Local conditional and marginal approach to parameter estimation in discrete graphical models

By Hélène Massam[*]and Nanwei Wang
Department of Mathematics and Statistics, York University,
Toronto, ON M3J 1P3, Canada

March 24, 2016

Abstract

Discrete graphical models are an essential tool in the identification of the relationship between variables in complex high-dimensional problems. When the number of variables p is large, computing the maximum likelihood estimate (henceforth abbreviated mle) of the parameter is difficult. A popular approach is to estimate the composite mle rather than the mle, that is the value of the parameter that maximizes the product of local conditional likelihoods centered around each vertex v of the graph underlying the model. A more recent development is to have the components of the composite likelihood be local marginal likelihoods centered around each v.

The purpose of this paper is to first show that the estimates obtained through local conditional and marginal likelihoods are identical. Second, we study the asymptotic properties of the composite mle obtained by averaging of the local estimates: this is done under the double asymptotic regime when both p and N go to infinity and compare the rate of convergence to the true parameter with that of the global mle under the same conditions. We also look at the simple asymptotic regime where p fixed and thus recover results by Liu and Ihler (2012).

Key words: discrete graphical models, distributed estimation, local conditional, local marginal, composite likelihood, "large p, large N" asymptotics. *AMS 2000 Subject classifications.* 62H17 (Primary), 62M40.

[*] H. Massam gratefully acknowledges support from NSERC Discovery Grant No A8947.

1 Introduction

Discrete graphical models are an essential tool in the analysis of complex high-dimensional categorical data. Let $V = \{1, \dots, p\}$ be a finite index set. Let $G = (V, E)$ be an undirected graph where E is the set of undirected edges in $V \times V$. Then the distribution of $X = (X_v, v \in V)$ is said to be Markov with respect to G if X_v is independent of X_u given $X_{V \setminus \{u,v\}}$ whenever the edge (v, u) is not in E. The set of distributions Markov with respect to a given graph G is called a graphical model. When the variables X_v take values in a finite set $I_v, v \in V$, the graphical model is said to be discrete. These models are used extensively to represent interactions between individuals in physical or human networks. Each data point is classified according to its values of $X_v = i_v,\ i_v \in I_v, v \in V$ and the data is thus gathered in a p-dimensional contingency table with cells $i = (i_v, v \in V)$ and cell counts $n(i), i \in I = \prod_{v \in V} I_v$. As we shall recall in Section 2, the density of the cell counts can be written under a natural exponential family form as

$$f(t; \theta) = \exp\{\langle \theta, t \rangle - Nk(\theta)\} \tag{1.1}$$

where t is a vector of marginal cell counts, $\langle \theta, t \rangle$ denotes the inner product of $t = t(x)$ and θ is the canonical loglinear parameter.

For a given data set, the first task is to learn the underlying graph and once the underlying graph has been learnt, the second task is to estimate the parameter θ of the model. In this paper we will be concerned with the maximum likelihood estimate (mle) of θ. When p is large, to obtain the mle of θ through a simple maximization of the likelihood function is impossible because of the dimension of the parameter θ and the complexity of the cumulant generating function $k(\theta)$ in (1.1). 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 $x^{(1)}, \dots, x^{(N)}$, a composite likelihood is typically the product of local conditional likelihoods, coming from the local conditional probability of X_v given $X_{\mathcal{N}_v},\ v \in V$, which we can write as

$$L^{PS}(\theta) = \prod_{v \in V} \prod_{k=1}^{N} p(X_v = x_v^{(k)} | X_{\mathcal{N}_v} = x_{\mathcal{N}_v}^{(k)}; \theta^{v,PS}) \tag{1.2}$$

where $\mathcal{N}_v$ indices the set of neighbours of v in G, and $\theta^{v,PS}$ is a subvector of θ.

Further in the vein of composite likelihood, recent research has focused on studying each local model and combining all the local results to yield a global estimate of either the underlying structure G or the parameter θ. For example, for model selection, with p large, Ravikumar et al. (2010) introduced a local approach to discrete graphical model selection by looking at the regularized local conditional likelihood of X_v given $X_{V\setminus\{v\}}$, that is, due to the Markov properties of the model, given $X_{\mathcal{N}_v}$. The aim is to identify the components of $\theta^{v,PS}$, related to the interaction of v and its neighbours, that are not equal to zero. For parameter estimation, with p large and G given, the local approach has been used for Gaussian graphical models by Wiesel and Hero (2012) who consider the composite likelihood based on local conditional likelihoods of X_v given $X_{\mathcal{N}_v}, v \in V$. To obtain the maximum composite likelihood estimate, the estimates obtained through each local likelihood are combined using the ADDM optimization technique. For discrete models, Liu and Ihler (2012) study the asymptotic properties, for p fixed and N going to infinity, of a maximum composite likelihood estimate obtained through either an optimal linear combination of the estimates of the components of $\theta^{v,PS}$ from different local conditional models (linear consensus) or through the choice of a "best", in some sense, such estimate (maximum consensus).

