(y^t W^n)^2 \leq \frac{1}{N}\sum_{n=1}^{N} |\Delta^t W^n| (y^t W^n)^2, \\ |\Delta^t W^n| &\leqslant \sqrt{d}||\Delta||_F = \sqrt{d}\delta \ . \end{aligned} \tag{B.6.9}$$

and, by definition of the operator norm and from Assumption (B),

$$\frac{1}{N}\sum_{n=1}^{N} (y^t W^n)^2 \leq ||\frac{1}{N}\sum_{n=1}^{N} W^n (W^n)^t||_2 < D_{max} \ . \tag{B.6.10}$$

From (B.6.8), (B.6.9) and (B.6.10), we obtain $\max_{\alpha\in[0,1]} ||A||_2 \leq D_{max}\sqrt{d}\delta$ and therefore

$$q \geq C_{min} - D_{max}\sqrt{d}\delta \ .$$

Substituting this into (B.6.6), we get

$$Q_2 \geq \frac{1}{2}\delta^2 (C_{min} - D_{max}\sqrt{d}\delta). \tag{B.6.11}$$

From the two inequalities (B.6.5) and (B.6.11), it follows that

$$Q(\Delta) \geq Q_2 - |Q_1| \geq \frac{1}{2}\delta^2 (C_{min} - D_{max}\sqrt{d}\delta) - \epsilon\sqrt{d}\delta. \tag{B.6.12}$$

To simplify the problem, we can choose δ such that $C_{min} - D_{max}\sqrt{d}\delta \geq \frac{C_{min}}{2}$, that is, $\delta \leq \frac{C_{min}}{2D_{max}\sqrt{d}}$. Then inequality (B.6.12) becomes

$$Q(\Delta) \geq \frac{C_{min}\delta^2}{4} - \epsilon\sqrt{d}\delta$$

and $Q(\Delta)$ is positive if we let $\delta = \frac{5\sqrt{d}\epsilon}{C_{min}}$. Moreover $\delta \leq \frac{C_{min}}{2D_{max}\sqrt{d}}$ yields the following bound of ϵ:

$$\epsilon \leq \frac{C_{min}^2}{10D_{max}d}.$$

We have therefore shown that (B.6.3) holds for $\delta = \frac{5\sqrt{d}\epsilon}{C_{min}}$ and the lemma is proved. □

In the next lemma, we make use of the Hoeffding inequality (seeHoeffding (1963), Theorem 2) which states the following. If $X_1, X_2, \cdots, X_n$ are independent and $a_i \leq X_i \leq b_i (i = 1, 2, \cdots, n)$, then for $\epsilon > 0$

$$p(|\bar{X} - \mu| \geq \epsilon) \leq 2\exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{B.6.13}$$

Lemma B.6.2. *Let $t_{J^{v,PS}}, k^{v,PS}$ and d_v be as defined above. For any $\epsilon > 0$, we have*

$$p(\{\max_{v\in V}\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^{v,*})\|_\infty \geq \epsilon\}) \leq 2|J|\exp(-2N\epsilon^2). \tag{B.6.14}$$

Proof. For $j \in J^{v,PS}$, we clearly have

$$E_{\theta^*}\left(\frac{\partial l(\theta)}{\partial \theta_j}\right) = E_{\theta^*}\left(t_j - \frac{\partial k(\theta)}{\partial \theta_j}\right) = E_{\theta^*}\left(\frac{1}{N}\sum_{n=1}^{N} f_j(x_v^{(n)}, x_{\mathcal{N}_v}^{(n)}) - p(x_v = j_v|x_{\mathcal{N}_v}^n) f_j(x_v = j_v, x_{\mathcal{N}_v}^{(n)})\right) = 0$$

We note that since $x_{\mathcal{N}_v}^{(n)}$ is given and $f_j(x_v^{(n)}, x_{\mathcal{N}_v}^{(n)})$ takes values 0 or 1, we have $E(f_j(x_v^{(n)}, x_{\mathcal{N}_v}^{(n)})) = p(x_v = j_v|x_{\mathcal{N}_v}^n) f_j(x_v = j_v, x_{\mathcal{N}_v}^{(n)})$ and by Hoeffding's inequality (B.6.13), we have

$$p(|t_j - k_j'(\theta^*)| \geq \epsilon) \leq 2\exp -\frac{2N^2\epsilon^2}{2N} = 2\exp(-2N\epsilon^2)$$

Since $\{\max_{v\in V}\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\|_\infty \leq \epsilon\} = \cap_{j\in\cup J^{v,PS}}\{\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\| \leq \epsilon\}$, we have

that

$$\begin{aligned} P(\max_{v\in V}\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\|_\infty \le \epsilon) &= 1 - P(\cup_{j\in\cup J^{v,PS}}\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\| \ge \epsilon) \\ &\ge 1 - \sum_{j\in\cup J^{v,PS}} P(\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\| \ge \epsilon), \\ &\ge 1 - 2|J|\exp(-2N\epsilon^2) \end{aligned}$$

which proves the lemma.

Proof of Theorem 5.2.1

Let $\epsilon = C\sqrt{\frac{\log p}{N}}$, where C is a constant that we will choose later in this proof. From Lemma B.6.2, we have

$$p(\max_{v\in V}\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\|_\infty \ge C\sqrt{\frac{\log p}{N}}) \le 2|J|\exp(-2C^2\log p) = \frac{2|J|}{p^{2C^2}} \tag{B.6.15}$$

From Lemma B.6.1, for $\epsilon = C\sqrt{\frac{\log p}{N}} \le \frac{C_{min}^2}{10D_{max}d_v}$, i.e. for $N \ge (\frac{10CD_{max}d_v}{C_{min}^2})^2\log p$, we have

$$\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\|_\infty \le \epsilon \le \frac{C_{min}^2}{10D_{max}d_v} \Rightarrow \|\hat{\theta}^{v,PS} - \theta^{v,*}\|_F \le \frac{5\sqrt{d_v}\epsilon}{C_{min}}.$$

The MCLE $\bar{\theta}$ obtained by the local averaging of the $\hat{\theta}^{v,PS}$ from each conditional model can then be bounded as follows:

$$\begin{aligned} \|\bar{\theta} - \theta^*\|_F &\le \big(\textstyle\sum_{v\in V}\|\hat{\theta}^{v,PS} - \theta^{v,*}\|_F^2\big)^{\frac{1}{2}} \\ &\le (\textstyle\sum_{v\in V}(\frac{5\sqrt{d_v}C\sqrt{\frac{\log p}{N}}}{C_{min}})^2)^{\frac{1}{2}} = \frac{5C}{C_{min}}\sqrt{\frac{\sum_{v\in V} d_v\log p}{N}} \end{aligned}$$

Therefore under the condition $N \ge \max_{v\in V}\ (\frac{10CD_{max}d_v}{C_{min}^2})^2\log p$, we have

$$p(\|\bar{\theta} - \theta^*\|_F \le \frac{5C}{C_{min}}\sqrt{\frac{\sum_{v\in V} d_v\log p}{N}}) \ge p(\max_{v\in V}\|t_{J^{v,PS}} - k'^{v,PS}(\theta^*)\|_\infty \le C\sqrt{\frac{\log p}{N}}) \ge 1 - \frac{2|J|}{p^{2C^2}}$$

with the last inequality due to (B.6.15).