For the estimation of the precision matrix in graphical Gaussian models, Meng et al. (2014) depart from the ideas of the two papers just mentioned, in two ways. First, they do not consider local conditional models but rather local marginal models. Second they do not look only at "one-hop" marginal models, i.e., models built on v and its neighbours $\mathcal{N}_v$ but they consider "two-hop" local marginal models that is marginal models with vertex set a vertex v, its neighbours and the neighbours of the neighbours. With the two-hop local marginal likelihoods, they achieve such accuracy that, to obtain the overall estimate of the parameter, they need not use a method more sophisticated than simple averaging of the various local marginal likelihood estimates. While they prove that for the one-hop case, the estimates obtained from local marginal models are identical to those obtained from local conditional models, they do not make the same statement for the estimates obtained from maximizing two-hop local marginal and conditional likelihoods.

In this paper, we are concerned with the maximum composite likelihood estimation of θ in (1.1) for discrete graphical models and our purpose is twofold. First we extend the local marginal method of Meng et al. (2014) to discrete graphical models and show that, actually the estimates of the parameters obtained from these local marginal likelihoods are equal to the estimates obtained from the more traditional local conditional likelihoods and this holds whether we are looking at one-hop or two-hop neighbourhoods. Given the complexity of computations for local marginal likelihoods, we suggest one should therefore work only with local conditional likelihoods. We then define our maximum composite likelihood estimate of θ in the following way: if a component θ_j of θ is obtained from one local conditional model only, then this will be the estimate of θ_j. If the same component θ_j is obtained from m_j different local conditional models, then the estimate of θ_j is the average of the estimates obtained from the m_j local marginal models. This is a particular case of "linear consensus" as defined in Liu and Ihler (2012). The second aim of our paper is to 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, and, when both p and N tend to infinity.

In Section 3.1, we first recall the definition of composite likelihood based on conditional local likelihoods. The most important feature is that the parameter of each local conditional likelihood which we denote $\theta^{v,PS}$ (see (1.2)) is a subvector of the parameter θ of the global model in (1.1). Then, following what was done in Meng et al. (2014), we define a relaxed local marginal likelihood and show that the parameter of this local likelihood, denoted $\theta^{\mathcal{M}_{l,v}}$ contains $\theta^{v,PS}$ also. In Section 3.4, we show our first main result, Theorem 3.1, which states that the estimate of $\theta^{v,PS}$ obtained from local marginal and conditional likelihoods are identical. We illustrate this results with numerical examples. It is interesting to note, at this point, that Mizrahi et al. (2014) who developped, also for discrete models, a local marginal composite likelihood method centered around cliques rather than vertices find that the performance of their new method is "basically indistinguishable from that of the pseudolikelihood". Though we have not verified it analytically, we conjecture that the estimates obtained by their LAP-D and LAP-E method are equal to the estimates obtained by pseudolikelihood.

In Section 4, we then look at the properties of our maximum composite likelihood estimate of θ. We study its asymptotic properties under the classical and double asymptotic regime. Our main result, Theorem 4.2, states that, when both p and N go to infinity, under certain conditions, Conditions A and B, for $\frac{N}{\log p}$ large enough, our estimate is close to the true value of the parameter with high probability. Conditions A and B are similar to the "Dependency" condition of Ravikumar et al. (2010) for model selection. The Dependency condition are conditions on the variance function, or Fisher information matrix, of the local conditional model that roughly state that the maximum eigenvalue of this variance function is bounded above and the minimum eigenvalue is bounded away from zero. Our Conditions A and B impose the same type of condition but on the sum, over $v \in V$, of the local variance functions. Our result under the classical regime, Theorem 4.1, where p is fixed coincides with Theorem 4.1 of Liu and Ihler, 2012 and is given here for the sake of completness .

Before proceeding to the next section, we ought to make some important remarks. First when computing the estimates from the local conditional likelihoods, we need to make sure that they exist, that is that there exists finite estimates of $\theta^{v,PS}$ that maximize the local conditional likelihood. If they do not exist, our maximization software may return values that are erroneous. It may happen also that the global maximum likelihood estimate of θ does not exist and yet the local estimates of $\theta^{v,PS}$ exist and we can obtain a maximum composite likelihood estimate of the parameter. We expand on these points in Lemma 3.3. Techniques to identify the existence of the global maximum likelihood estimate of θ have been developed in Fienberg and Rinaldo (2012) and Wang et al. (2016). In the present paper, we will assume that all local estimates exist.

Our second remark is that in the sequel, we will only consider graphs that are not reducible: a graph G is reducible if there exist three disjoint subsets A, B, C of V with $V = A \cup B \cup C$ such that every path from A to B goes through C and such that the graph G_C induced from G by C is complete, i.e. every vertex in C is linked to any other vertex in C by an edge. If G can be so decomposed, then, we decompose each component $G_{A\cup C}$ and $G_{C\cup B}$ and so on until the smallest components thus obtained are prime components, i.e. nondecomposable induced subgraphs that are maximal with respect to inclusion. It is easy to show that the prime components thus obtained can be ordered into a perfect sequence $P_1, \dots, P_k$ of components that, for any $i = 2, \dots, k$, there exists $j < i$ such that

$$P_i \cap \Big(\cup_{l=1}^{i-1} P_l \Big) \subset P_j \quad \text{and} \quad S_j = P_j \cap \Big(\cup_{l=1}^{j-1} P_l \Big) \text{ is complete.}$$

In that case, it is well-known, that the cell probabilities $p(i) = P(X = i),\ i \in I$ can be expressed analytically in terms of the cell probabilities $p^{P_l}(i_{P_l})$ and $p^{S_l}(i_{S_l})$ in the P_l-marginal and S_l-marginal, $l = 1, \dots, k$ models respectively, as follows

$$p(i) = \frac{\prod_{l=1}^{k} p^{P_l}(i_{P_l})}{\prod_{l=2}^{k} p^{S_l}(i_{S_l})}.$$

Since, as we shall see in the next section, knowing $p(i), i \in I$ in a given model is equivalent to knowing θ as in (1.1), it is sufficient to work on the induced graphs $G_{P_l}, l = 1, \dots, k$. Thus in the sequel, all graphs considered in this paper are irreducible prime graphs. In this case, there is no possibility to see cuts in the natural exponential family (1.1), that is no possibility to split the parameter θ into functionally independent components and the task at hand is to estimate θ.

2 Preliminaries

2.1 Discrete graphical and hierarchical loglinear models

Let p, V and $X = (X_v,\ v \in V)$ be as described in Section 1 above. If N individuals are classified according to the p criteria, the resulting counts are gathered in a contingency table such that

$$I = \prod_{v \in V} I_v$$

is the set of cells $i = (i_v,\ v \in V)$. For $D \subset V$, i_D denotes the marginal cell $i_D = (i_v, v \in D)$ with $i_v \in I_v$. Let $\mathcal{D}$ be a family of non empty subsets of V such that $D \in \mathcal{D}$, $D_1 \subset D$ and $D_1 \neq \emptyset$ implies $D_1 \in \mathcal{D}$. In order to avoid trivialities we assume $\cup_{D\in\mathcal{D}} D = V$. The family $\mathcal{D}$ is called the generating class of the hierarchical loglinear model. We denote by $\Omega_{\mathcal{D}}$ the linear subspace of $y \in \mathbb{R}^I$ such that there exist functions $\theta_D \in \mathbb{R}^I$ for $D \in \mathcal{D}$ depending only on i_D and such that $y = \sum_{D\in\mathcal{D}} \theta_D$, that is

$$\Omega_{\mathcal{D}} = \{y \in \mathbb{R}^I :\ \exists \theta_D \in \mathbb{R}^I, D \in \mathcal{D} \text{ such that } \theta_D(i) = \theta_D(i_D) \text{ and } y = \sum_{D\in\mathcal{D}} \theta_D\}$$

The hierarchical model generated by $\mathcal{D}$ is the set of probabilities $p = (p(i))_{i\in I}$ on I such that $p(i) > 0$ for all i and such that $\log p \in \Omega_{\mathcal{D}}$.

The class of discrete graphical models Markov with respect to an undirected graph G is a subclass of the class of hierarchical discrete loglinear models. Indeed, let $G = (V, E)$ be an undirected graph where V is the set of vertices and $E \subset V \times V$ denotes the set of undirected edges. We say that the distribution of X is Markov with respect to G if $(v_1, v_2) \notin E$ implies

$$X_{v_1} \perp X_{v_2} |\ X_{V\setminus\{v_1,v_2\}}.$$

Let $\mathcal{D}$ be the set of all cliques (not necessarily maximal) of the graph G. If the distribution of $X = (X_1, \ldots, X_p)$ is Multinomial$(1, p(i), i \in I)$ Markov with respect to the graph G, and if we assume that all $p(i), i \in I$, are positive, then, by the Hammersley-Clifford theorem, $\log p(i)$ is a linear function of parameters dependent on the marginal cells $i_D, D \in \mathcal{D}$ only, and therefore the graphical model is a hierarchical loglinear model with generating set the set $\mathcal{D}$ of cliques of G. The reader is referred to Darroch & Speed (1983), Lauritzen (1996) or Letac & Massam (2012) for a detailed description of the hierarchical loglinear model and the subclass of discrete graphical loglinear models.

We now set our notation and recall some basic results for discrete hierarchical loglinear models. The following notation and results can be found in Letac & Massam (2012) and the corresponding supplementary file.

Among all the values that X_v can take in $I_v, v \in V$, we call one of them 0. For a cell $i \in I$, we define its support $S(i)$ as

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

and we define also the following subset J of I

$$J = \{j \in I,\quad S(j) \in \mathcal{D}\}. \tag{2.1}$$

From here on, we will call this set the J-set of the model. For $i \in I$ and $j \in J$, we define the symbol

$$j \triangleleft i$$

to mean that $S(j)$ is contained in $S(i)$ and that $j_{S(j)} = i_{S(j)}$. The relation $\lhd$ has the property that if $j, j' \in J$ and $i \in I$, then

$$j \lhd j' \text{ and } j' \lhd i \Rightarrow j \lhd i.$$

The loglinear parametrization that we use for the multinomial is the so-called baseline parametrization with general expression, for $i \in I, S(i) = E \subset V$,

$$\theta_i = \sum_{F \subset E} (-1)^{|E|-|F|} \log p(i_F, 0_{V \setminus F}) \; . \tag{2.2}$$