The theorem would make no sense if the probability of the convergence rate was negative, and thus C must satisfy

$$1 - \frac{2|J|}{p^{2C^2}} > 0 \Rightarrow C \ge \sqrt{\frac{\log(2|J|)}{2\log p}}\ .$$

□

B.7 Proof of Theorem 5.2.2

We first need to prove a series of lemmas. We recall our two assumptions:

$$(A') \ \text{ there exists } \ D_{max} > 0 \ \text{ such that } \ \lambda_{max}\Big(\sum_{i\in I} f_i \otimes f_i\Big) \leq D_{max},$$

$$(B') \ \ 0 < \kappa^* = \lambda_{min}\Big[k^{''}(\theta^*)\Big].$$

Assumption $A^{'}$ yields an upper bound for the maximum eigenvalue of the Fisher information matrix as stated in the following lemma.

Lemma B.7.1. *If assumption $A^{'}$ is satisfied, then*

$$\lambda_{max}(k^{''}(\theta^*)) \leq D_{max}$$

Proof. First, the diagonal elements of $k^{''}(\theta^*)$ are $\{P_j^* - P_j^{*2} | j \in J\}$, therefore, since $P_j^* - P_j^{*2} \leq \frac{1}{4}$, we have

$$\lambda_{max}(k^{''}(\theta^*)) \leq \sum_{j\in J} P_j^* - P_j^{*2} \leq \frac{|J|}{4}.$$

Since for a symmetric matrix A, $\lambda_{max}(A) = \max_{||y||=1} y^t A y$, we have

$$\lambda_{max}\Big(\sum_{i\in I} f_i \otimes f_i\Big) \geq \frac{1}{|J|}\sum_{i=1}^{|J|}\sum_{j=1}^{|J|} a_{ij},$$

where a_{ij} are the entries of $\sum_{i\in I} f_i \otimes f_i$. The sum of the elements in matrix $f_i \otimes f_i$ is $|\{j | j \lhd i\}|^2$ and therefore

$$\sum_{i=1}^{|J|}\sum_{j=1}^{|J|} a_{ij} = \sum_{i\in I} |\{j | j \lhd i\}|^2 \geq |J|^2.$$

Thus

$$\lambda_{max}\Big(\sum_{i\in I} f_i \otimes f_i\Big) \geq |J| \geq \frac{|J|}{4} \geq \lambda_{max}(k^{''}(\theta^*)),$$

and

$$\max_{v\in V} \lambda_{max}(k^{''}(\theta^*)) \leq D_{max}$$

□

The next lemma gives an upper bound for the square error of the MLE $\hat{\theta}^G$ in the global model.

Lemma B.7.2. *Let $t = \{t_j | j \in J\}$ be the vector of marginal cell counts, and let $P(\theta^*) \in R^{|J|}$ be the vector of marginal cell probabilities in the global model at the true value of the parameter θ^*. If*

$$\|\frac{t}{N} - k^{'}(\theta^*)\|_\infty \leq \epsilon \leq \frac{\kappa^{*2}}{40|J|D_{max}}, \tag{B.7.1}$$

then

$$\|\hat{\theta}^G - \theta^*\|_F \leq \frac{5\sqrt{|J|}\epsilon}{\kappa^*} \ . \tag{B.7.2}$$

Proof. From the log-likelihood function of our discrete graphical model, we have $\frac{t}{N} = k^{'}(\hat{\theta})$. Consider the function $Q(\Delta) = l(\theta^*) - l(\theta^* + \Delta)$, $\Delta \in R^{|J|}$. Clearly, $Q(0) = 0$ and $Q(\hat{\Delta}) \leq Q(0) = 0$, where $\hat{\Delta} = \hat{\theta}^G - \theta^*$.

Define $C(\delta) = \{\Delta | \quad \|\Delta\|_2 = \delta\}$. Since $Q(\Delta)$ is a convex function of Δ, if we can prove

$$\inf_{\Delta \in C(\delta)} Q(\Delta) > 0,$$

it will follow that $\hat{\Delta}$ must lie in the sphere defined by $C(\delta)$. Therefore

$$\|\hat{\Delta}\|_2 \leq \delta.$$

We now try to find a suitable radius δ for which $Q(\delta) > 0$.

For $\Delta \in C(\delta)$:

$$\begin{aligned} Q(\Delta) &= l(\theta^*) - l(\theta^* + \Delta) \\ &= < \tfrac{t}{N}, \theta^* > -k(\theta^*) - (< \tfrac{t}{N}, \theta^* + \Delta > -k(\theta^* + \Delta)) \\ &= k(\theta^* + \Delta) - k(\theta^*) - < \tfrac{t}{N}, \Delta > \end{aligned} \tag{B.7.3}$$

Inserting the Taylor expansion of $k(\theta^* + \Delta)$ around θ^*,

$$k(\theta^* + \Delta) - k(\theta^*) =< k^{'}(\theta^*), \Delta > + \Delta^T [\int_0^1 (1-\alpha) k^{''}(\theta^* + \alpha\Delta) d\alpha] \Delta,$$

into (B.7.3), we obtain

$$Q(\Delta) = \underbrace{< k^{'}(\theta^*) - \frac{t}{N}, \Delta >}_{Q_1} + \underbrace{\Delta^T [\int_0^1 (1-\alpha) k^{''}(\theta^* + \alpha\Delta) d\alpha] \Delta}_{Q_2} \tag{B.7.4}$$

For Q_1, we have

$$\begin{aligned} |Q_1| = | < k^{'}(\theta^*) - \tfrac{t}{N}, \Delta > | &\leq \|k^{'}(\theta^*) - \tfrac{t}{N}\|_\infty ||\Delta||_1 \\ &\leq \epsilon\sqrt{|J|}\|\Delta\|_2 = \epsilon\sqrt{|J|}\delta \end{aligned} \tag{B.7.5}$$

For Q_2, we have

$$\begin{aligned} Q_2 &\geq \|\Delta\|_2^2 \lambda_{min}(\textstyle\int_0^1 (1-\alpha) k^{''}(\theta^* + \alpha\Delta) d\alpha) \\ &\geq \|\Delta\|_2^2 \textstyle\int_0^1 (1-\alpha) \lambda_{min}(k^{''}(\theta^* + \alpha\Delta)) d\alpha \\ &\geq \tfrac{1}{2}\|\Delta\|_2^2 \min_{\alpha\in[0,1]} \lambda_{min}(k^{''}(\theta^* + \alpha\Delta)) \end{aligned} \tag{B.7.6}$$