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

$$\begin{aligned} \theta_j &= \sum_{j' \in J,\, j' \lhd j} (-1)^{|S(j)|-|S(j')|} \log \frac{p(j')}{p(0)},\; j \in J \\ \log p(i) &= \theta_0 + \sum_{j \in J, j \lhd i} \theta_j,\; i \in I \qquad (2.3) \\ \log p(0) &= \theta_0. \qquad (2.4) \end{aligned}$$

One then readily derives the density of the multinomial $\mathrm{M}(N, p(i), i \in I)$ of the cell counts $\underline{n} = (n(i), i \in I)$, Markov with respect to G to be, up to a multiplicative constant, equal to

$$f(t;\, \theta) = \exp\{\langle t, \theta \rangle - Nk(\theta)\},\; \theta \in R^J \tag{2.5}$$

with $\theta = (\theta_j, j \in J),\; t = t(\underline{n}) = (t(j),\; j \in J)$ where $t(j) = n(j_{S(j)})$ are the $j_{S(j)}$-marginal cell counts and

$$k(\theta) = \log \Big(\sum_{i \in I} \exp \sum_{j \in J, j \lhd i} \theta_j \Big) = \log \Big(1 + \sum_{i \in I \setminus \{0\}} \exp \sum_{j \in J, j \lhd i} \theta_j \Big) \; . \tag{2.6}$$

For $\theta \in R^J$, these distributions form a natural exponential family of dimension J generated by a measure μ which we will now identify. Let $e_j, j \in J$ be the canonical basis of R^J and, for $i \in I$, let

$$f_i = \sum_{j \in J, j \lhd i} e_j. \tag{2.7}$$

Then (2.3) and (2.4) can be written in matrix form as

$$\log p \;=\; A\tilde{\theta} \tag{2.8}$$

where $\tilde{\theta}^t = (\theta_0, \theta^t)$, A is an $(|I|) \times (1 + |J|)$ matrix. We call A the design matrix of the model. The rows of A are indexed by $i \in I$ and equal to $\tilde{f_i}^t = (1, f_i^t) \in \mathbb{R}^{J+1}$. It is immediate to see that the Laplace transform of the generating measure μ is

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

and therefore the measure μ generating (2.5) is

$$\mu(dx) = \sum_{i \in I} \delta_{f_i}(x). \tag{2.9}$$

This exponential family is concentrated on the convex hull of $f_i, i \in I$, which is a bounded set of $\mathbb{R}^J$, and therefore the set of parameters θ for which $k(\theta)$ is finite is the whole space $\mathbb{R}^J$. From the definition of X, $f_i, i \in I$ and $t = (t(j_{S(j)}, j \in j)$, it is easy to see that $(N, t(j),\ j \in J)^t = A^t n = \sum_{i \in I} n(i) \tilde{f_i}$ and the vector of sufficient statistics t, which we also write as $t = t_J$ to emphasize its length, is such that

$$\frac{t_J}{N} = \Big(\frac{t(j)}{N}, j \in J\Big)^t = \sum_{i \in I \setminus \{0\}} \frac{n(i)}{N} f_i = \sum_{i \in I} \frac{n(i)}{N} f_i \tag{2.10}$$

and thus belongs to the convex hull of $(f_i)_{i \in I}$. The $(f_i)'s$ are the extreme points of the closure of the convex hull of the $f_i, i \in I$.

3 The conditional and marginal composite maximum likelihood estimators

When the dimension of the discrete graphical model is large, computing the maximum likelihood estimate of θ in (2.5) is challenging, if not impossible. As mentioned in the introduction, a recent approach to this problem has been local with the use of a composite likelihood which is equal to 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 G. Recently, for Gaussian high-dimensional graphical models, Wiesel & Hero (2012) and Meng & al. (2013, 2014) worked with a different composite likelihood which is the product, over all vertices $v \in V$, of local marginal likelihoods. In this section, we will first recall the definition of the conditional composite likelihood estimate, then extend the marginal composite likelihood 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. Since the computational complexity of the marginal computations is exponential in the number of vertices in the neighbourhood of v while the conditional computations are linear in this number, there is no advantage in working with marginal composite likelihoods.

3.1 The conditional composite likelihood function

We first define the standard conditional composite likelihood function. For $i = (i_v, v \in V)$, let $i^{(1)}, \dots, i^{(N)}$ be a sample of size N from the distribution of X Markov with respect to G. We recall that the global likelihood function is

$$L(\theta) \propto \prod_{k=1}^{N} p(X_v = i_v^{(k)}, v \in V) = \exp\{\langle \theta, t \rangle - Nk(\theta)\} \tag{3.1}$$

where $k(\theta)$ is as in (2.6).

For a given vertex $v \in V$, let $\mathcal{N}_v$ 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

$$L^{v,PS}(\theta) = \prod_{k=1}^{N} p(X_v = i_v^{(k)} | X_{\mathcal{N}_v} = i_{\mathcal{N}_v}^{(k)}; \theta) \tag{3.2}$$

and the superscript PS stands for "pseudo-likelihood", the name often given to the conditional composite lilelihood (Besag, 1974). As given by (2.3), for a given cell i, we have

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

The set J is as defined in (2.1) for the global model. Let

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

Then for $i_v \neq 0$, we have

$$\begin{aligned} p(X_v = i_v|\ X_{\mathcal{N}_v} = i_{\mathcal{N}_v}) = p(X_v = i_v|\ X_{V\setminus\{v\}} = i_{V\setminus\{v\}}) &= \frac{p(X_V = i_V)}{p(X_{V\setminus\{v\}} = i_{V\setminus\{v\}})} \\ = \qquad & \frac{e^{\theta_0 + \sum_{j \lhd i,\ j \in J^{PS_v}} \theta_j + \sum_{j \lhd i,\ S(j) \subseteq \mathcal{N}_v} \theta_j + \sum_{j \lhd i,\ S(j) \not\subseteq v \cup \mathcal{N}_v} \theta_j}}{\sum_{k \in I|\ k_{V\setminus\{v\}} = i_{V\setminus\{v\}}} \left(e^{\theta_0 + \sum_{j \lhd k,\ j \in J^{PS_v}} \theta_j + \sum_{j \lhd k,\ S(j) \subseteq \mathcal{N}_v} \theta_j + \sum_{j \lhd k,\ S(j) \not\subseteq v \cup \mathcal{N}_v} \theta_j} \right)} \\ = \qquad & \frac{e^{\sum_{j \lhd i,\ j \in J^{PS_v}} \theta_j}}{1 + \sum_{k \in I|\ k_{V\setminus\{v\}} = i_{V\setminus\{v\}},\ k_v \neq 0} e^{\sum_{j \lhd k,\ j \in J^{PS_v}} \theta_j}} \qquad (3.3) \end{aligned}$$

and

$$p(X_v = 0|\ X_{V\setminus\{v\}} \quad = \quad i_{V\setminus\{v\}}) = \frac{1}{1 + \sum_{k \in I|\ k_{V\setminus\{v\}} = i_{V\setminus\{v\}},\ k_v \neq 0} e^{\sum_{j \lhd k,\ j \in J^{PS_v}} \theta_j}} \qquad (3.4)$$

Equality (3.3) 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 3.1 *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, in order to emphasize that the parameter*

$$\theta^{v,PS} = (\theta_j,\ j \in J^{PS_v}), \quad v \in V \qquad (3.5)$$

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

We now define the two-hop conditional composite likelihood function.

Definition 3.1 *For a given $v \in V$, we will say that $\mathcal{M}_v$ is a one-hop neighbourhood 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 neighbourhood if it comprises v, its immediate neighbours and the neighbours of the immediate neighbours in G. 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. For simplicity of notation, we will denote both the one-hop and two-hop neighbourhoods by $\mathcal{M}_v$.

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

$$L^{v,PS_2}(\theta) = \prod_{k=1}^{N} p(X_v = i_v^{(k)}, X_{\mathcal{N}_v} = i_{\mathcal{N}_v}^{(k)} | X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}^{(k)}). \tag{3.6}$$

The expression of $p(X_v = i_v^{(k)}, X_{\mathcal{N}_v} = i_{\mathcal{N}_v}^{(k)} | X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}^{(k)})$ is the same as (3.3) and (3.4) 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 3.1, we note that

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

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

3.2 The marginal composite likelihood

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

$$L^{\mathcal{M}}(\theta) = \prod_{v\in V}\prod_{k=1}^{N} p(X_{\mathcal{M}_v} = i_{\mathcal{M}_v}^{(k)}) = \prod_{v\in V} L^{\mathcal{M}_v}(\theta). \tag{3.7}$$

where $L^{\mathcal{M}_v}(\theta) = \prod_{k=1}^{N} p(X_{\mathcal{M}_v} = i_{\mathcal{M}_v}^{(k)})$. 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

$$f(t^{\mathcal{M}_v};\ \theta^{\mathcal{M}_v}) = \exp\{\langle t^{\mathcal{M}_v}, \theta^{\mathcal{M}_v}\rangle - Nk^{\mathcal{M}_v}(\theta^{\mathcal{M}_v})\} \tag{3.8}$$

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 cumulant generating function.

In order to identify the $\mathcal{M}_v$-marginal model, we first establish the relationship between θ and $\theta^{\mathcal{M}_v}$. In the sequel, the symbol j will be understood to be an element of $I_{\mathcal{M}_v}$ when used in the notation $\theta_j^{\mathcal{M}_v}$ while it will be understood to be the element of J obtained by padding it with entries $j_{V\setminus\mathcal{M}_v} = 0$ when 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. Proofs are given in the Appendix.

Lemma 3.1 *Let $\mathcal{M}_v$ be the one-hop or two-hop neighbourhood 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:*

$$\theta_j^{\mathcal{M}_v} = \qquad \theta_j + \sum_{j' \mid j' \lhd_0 j} (-1)^{|S(j)-S(j')|} \log \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v} = j'} \exp \sum_{\substack{k \mid k \lhd i \\ k \ntrianglelefteq j'}} \theta_k \Big) . \tag{3.9}$$

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 0 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:

$$\mathcal{B}_v = \{w \in \mathcal{M}_v \mid \exists w' \in \mathcal{M}_v^c \text{ with } (w, w') \in E\}. \tag{3.10}$$

We have the following result.

Lemma 3.2 *Let $\mathcal{M}_v$ be the one-hop or two-hop neighbourhood 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 (3.9) 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 $\theta_i = 0$.