We now need to bound the term $\min_{\alpha\in[0,1]} \lambda_{min}[k^{''}(\theta^* + \alpha\Delta)]$. The Fisher information matrix is

$$\begin{aligned} k^{''}(\theta) &= \tfrac{\sum_{i\in I} \exp <\theta, f_i>}{L(\theta)} (f_i \otimes f_i) - (\tfrac{\sum_{i\in I} \exp <\theta, f_i>}{L(\theta)} f_i) \otimes (\tfrac{\sum_{i\in I} \exp <\theta, f_i>}{L(\theta)} f_i) \\ &= \tfrac{\sum_{i\in I} \exp <\theta, f_i>}{L(\theta)} (f_i \otimes f_i) - P(\theta) \otimes P(\theta) \end{aligned}$$

where $P(\theta) = k'(\theta)$ is the vector of marginal probabilities. Therefore

$$k^{''}(\theta^* + \alpha\Delta) = \underbrace{\frac{\sum_{i\in I} \exp < \theta^* + \alpha\Delta, f_i >}{L(\theta^* + \alpha\Delta)} (f_i \otimes f_i)}_{T_1} - \underbrace{P(\theta^* + \alpha\Delta) \otimes P(\theta^* + \alpha\Delta)}_{T_2} \tag{B.7.7}$$

A Taylor expansion of $\frac{e^{\langle\theta^* + \alpha\Delta, f_i\rangle}}{L(\theta^* + \alpha\Delta)}$ around $\alpha = 0$ is

$$\frac{e^{\langle\theta^* + \alpha\Delta, f_i\rangle}}{L(\theta^* + \alpha\Delta)} = \frac{e^{\langle\theta^*, f_i\rangle}}{L(\theta^*)} + \underbrace{\Big[\frac{e^{\langle\theta^* + \alpha^*\Delta, f_i\rangle}}{L(\theta^* + \alpha^*\Delta)} \langle f_i, \Delta\rangle - \frac{e^{\langle\theta^* + \alpha^*\Delta, f_i\rangle}}{L(\theta^* + \alpha^*\Delta)} \sum_{i\in I} \frac{e^{\langle\theta^* + \alpha^*\Delta, f_i\rangle}}{L(\theta^* + \alpha^*\Delta)} \langle f_i, \Delta\rangle\Big]}_{A_i} \tag{B.7.8}$$

for some $\alpha^* \in [0, \alpha]$. For each $i \in I$,

$$\begin{aligned} A_i &= \Big[p_i(\theta^* + \alpha^*\Delta)\langle f_i, \Delta\rangle - p_i(\theta^* + \alpha^*\Delta)\textstyle\sum_{i' \in I} p_{i'}(\theta^* + \alpha^*\Delta)\langle f_{i'}, \Delta\rangle\Big] \\ &= \Big[p_i(\theta^* + \alpha^*\Delta)\textstyle\sum_{j \lhd i} \Delta_j - p_i(\theta^* + \alpha^*\Delta)\sum_{j \in J} P(j_{S(j)})\Delta_j\Big] \\ &= \Big[\textstyle\sum_{j \lhd i} p_i(\theta^* + \alpha^*\Delta)(1 - P(j_{S(j)})\Delta_j - \sum_{j \not\lhd i} p_i(\theta^* + \alpha^*\Delta)P(j_{S(j)})\Delta_j\Big] \\ &= \langle \pi_i, \Delta\rangle, \end{aligned}$$

where

$$\pi_i = p_i(\theta^* + \alpha^*\Delta)\Big((1 - P(j_{S(j)}), \quad j \lhd i, \ -P(j_{S(j)}), \quad j \not\lhd i\Big).$$

Therefore, by Cauchy-Schwarz inequality, we have that

$$|A_i| \leq ||\pi_i||_2 \times ||\Delta||_2 \leq \sqrt{|J|}||\Delta||_2 = \sqrt{|J|}\delta.$$

Then T_1 can be written as

$$T_1 = \sum_{i \in I} \frac{e^{\langle \theta^*, f_i\rangle}}{L(\theta^*)}(f_i \otimes f_i) + \sum_{i \in I} A_i(f_i \otimes f_i)$$

For term T_2, there exists a $|J|$-dimensional vector u, such that

$$P(\theta^* + \alpha\Delta) = P(\theta^*) + u,$$

which means

$$u = P(\theta^* + \alpha\Delta) - P(\theta^*) = P(\theta^*)'\Delta + o(\Delta) = k''(\theta^*)\Delta + o(\Delta^2)$$

and thus $||u||_F \leq \lambda_{max}[k''(\theta^*)]||\Delta||_F + o(||\Delta||_F^2)$. Therefore using Lemma B.7.1 and the fact that the magnitude of $o(||\Delta||_F^2)$ is much smaller than the difference between $\lambda_{max}(k''(\theta^*))$ and $\lambda_{max}(\sum_{i \in I} f_i \otimes f_i)$, we have $||u||_F \leq D_{max}\delta$.

Now, T_2 can be written as

$$T_2 = P(\theta^*) \otimes P(\theta^*) + u \otimes P(\theta^*) + P(\theta^*) \otimes u + u \otimes u.$$

If we plug T_1, T_2 back into (B.7.7), we have

$$\begin{aligned} k''(\theta^* + \alpha\Delta) &= T_1 - T_2 \\ &= \textstyle\sum_{i\in I} \frac{e^{\langle \theta^*, f_i\rangle}}{L(\theta^*)}(f_i \otimes f_i) + \sum_{i\in I} A_i(f_i \otimes f_i) - P(\theta^*) \otimes P(\theta^*) \\ &\qquad -u \otimes P(\theta^*) - P(\theta^*) \otimes u - u \otimes u \\ &= k''(\theta^*) + \textstyle\sum_{i\in I} A_i(f_i \otimes f_i) - u \otimes P(\theta^*) - P(\theta^*) \otimes u - u \otimes u \end{aligned} \tag{B.7.9}$$

From the first inequality of Lemma A.0.3, we know that

$$\begin{aligned} \lambda_{min} k''(\theta^* + \alpha\Delta) &\geq \lambda_{min}\Big[k''(\theta^*) + \sum_{i\in I} A_i(f_i \otimes f_i)\Big] \\ &\quad +\lambda_{min}\Big[- P(\theta^*) \otimes P(\theta^*)\Big] + \lambda_{min}\Big[- u \otimes P(\theta^*)\Big] \\ &\quad +\lambda_{min}\Big[- P(\theta^*) \otimes u\Big] + \lambda_{min}\Big[- u \otimes u\Big] \\ &= \lambda_{min}\Big[k''(\theta^*) + \sum_{i\in I} A_i(f_i \otimes f_i)\Big] \\ &\quad -\lambda_{max}\Big[P(\theta^*) \otimes P(\theta^*)\Big] - \lambda_{max}\Big[u \otimes P(\theta^*)\Big] \\ &\quad -\lambda_{max}\Big[P(\theta^*) \otimes u\Big] - \lambda_{max}\Big[u \otimes u\Big] \end{aligned}$$

where we also use the fact that $P(\theta^*) \otimes u,\ u \otimes P(\theta^*),\ u \otimes u$ are rank one matrices with only one nonzero eigenvalue (positive or negative). From the second inequality of Lemma A.0.3, we know that

$$\lambda_{min}\Big[k''(\theta^*) + \sum_{i\in I} A_i(f_i \otimes f_i)\Big] \geq \lambda_{min}\Big[k''(\theta^*)\Big] - \|\textstyle\sum_{i\in I} A_i(f_i \otimes f_i)\|_2$$