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 loglinear 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 loglinear parameter is zero in the global model, it remains zero in the $\mathcal{M}_v$-marginal model.

3.3 A convex relaxation of the local marginal optimization problems

It is clear from (3.9) that even though maximizing the marginal likelihood from (3.8) is convex in $\theta^{\mathcal{M}_v}$, it is not convex in θ. We would therefore like to replace the problem of maximizing (3.8) non convex in θ by a convex relaxation problem. We know from *(1.)* of Lemma 3.2 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 3.2 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 what has been done for Gaussian graphical models in 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 loglinear 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 index l takes values $l = 1$ or $l = 2$ when $\mathcal{M}_v$ is respectively the one-hop or two-hop neighbourhood of v. The J-set of this local model is

$$J^{\mathcal{M}_{l,v}} = \{j \in J \mid S(j) \subset \mathcal{M}_v, S(j) \not\subset \mathcal{B}_v\} \cup \{i \in I \mid S(i) \subset \mathcal{B}_v\}\ . \tag{3.11}$$

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 loglikelihood

$$L^{\mathcal{M}_{l,v}}(\theta) = \prod_{k=1}^{N} p^{\mathcal{M}_{l,v}}(X_{\mathcal{M}_v} = i^{(k)}_{\mathcal{M}_v}) = \exp\{\langle \theta^{\mathcal{M}_{l,v}}, t^{\mathcal{M}_{l,v}} \rangle - N k^{\mathcal{M}_{l,v}}(\theta^{\mathcal{M}_{l,v}})\} \tag{3.12}$$

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 3.2 *The vector $\theta^{v,PS}$ defined in (3.5) is a subvector of $\theta^{\mathcal{M}_{l,v}}$. Therefore maximizing (3.12) for either $l =$ or $l = 2$ will yield an estimate of $\theta^{v,PS}$.*

Remark 3.3 *The $\mathcal{M}_{l,v},\ l = 1, 2$-marginal model is a hierarchical loglinear 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*

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

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 loglinear models rather than the more restrictive class of discrete graphical models.

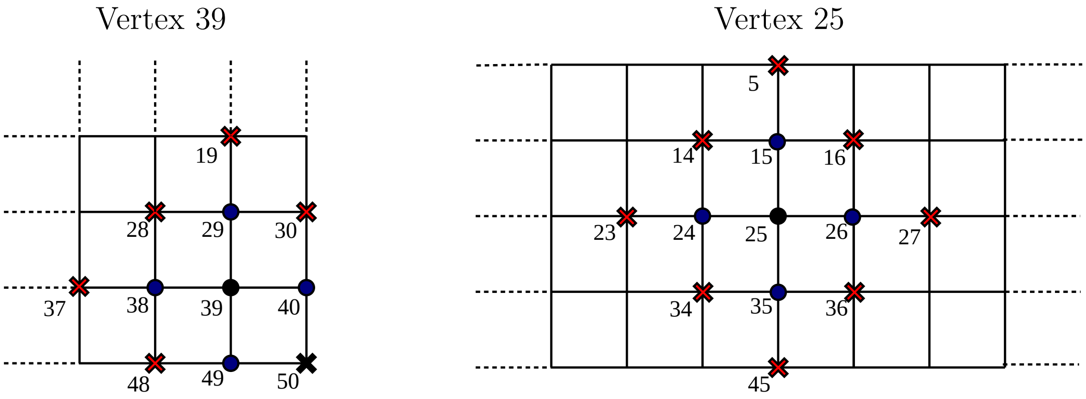

Figure 1: Two vertices in a 5×10 lattice: Theorem 3.1 applies for vertex 25 while it does not apply for vertex 39.

3.4 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 likelihood (3.12).

Theorem 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 (3.2). 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 (3.6).*

The proof is given in the Appendix.

At this point, we ought to make an important observation. In the case of the two-hop marginal likelihood, it may happen that the buffer $\mathcal{B}_v$ is no longer equal 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 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 1 where for the sake of comparison, we also look at vertex 25 for which $\mathcal{N}_{2v} = \mathcal{B}_v$ and Theorem 3.1 applies..

In Tables 1 and 2, we give the numerical values of the maximum likelihood estimate $\theta_j, j \in J^{\mathcal{M}_{2,v}}$ obtained by the four local model $PS, PS_2, \mathcal{M}_{1,v}$ 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 through 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 3.1 does not break down.

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

Remark 3.4 *The equality of the estimates holds also for the marginal estimates obtained by 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 verified 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 3.2 shows.*

3.5 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, given the computational complexity in the relaxed marginal model, we will work only with the local estimates obtained from local conditional lilkelihoods. 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

$$\hat{\theta}^v = \begin{cases} \hat{\theta}^{v,PS} & \text{if we work with } l^{v,PS} \\ (\hat{\theta}_j^{v,PS_2},\ S(j) \subset \{v\} \cup \mathcal{N}_v) & \text{if we work with } l^{v,PS_2}\ . \end{cases} \tag{3.13}$$

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

$$\bar{\theta} = (\bar{\theta}_j = \frac{\sum_{l=1}^{m_j} \hat{\theta}_j^{v_l}}{m_j},\ j \in J), \tag{3.14}$$

which from now on, we will abbreviate by "mcle".