Therefore

$$\begin{aligned} \min_{\alpha\in[0,1]} \lambda_{min}(k''(\theta^* + \alpha\Delta)) &\geq \lambda_{min}(k''(\theta^*)) - \max_{\alpha\in[0,1]} \|\sum_{i\in I} A_i(f_i \otimes f_i)\|_2 \\ &\qquad -\lambda_{max}(u \otimes P(\theta^*)) - \lambda_{max}(P(\theta^*) \otimes u) - \lambda_{max}(u \otimes u) \end{aligned}$$

Bound the terms in the above formula, one by one:

$$\lambda_{min}(k^{''}(\theta^*)) \geq \kappa^*.$$

We now want to control the spectral norm of the matrix $\sum_{i\in I} A_i(f_i \otimes f_i)$ for $\alpha \in [0,1]$. For any fixed $\alpha \in [0,1]$, and vector $y \in R^{J^{\mathcal{M}_{i,v}}}$ with $\|y\|_2 = 1$, we have

$$\begin{aligned} \|\textstyle\sum_{i\in I} A_i(f_i \otimes f_i)\|_2 &= \max_{\|y\|_2=1} y^{'}\Big(\textstyle\sum_{i\in I} A_i(f_i \otimes f_i)\Big)y = \max_{\|y\|_2=1} \textstyle\sum_{i\in I} A_i\Big(y^{'}(f_i \otimes f_i)y\Big) \\ &\leq \max_{\|y\|_2=1} \textstyle\sum_{i\in I} |A_i| y^{'}(f_i \otimes f_i)y. \end{aligned}$$

Recall that for any $\alpha \in [0,1]$, $|A_i| \leq \sqrt{|J|}\delta$. Moreover, we have $\lambda_{max}(\sum_{i\in I}(f_i \otimes f_i)) \leq D_{max}$ by assumption. Combining these two pieces, we get

$$\max_{\alpha\in[0,1]} \|\sum_{i\in I} A_i(f_i \otimes f_i)\|_2 \leq \sqrt{|J|}\delta \max_{\|y\|_2=1} y^{'} \sum_{i\in I}(f_i \otimes f_i)y = \sqrt{|J|}\delta\lambda_{max}(\sum_{i\in I}(f_i \otimes f_i)) \leq \sqrt{|J|}\delta D_{max}.$$

The eigenvalue of $P \otimes u$ and $u \otimes P$ as in Lemma A.0.2 is

$$\lambda_{max}(u \otimes P(\theta^*)) = \lambda_{max}(P(\theta^*) \otimes u) \leq |\langle P(\theta^*), u\rangle| \leq \|P(\theta^*)\|_F \|u\|_F,$$

and since we have $\|P(\theta^*)\|_F = \sqrt{\sum_{j=1}^{J} P_j(\theta^*)^2} \leq \sqrt{J}$ and $\|u\|_F \leq D_{max}\delta$, then

$$\lambda_{max}(u \otimes P(\theta^*)) \leq \sqrt{|J|} D_{max}\delta \ .$$

The eigenvalue of $u \otimes u$ as in Lemma A.0.1 is

$$\lambda_{max}(u \otimes u) = \langle u, u\rangle = \|u\|_F \|u\|_F \leq \sqrt{|J|} D_{max}\delta,$$

since $\|u\|_F = \|P(\theta^* + \alpha\Delta) - P(\theta^*)\|_F \leq \sqrt{J}$ (the difference of two positive quantities less than 1).

Combining all these pieces, we get the bound of the minimum eigenvalue of the Fisher information matrix:

$$\min_{\alpha\in[0,1]} \lambda_{min}(k^{''}(\theta^* + \alpha\Delta)) \geq \kappa^* - \sqrt{|J|}\delta D_{max} - 3\sqrt{|J|} D_{max}\delta.$$

Substitute this into (B.7.6) and we get

$$Q_2 \geq \frac{1}{2}\delta^2(\kappa^* - \sqrt{|J|}\delta D_{max} - 3\sqrt{|J|}D_{max}\delta) = \frac{1}{2}\delta^2(\kappa^* - 4\sqrt{|J|}\delta D_{max}). \tag{B.7.10}$$

From the two inequalities (B.7.5) and (B.7.10), we have

$$Q(\Delta) \geq Q_2 - |Q_1| \geq \frac{1}{2}\delta^2(\kappa^* - 4\sqrt{|J|}\delta D_{max}) - \epsilon\sqrt{|J|}\delta. \tag{B.7.11}$$

To simplify the problem, we can choose δ such that $\kappa^* - 4\sqrt{|J|}\delta D_{max} \geq \frac{\kappa^*}{2}$, so that $\delta \leq \frac{\kappa^*}{8\sqrt{|J|}D_{max}}$.

Then inequality (B.7.11) becomes

$$Q(\Delta) \geq \frac{\kappa^*\delta^2}{4} - \epsilon\sqrt{|J|}\delta \ .$$

$Q(\Delta)$ can be positive if we let $\delta = \frac{5\sqrt{|J|}\epsilon}{\kappa^*}$, which also gives us the following bound for ϵ:

$$\epsilon \leq \frac{\kappa^{*2}}{40|J|D_{max}} \ .$$

So, we have found a $\delta > 0$ such that $Q(\Delta) \geq 0$ and therefore

$$\|\hat{\theta}^G - \theta^*\|_F \leq \delta = \frac{5\sqrt{|J|}\epsilon}{\kappa^*},$$

when

$$\|\frac{t}{N} - P^*\|_\infty \leq \epsilon \leq \frac{\kappa^{*2}}{40|J|D_{max}}$$

where $P^* = k'(\theta^*)$. □

We can now proceed to proving Theorem 5.2.2

Proof of Theorem 5.2.2

Proof. Let $\epsilon = C\sqrt{\frac{\log p}{N}}$, where C is a constant to be chosen later in the proof. From Hoeffding's inequality (see main file), we have

$$P\{|\frac{t_j}{N} - P_j^*| \geq C\sqrt{\frac{\log p}{N}}\} \leq 2\exp(-2N\epsilon^2) = \frac{2}{p^{2C^2}}.$$