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 (3.5). If $S(j) = \{v\}$, clearly, the row of A corresponding to $\bar{\theta}_j$ has all its entries equal to 0 except for one entry equal to 1 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 0 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

$$\bar{\theta} = \begin{pmatrix} \bar{\theta}_a \\ \bar{\theta}_{ab} \\ \bar{\theta}_b \\ \bar{\theta}_{bd} \\ \bar{\theta}_c \\ \bar{\theta}_{cd} \\ \bar{\theta}_d \\ \bar{\theta}_{db} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 \end{pmatrix} \begin{pmatrix} \hat{\theta}_a^a \\ \hat{\theta}_{ab}^a \\ \hat{\theta}_{ac}^a \\ \hat{\theta}_b^b \\ \hat{\theta}_{ab}^b \\ \hat{\theta}_{bd}^b \\ \hat{\theta}_c^c \\ \hat{\theta}_{ca}^c \\ \hat{\theta}_{cd}^c \\ \hat{\theta}_d^d \\ \hat{\theta}_{bd}^d \\ \hat{\theta}_{cd}^d \end{pmatrix}.$$

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

$$A_{j,k} = \begin{cases} \frac{1}{m_j} & \text{if} \quad j_{v_l \cup \mathcal{N}_{v_l}} = k \in J^{v_l,PS},\ l = 1, \ldots, m_j \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

We have now defined our mcle which we use to replace the global mle maximizing (3.1). It is natural to ask whether the mcle exists when the global mle exists and conversely whether 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 the corresponding cell probabilities $p(i)$ and $p(0)$ as given by (2.3) and (2.4) 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 $(p(i), i \in I$, as the following lemma shows, the global mle may not exist but we may accept still accept the mcle as an estimate of the parameter.

Lemma 3.3 *For a discrete log-linear model, if the global mle exists, then the mcle exists. However, the mcle may exist and yet the global mle does not.*

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

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

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 and Massam (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) references therein, this implies that the global mle does not exist. The facets corresponding to the local models built on $v = a$ have equation

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

We can verify immediately that none of these equations are satisfied with the given data and therefore the mle iof $\theta^{v,PS}$ inn the a-local model. Similarly the mle of $\theta^{v,PS}$, $v = b, c, d$ exist and thus the mcle exist. □

4 Asymptotic properties of the maximum composite likelihood estimate

In this section, 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 in the case p is fixed have been given by Liu and Ihler (2012), we give our result here in Section 4.1 for completeness in our own notation.

4.1 The classical asymptotic regime

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

Theorem 4.1 *The mcle $\bar{\theta}$ as defined in (3.14) is asymptotically consistent and*

$$\sqrt{N}(\bar{\theta} - \theta^*) \to N(0, AGA^t) \tag{4.1}$$

where A is as defined in (3.15), G is the square $\sum_{v \in V} |J^{v,PS}|$-dimensional matrix with (v_l, v_m)-block entry

$$G_{v_l,v_m} = I^{-1}(\theta^{v_l,*}) E\big(\frac{\partial l(\theta^{*v_l})}{\partial \theta^{*v_l}} \left(\frac{\partial l(\theta^{*v_m})}{\partial \theta^{*v_m}}\right)^t\big) I^{-1}(\theta^{*v_m}), \tag{4.2}$$

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

$$NE(\|\bar{\theta}_j - \theta^*_j\|^2) \xrightarrow{N \to \infty} \sum_{l=1}^{m_j} \frac{1}{m_j^2} [I^{vl}(\theta^{vl,*})]^{-1}_{j,j} + \sum_{l_1=1}^{m_j} \sum_{l_2=l_1+1}^{m_j} \frac{2}{m_j^2} [G_{v_{l_1},v_{l_2}}]_{j,j} \tag{4.3}$$

In the expression of the mean square error (4.3) 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 $\hat{\theta}^v_j$ 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 model Markov with respect to the four cycle G as described above. 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 2.

4.2 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 4.2 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 4.3, 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 (3.3) as

$$p(x_v|x_{N_v}) = \frac{\exp\{\sum_{j \in J^{v,PS}} \theta_j f_j(x_v, x_{N_v})\}}{1 + \sum_{y_v \in I_v \setminus \{0\}} \exp\{\sum_{j \in J^{v,PS}} \theta_j f_j(y_v, x_{N_v})\}} . \tag{4.4}$$

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 loglikelihood by the factor $\frac{1}{N}$. Then the v-local conditional loglikelihood function is