Applying the union bound yields

$$P(||(\frac{t}{N} - P^*)_J||_\infty \geq C\sqrt{\frac{\log p}{N}}) \leq \sum_{j\in J} P\{|\frac{t_j}{N} - P_j^*| \geq C\sqrt{\frac{\log p}{N}}\} \leq \frac{2|J|}{p^{2C^2}}.$$

From Lemma B.7.2:

$$\|(\frac{t}{N} - P^*)_J\|_\infty \leq C\sqrt{\frac{\log p}{N}} \Rightarrow \|\hat{\theta} - \theta^*\|_F \leq \frac{5C}{\kappa^*}\sqrt{\frac{|J|\log p}{N}},$$

when

$$\begin{aligned} \epsilon &= C\sqrt{\tfrac{\log p}{N}} \leq \tfrac{\kappa^{*2}}{40|J|D_{max}} \\ N &\geq (\tfrac{40C|J|D_{max}}{\kappa^{*2}})^2 \log p. \end{aligned}$$

Therefore when $N \geq (\frac{40C|J|D_{max}}{\kappa^{*2}})^2 \log p$,

$$p(\|\hat{\theta} - \theta^*\|_F \leq \frac{5C}{\kappa^*}\sqrt{\frac{|J|\log p}{N}}) \geq p(\|(\frac{t}{N} - P^*)_J\|_\infty \leq C\sqrt{\frac{\log p}{N}}) \geq 1 - \frac{2|J|}{p^{2C^2}}.$$

The theorem would not make sense if the probability of the convergence rate is negative. It follows we need to have

$$1 - \frac{2|J|}{p^{2C^2}} > 0 \Rightarrow C \geq 2\sqrt{\frac{\log 2|J|}{\log p}}.$$

□

B.8 Proof of Theorem 6.0.1

Theorem 6.0.1 goes back to Barndorff-Nielsen (2014), who studies the closure of much more general exponential families. The case of a discrete exponential family is much easier.

For a probability measure p on I given, let $supp(p)$ be the support of p. The theorem follows from the following lemmas:

Lemma B.8.1. *Let $p \in \overline{\mathcal{E}_A}$. Then $p \in \mathcal{E}_{A,supp(p)}$.*

Lemma B.8.2. *Let $p \in \overline{\mathcal{E}_A}$. Then $\mathcal{E}_{A,supp(p)} \subseteq \overline{\mathcal{E}_A}$.*

Lemma B.8.3. *Let $p \in \overline{\mathcal{E}_A}$. Then $supp(p)$ is facial.*

Lemma B.8.4. *If F is facial, then there exists $p \in \overline{\mathcal{E}_A}$ with $supp(p) = F$.*

Indeed, Lemma B.8.1 shows that $\overline{\mathcal{E}_A} \subseteq \bigcup_F \mathcal{E}_{A,F}$, where the union is over all support sets F. Lemma B.8.2 shows the converse containment is also true, so that $\overline{\mathcal{E}_A} = \bigcup_F \mathcal{E}_{A,F}$. It remains to see that a subset $F \subseteq I$ is a support set if and only if F is facial. This follows from Lemma B.8.3 and B.8.4.

In the proofs of Lemma B.8.1 to B.8.4, we need the following easy lemma for which we don't provide the proof:

Lemma B.8.5. *$p \in \mathcal{E}_A$ if and only if* $\log(p) \perp \ker A$.

Proof of Lemma B.8.1. Let $p = \lim_{k\to\infty} p_k$, where $p_k \in \mathcal{E}_A$, and let $F = supp(p)$. Then $\mathcal{E}_{A,F}$ is the exponential family $\mathcal{E}_{A_F}$, where A_F consists of the columns of A indexed by F. Any $v \in \ker A_F$ can be extended by zeros to $v' \in \ker A$. By Lemma B.8.5,

$$0 = \langle \log(p_k), v' \rangle = \sum_{i \in F} \log(p_k(i)) v(i) \to \langle \log(p), v \rangle.$$

Thus, $\log(p) \perp \ker A_F$, which implies $p \in \mathcal{E}_{A,F}$. □

Proof of Lemma B.8.2. Let $p = \lim_{k\to\infty} p_k$, where $p_k \in \mathcal{E}_A$, let $F = supp(p)$, and let $q \in \mathcal{E}_{A,F}$. Then there exists parameter θ with $\log(q(i)) - \log(p(i)) = \langle \theta, f_i \rangle$ for all $i \in F$. For any k, there exists a positive constant c_k such that $q_k := c_k p_k \exp(\langle \theta, A \rangle) \in \mathcal{E}_A$. Then $q_k \to q$ as $k \to \infty$, and so $q \in \overline{\mathcal{E}_A}$. □

Proof of Lemma B.8.3. Let $p = \lim_{k\to\infty} p_k$, where $p_k \in \mathcal{E}_A$, and let $F = F_A(supp(p))$. Then $x = \frac{1}{|supp(p)|} \sum_{i \in supp(p)} f_i$ is an interior point of the face corresponding to F, and thus there exist

positive coefficients $\lambda_i > 0$, $i \in F$, with $x = \sum_{i \in F} \lambda_i f_i$. The vector $v = (v_i, i \in I)$ defined by

$$v_i = \begin{cases} \frac{1}{|supp(p)|} - \lambda_i, & i \in supp(p), \\ -\lambda_i, & i \in F \setminus supp(p), \\ 0, & i \notin F, \end{cases}$$

satisfies $Av = x - x = 0$. By Lemma B.8.5, $\log(p_k) \perp v$ for all k. In particular,

$$\sum_{i \in F \setminus supp(p)} \lambda_i \log(p_k(i)) = \sum_{i \in supp(p)} \log(p_k(i)) v_i \to \sum_{i \in supp(p)} \log(p(i)) v_i.$$

On the other hand, note that each coefficient λ_i for $i \in F \setminus supp(p)$ on the left hand side is positive, while $\log(p_k(i)) \to -\infty$ for $i \notin supp(p)$. This shows that $F \setminus supp(p) = \emptyset$. □

Proof of Lemma B.8.4. If F is facial, there exist $g \in \mathbf{R}^h$ and $c \in \mathbf{R}$ with $\langle g, f_i \rangle \geq c$ for all $i \in I$ and $\langle g, f_i \rangle = c$ if and only if $i \in F$. Let $\theta^{(s)} = -s \cdot g$. Then

$$k_F(\theta_{(s)}) + sc = \log \sum_{i \in I} \exp(-s\langle g, f_i \rangle + sc) \to \log |F|,$$

and so

$$\log p_{\theta^{(s)}}(i) = -s\langle g, f_i \rangle - k_F(\theta_{(s)}) = (sc - s\langle g, f_i \rangle) - (k_F(\theta_{(s)}) + sc)$$

$$\to \begin{cases} -\log |F|, & \text{if } i \in F, \\ -\infty, & \text{if } i \notin F, \end{cases}$$

as $s \to \infty$. Thus, $p_{\theta^{(s)}}$ converges to the uniform distribution on F. □

B.9 Proof of Theorem 6.0.2

By definition, any EMLE p_* belongs to the closure of the model. According to Theorem 6.0.1, the support of p_* is facial. If $supp(p)$ does not contain $supp(n)$, then the log-likelihood goes to

minus infinity, $\tilde{l}(p) = -\infty$, and so p does not maximize the likelihood, therefore, $supp(p_*)$ is a facial set containing $supp(n)$. Thus, $F_t \subseteq supp(p_*)$.

By Lemma B.8.1, p_* belongs to $\mathcal{E}_{\Delta,supp(p_*)}$, which is parametrized by a vector θ, see Theorem 6.0.1. On $\mathcal{E}_{\Delta,supp(p_*)}$, the log-likelihood function in terms of this parameter θ is

$$l_F(\theta) = \sum_{j \in J} \theta_j t_j - N k_F(\theta).$$

l_F is strictly concave, and so it has a unique maximum. The critical equations are

$$Ap_* = \frac{t}{N},$$

proving the first property. Note that these equations are independent of the parameters and the support of p_*. We now show that any solution to these equations is supported on the same face of $\mathbf{P}$ as $\frac{t}{N}$.

Let p be a probability distribution on I such that $supp(p)$ does not contain F_t. This means that there is a linear inequality $\langle g, t \rangle \geq c$ that is valid on $\mathbf{P}$ and such that

- $\langle g, f_i \rangle = c$ for all $i \in F_t$;
- $\langle g, f_i \rangle > c$ for some $i \in supp(p)$.

Then

$$\langle g, Ap \rangle = \sum_i \langle g, f_i \rangle p(i) > c = \frac{1}{N} \sum_i n(i) \langle g, f_i \rangle = \langle g, \frac{t}{N} \rangle,$$

which implies $Ap \neq \frac{t}{N}$. This shows $supp(p_*) \subseteq F_t$ and finishes the proof of $supp(p_*) = F_t$.

We have now shown the two properties, and it remains to argue that the EMLE is unique. But this follows from the fact that $supp(p_*)$ is equal to F_t, and l_F is strictly convex, such that the likelihood has a unique maximizer on $\mathcal{E}_{\Delta,F_t}$.

C Example: Two binary random variables

Consider two binary random variables, and let $\Delta = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$. The hierarchical model $\mathcal{E}_\Delta$ is the *saturated model*; that is, it contains all possible probability distributions with full support. Then

$$\tilde{A} = \begin{array}{c} \overbrace{}^{f_{00}} \overbrace{}^{f_{01}} \overbrace{}^{f_{10}} \overbrace{}^{f_{11}} \\ \left(\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{array}\right) \end{array} \begin{array}{l} \theta_{00} \\ \theta_{01} \\ \theta_{10} \\ \theta_{11} \end{array}$$

The marginal polytope is a 3-simplex (a tetrahedron) with facets

$$\mathbf{F}_{00} : 1 - t_{01} - t_{10} + t_{11} \geq 0, \quad \mathbf{F}_{01} : t_{01} - t_{11} \geq 0,$$

$$\mathbf{F}_{10} : t_{10} - t_{11} \geq 0, \quad \mathbf{F}_{11} : t_{11} \geq 0.$$

Each of the corresponding facets contains three columns of $\tilde{A}$. Facet $\mathbf{F}_i$ in the above list does not contain the column f_i of $\tilde{A}$.

The EMLE of the saturated model is just the empirical distribution; that is, $p_* = \frac{1}{N}n$. Suppose that t lies on the facet $\mathbf{F}_{00}$ (i.e. $n = (0, n_{01}, n_{10}, n_{11})$ with $n(01), n(10), n(11) > 0$). If $p_{\theta^{(s)}} \to p_*$,

then $p_{\theta^{(s)}}(00) \to 0$, while all other probabilities converge to a non-zero value. It follows that

$$\theta_{00}^{(s)} = \log p_{\theta^{(s)}}(00) \to -\infty,$$
$$\theta_{01}^{(s)} = \log \frac{p_{\theta^{(s)}}(01)}{p_{\theta^{(s)}}(00)} \to +\infty,$$
$$\theta_{10}^{(s)} = \log \frac{p_{\theta^{(s)}}(10)}{p_{\theta^{(s)}}(00)} \to +\infty,$$
$$\theta_{11}^{(s)} = \log \frac{p_{\theta^{(s)}}(11)p_{\theta^{(s)}}(00)}{p_{\theta^{(s)}}(01)p_{\theta^{(s)}}(10)} \to -\infty.$$

On the other hand, $\theta_{01}^{(s)} + \theta_{00}^{(s)} = \log p_{\theta^{(s)}}(01)$ converges to a finite value, as do $\theta_{10}^{(s)} + \theta_{00}^{(s)} = \log p_{\theta^{(s)}}(10)$ and $\theta_{11}^{(s)} + \theta_{01}^{(s)} = \log p_{\theta^{(s)}}(11)/p_{\theta^{(s)}}(10)$.

Proceeding similarly for the other facets, one can show for the limits $\theta_{ij} := \lim_{s\to\infty} \theta_{ij}^{(s)}$:

	θ_{00}	θ_{01}	θ_{10}	θ_{11}	finite parameter combinations:
$\mathbf{F}_{00}$	$-\infty$	$+\infty$	$+\infty$	$-\infty$	$\theta_{01}^{(s)} + \theta_{00}^{(s)}$, $\theta_{10}^{(s)} + \theta_{00}^{(s)}$, $\theta_{11}^{(s)} + \theta_{01}^{(s)}$
$\mathbf{F}_{01}$	finite	$-\infty$	finite	$+\infty$	$\theta_{00}^{(s)}$, $\theta_{10}^{(s)}$, $\theta_{01}^{(s)} + \theta_{11}^{(s)}$
$\mathbf{F}_{10}$	finite	finite	$-\infty$	$+\infty$	$\theta_{00}^{(s)}$, $\theta_{01}^{(s)}$, $\theta_{10}^{(s)} + \theta_{11}^{(s)}$
$\mathbf{F}_{11}$	finite	finite	finite	$-\infty$	$\theta_{00}^{(s)}$, $\theta_{10}^{(s)}$, $\theta_{01}^{(s)}$

Each line of the last column contains three combinations of the parameters $\theta_i^{(s)}$ that converge to a finite value. Any other parameter combination that converges is a linear combination of these three. This can be seen by using coordinates μ_i introduced in Section 8.2 and applying Lemma 8.2.1. For example, on the facet $\mathbf{F}_{01}$, consider the parameters

$$\mu_{10} = \log p(10)/p(00) = \theta_{10}, \qquad \mu_{11} = \log p(11)/p(00) = \theta_{10} + \theta_{01} + \theta_{11},$$
$$\mu_{01} = \log p(01)/p(00) = \theta_{01}.$$

Then μ_{10} and μ_{11} are identifiable parameters on $\mathcal{E}_{F_{01}}$, and μ_{01} diverges close to $\mathbf{F}_{01}$. By Lemma 8.2.1, the linear combinations that are well-defined are $\mu_{10} = \theta_{10}$ and $\mu_{11} = \theta_{10} + (\theta_{01} + \theta_{11})$. The above

table also lists θ_{00}, which is not a linear combination of other parameter, but θ_{00} is not free.

We obtain similar results for facets $\mathbf{F}_{01}$ and $\mathbf{F}_{11}$. The results are summarized in the following table:

facet	μ_{01}	μ_{10}	μ_{11}
$\mathbf{F}_{01}$	$-\infty$	finite	finite
$\mathbf{F}_{10}$	finite	$-\infty$	finite
$\mathbf{F}_{11}$	finite	finite	$-\infty$

Of course, by definition of μ_is, we cannot consider facet $\mathbf{F}_{00}$ where $n(00) = 0$. To study $\mathbf{F}_{00}$, we have to choose another zero cell and redefine the parameters μ_i.

The situation is more complicated for faces smaller than facets, because sending a single parameter to plus or minus infinity can be enough to send the distribution to a face F of higher codimension, as we will see below. The remaining parameters then determine the position within $\mathcal{E}_{\Delta,F}$. Thus, in this case there are more remaining parameters than the dimension of $\mathcal{E}_{\Delta,F}$.

For example, the data vector $n = (n_{00}, 0, n_{10}, 0)$ (with $n_{00}, n_{10} > 0$) lies on the face $\mathbf{F} = \mathbf{F}_{01} \cap \mathbf{F}_{11}$ of codimension two. If $p_{\theta^{(s)}} \to p_*$, then

$$\begin{aligned}
\theta_{00}^{(s)} &= \log p_{\theta^{(s)}}(00) \to \log \frac{n_{00}}{N}, \\
\theta_{01}^{(s)} &= \log \frac{p_{\theta^{(s)}}(01)}{p_{\theta^{(s)}}(00)} \to -\infty, \\
\theta_{10}^{(s)} &= \log \frac{p_{\theta^{(s)}}(10)}{p_{\theta^{(s)}}(00)} \to \log \frac{n_{10}}{n_{00}}.
\end{aligned}$$

However, the limit of $\theta_{11}^{(s)} = \log \frac{p_{\theta^{(s)}}(11) p_{\theta^{(s)}}(00)}{p_{\theta^{(s)}}(01) p_{\theta^{(s)}}(10)}$ is not determined. The only constraint is that $\theta_{11}^{(s)}$ cannot go to $+\infty$ faster than $\theta_{01}^{(s)}$ goes to $-\infty$, since $p_{\theta_{11}^{(s)}} = \exp(\theta_{00}^{(s)} + \theta_{01}^{(s)} + \theta_{10}^{(s)} + \theta_{11}^{(s)})$ has to converge to zero.

With the same data vector $n = (n_{00}, 0, n_{10}, 0)$, suppose we use a numerical algorithm to optimize the likelihood function by optimizing parameters θ_j in turn. To be precise, we order the parameters

θ_j in some way. For simplicity, say that the parameters are $\theta_1, \theta_2, \ldots, \theta_h$. Then we let

$$\theta_j^{(k+1)} = \arg\max_{y \in \mathbf{R}} l(\theta_1^{(k+1)}, \ldots, \theta_{j-1}^{(k+1)}, y, \theta_{j+1}^{(k)}, \ldots, \theta_h^{(k)})$$

This is called the *non-linear Gauss-Seidel method*. Let us choose the ordering $\theta_{01}, \theta_{10}, \theta_{11}$, where $\theta_{00} = -k(\theta)$ is not a free parameter. We start at $\theta_{01}^{(0)} = \theta_{10}^{(0)} = \theta_{11}^{(0)} = 0$. In the first step, we only look at θ_{01}. That is, we want to solve

$$\begin{aligned} 0 = \frac{\partial}{\partial \theta_{01}} l(\theta) &= -\frac{\exp(\theta_{01}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(0)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})} \\ &= -\frac{2\exp(\theta_{01}^{(1)})}{1 + 2\exp(\theta_{01}^{(1)})}. \quad \text{(C.0.1)} \end{aligned}$$

Clearly, the derivative is negative for any finite value of $\theta_{01}^{(1)}$, and thus the critical equation has no finite solution. If we try to solve this equation numerically, we will find that $\theta_{01}^{(1)}$ will be a large negative number. Next, we look at θ_{10}. We fix the other variables and try to solve

$$\begin{aligned} 0 = \frac{\partial}{\partial \theta_{10}} l(\theta) &= \frac{n_{10}}{N} - \frac{\exp(\theta_{10}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(0)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(0)})} \\ &\approx \frac{n_{10}}{N} - \frac{\exp(\theta_{10}^{(1)})}{1 + \exp(\theta_{10}^{(1)})}, \end{aligned}$$

where we have used that $\theta_{01}^{(1)}$ is a large negative number. This equation always has the unique solution

$$\theta_{10}^{(1)} \approx \log \frac{n_{10}}{N - n_{10}}.$$

Finally, we look at θ_{11}. We have to solve

$$0 = \frac{\partial}{\partial \theta_{11}} l(\theta) = -\frac{\exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(1)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(1)})}.$$

This equation has no solution, and therefore the numerical solution for $\theta_{11}^{(1)}$ should be close to numerical minus infinity. However, since $\theta_{01}^{(1)}$ is already close to $-\infty$, the equation is already

approximately satisfied. Thus, there is no need to change θ_{11}. In simulations, we observed that usually $\theta_{11}^{(1)}$ is negative, but not as small as $\theta_{01}^{(1)}$. In theory, we would have to iterate and now optimize θ_{01} again. But the values will not change much, since the critical equations are already satisfied to a high numerical precision after one iteration.

It is not difficult to see that the result is different if we change the order of the variables. If θ_{11} is optimized before θ_{01}, then θ_{11}^1 will in any case be a large negative number.

For general data, the derivative of with respect to θ_{01} (equation (C.0.1)) takes the form

$$\frac{\partial}{\partial\theta_{01}} l(\theta) = \frac{t_{01}}{N} - \frac{\exp(\theta_{01}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(0)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})}.$$

Setting this derivative to zero and solving for $\theta_{01}^{(1)}$ leads to a linear equation in $\theta_{01}^{(1)}$ with symbolic solution

$$\theta_{01}^{(1)} = \log \frac{1 + \exp(\theta_{10}^{(0)})}{1 + \exp(\theta_{10}^{(0)} + \theta_{11}^{(0)})} \frac{\frac{t_{01}}{N}}{1 - \frac{t_{01}}{N}}.$$

In fact, for any hierarchical model, the likelihood equation is linear in any single parameter θ_j, as long as all other parameters are kept fixed (more generally this is true when the design matrix A is a 0-1-matrix). Instead of optimizing the likelihood numerically with respect to one parameter, it is possible to use these symbolic solutions. This leads to the Iterative Proportional Fitting Procedure (IPFP). In our example, the IPFP would lead to a division by zero right in the first step, which indicates that the MLE does not exist.

D Parametrizations adapted to facial sets

Let us briefly discuss how to remedy problems 1. to 3. presented at the beginning of Section 8.2. The idea behind the remedy for 1. and 2. is to define parameters μ_i, $i \in L$, of $\mathcal{E}_A$, such that a subset $L_t \subseteq L$ of μ_i parametrizes $\mathcal{E}_{F_t,A}$ in a consistent way. Denote by $A^\mu = (a^\mu_{j,i}, j \in L, i \in I)$ the design matrix of $\mathcal{E}_A$ corresponding to the new parameters μ. Then the necessary conditions are:

(∗) Let $A^\mu_{L_t,F_t} := (a^\mu_{j,i}, j \in L_t, i \in F_t)$ be the submatrix of A^μ with rows indexed by L_t and columns indexed by L_t, and denote by $\tilde{A}^\mu_{L_t,F_t}$ the same matrix with an additional row of ones. The rank of $\tilde{A}^\mu_{L_t,F_t}$ is equal to $|L_t| + 1$, the number of its rows (and thus, $A^\mu_{L_t,F_t}$ has rank $|L_t|$).

(∗∗) $a^\mu_{j,i} = 0$ for all $i \in F_t$ and $j \in L \setminus L_t$.

In fact, (∗∗) implies that $A^\mu_{L_t,F_t}$ is the design matrix of $\mathcal{E}_{A,F_t}$, since the parameters μ_i with $i \notin L_t$ do not play a role in the parametrization $\mu \mapsto p_{F_t,\mu}$. Moreover, (∗) implies that the parametrization $\mu \mapsto p_{F_t,\mu}$ is identifiable. In this sense, we have remedied problem 1. from the beginning of the section.

Since the matrix $\tilde{A}^\mu_{L_t,F_t}$ has full row rank, it has a right inverse matrix $\tilde{C}$, such that $\tilde{A}^\mu_{L_t,F_t}\tilde{C} = I_{|L_t|+1}$ equals the identity matrix of size $|L_t| + 1$. Recall that

$$\log p_{F_t,\mu}(i) = \langle \mu^t, f^\mu_i \rangle - k_F(\mu),$$

$$\log p_\mu(i) = \langle \tilde{\mu}^t, f^\mu_i \rangle - k(\mu),$$

for any parameter vector μ and all $i \in F_t$. Since f_i^μ are the columns of A^μ and since the components of f_i^μ corresponding to $L \setminus L_t$ vanish according to $(**)$, we may apply matrix C obtained from $\tilde{C}$ by dropping the row corresponding to k_F or k and obtain

$$(\log p_\mu)C = \mu_{L_t} \quad \text{and} \quad (\log p_{F_t,\mu})C = \mu_L. \tag{D.0.1}$$

When $p_{\mu^{(s)}}$ is a sequence in $\mathcal{E}_A$ with limit p_μ in $\mathcal{E}_{F_t,A}$, then (D.0.1) shows that $\mu_i^{(s)} \to \mu_i$ for $i \in L_t$. In this sense, we have remedied problem 2.

Finally, we solve problem 3. Suppose that we have chosen parameters μ_L as in Section 8.2, and let A^{μ_L} be the design matrix with respect to these parameters. Then $(A^{\mu_L})_{j,i} = 0$ if $i \in F_t$ and $j \notin L_t$. Moreover, for $j \in L_t$, the jth column of A_{μ_L} has a single non-vanishing entry (equal to one) at position j. Suppose that F_t corresponds to a face $\mathbf{F}_t$ of codimension c. Then there are c facets of $\mathbf{P}$ whose intersection is $\mathbf{F}_t$. Thus, following the notation introduced in Remark 2.3.1, there exist c inequalities

$$\langle \tilde{g}_1, \tilde{x} \rangle \geq 0, \quad \ldots, \quad \langle \tilde{g}_c, \tilde{x} \rangle \geq 0 \tag{D.0.2}$$

that together define $\mathbf{F}_t$. In this case, vectors $\tilde{g}_1, \ldots, \tilde{g}_c$ are linearly independent and satisfy $\langle \tilde{g}_j, \tilde{f}_i \rangle = 0$; thus, they are a basis of the kernel of $(\tilde{A}^{\mu_L}_{F_t})^t$. It follows that the kth component of g_j, denoted by $g_{j,k}$, vanishes if $k \in L_t$; that is, the inequalities (D.0.2) do not involve the variables corresponding to L_t. Let G be the square matrix, indexed by $L \setminus L_t$ with entries $g_{j,k}$, $j, k \in L \setminus L_t$. Then the square matrix

$$\tilde{G} = \begin{pmatrix} 1 & 0 \\ 0 & G \end{pmatrix}$$

is invertible. We claim that parameters $\lambda = \tilde{G}^{-1}\mu_L$ are what we are looking for.

The design matrix with respect to the parameters λ is $A^\lambda = \tilde{G}A^{\mu_L}$. For any $j \notin L_t$,

$$A^\lambda_{j,i} = 0, \quad \text{if } i \in F_t, \qquad \text{and} \quad A^\lambda_{j,i} = \langle \tilde{g}_j, \tilde{f}_i \rangle \geq 0, \quad \text{if } i \notin F_t.$$

This implies the following properties:

1. If all parameters λ_j with $j \notin L_t$ are sent to $-\infty$, then p_λ tends towards a limit distribution with support F_t.

2. The coefficient of λ_j in any log-probability is non-negative, so there is no cancellation of $\pm\infty$.

So far, we only used the fact that vectors $\tilde{g}_j$ define valid inequalities for the face $\mathbf{F}_t$. Suppose that we choose $\tilde{g}_j$ in such a way that each inequality $\langle \tilde{g}_j, \tilde{x} \rangle \geq 0$ defines a facet. The intersection of less than c facets is a face that strictly contains $\mathbf{F}_t$. This implies that for each j, there exists $i_j \in I \setminus F_t$ such that f_{i_j} satisfies

$$\langle \tilde{g}_j, \tilde{f}_{i_j} \rangle > 0, \qquad \text{and} \quad \langle \tilde{g}_{j'}, \tilde{f}_{i_j} \rangle = 0 \text{ for all } j' \neq j.$$

This implies

$$A^\lambda_{j,i_j} > 0, \qquad \text{and} \quad A^\lambda_{j'}, i_j = 0 \text{ for all } j' \neq j.$$

This implies the following:

3. If $\lambda^{(s)}_j$ are sequences of parameters such that $p_{\lambda^{(s)}}$ tends towards a limit distribution with support F_t, then $\lambda^{(s)}_j \to -\infty$ for all $j \notin L_t$.

It is not difficult to see that, conversely, any parametrization that satisfies these three properties comes from facets defining the face $\mathbf{F}_t$.