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

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

$$t_{J^{v,PS}} = [t_1, t_2, \cdots, t_{d_v}] \tag{4.5}$$

and

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

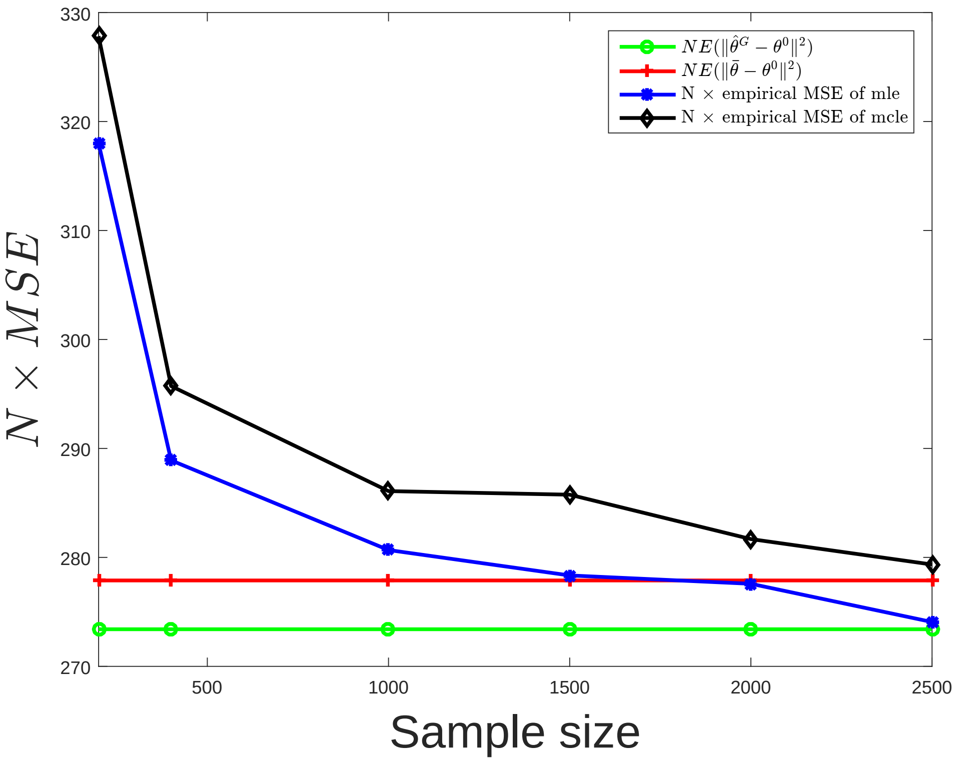

Figure 2: Empirical and theoretical mean square errors for the global mle and the mcle for the four-cycle graphical model.

where

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

Then the loglikelihood function is

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

Its first derivative is

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

with

$$\frac{\exp\{\sum_{j \in J^{v,PS}} \theta_j f_j(k_v, x^{(n)}_{N_v})\}}{Z^{n,v}(\theta)} = p(X_v = k_v | x^{(n)}_{N_v}) \tag{4.6}$$

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

$$z_{y_v}(\theta) = \sum_{j \in J^{v,PS}} \theta_j f_j(y_v, x^{(n)}_{N_v}). \tag{4.7}$$

For $k_v = l_v$, using (4.6), we obtain

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

if $k_v \neq l_v$, then

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

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

$$\eta^{n,v}_{k,l}(\theta, x^{(n)}_{N_v}) = \begin{cases} \frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)} - (\frac{\exp z_{k_v}(\theta)}{Z^{n,v}(\theta)})^2, & \text{if } k_v = l_v \\ -\frac{\exp z_{k_v}(\theta) \exp z_{l_v}(\theta)}{(Z^{n,v}(\theta))^2}, & \text{if } k_v \neq l_v \ . \end{cases} \tag{4.8}$$

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

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

where $\circ$ denotes the Hadamard product of two matrices. We make two assumptions on the behaviour of the cumulant 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

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

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

Theorem 4.2 *Assume conditions (A) and (B) hold. If the sample size N and $|V| = p$ satisfy*

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

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

$$\|\bar{\theta} - \theta^*\|_F \leq \frac{5C}{C_{min}}\sqrt{\frac{\sum_{v\in V} d_v \log p}{N}} \tag{4.9}$$

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

The proof is given in the Appendix. 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 4.3 *Assume conditions* (A') *and* (B') *hold. If* N *and* p *satisfy the condition*

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

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*

$$\|\hat{\theta}^G - \theta^*\|_F \leq \frac{5C}{\kappa^*}\sqrt{\frac{|J|\ \log p}{N}} \tag{4.10}$$

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

The proof is provided in the Supplementary file. Comparing Theorems 4.2 and 4.3, 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 $\frac{N}{\log p} = \mathcal{O}(\max_{v\in V}(d_v^2))$, $\|\bar{\theta} - \theta^*\|_F = \mathcal{O}(\sqrt{\frac{\sum_{v\in V} d_v \log p}{N}})$. This implies that for the mcle, the requirement on the sample size N are 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 $\frac{\sum_{v\in V} d_v}{|J|} = 1 + \frac{|E|}{p+|E|} \leq 2$. Of course, the size of the v-local model can grow with p (see Ravikumar et al., 2011) but like in Meng et al. (2014), since we are concerned with parameter estimation, we assume that the graph structure is known, that is $\sum_{v\in V} d_v$ and $|J|$ are known.

5 Conclusion

In this paper, we have made a detailed study of the maximum composite likelihood estimate of the parameter in a discrete graphical model, obtained through simple averaging of the mle of the parameters of local likelihoods. A basic result is that the components of $\theta^{v,PS} = (\theta_j \mid j \in J : \; v \in S(j)), \; v \in V$ are parameters of the global model, more precisely $(\theta^{v,PS}, v \in V) = \theta = (\theta_j, j \in J)$, and also that $\theta^{v,PS}$ is a subvector of the parameter vector of the local conditional likelihood as well as of the local marginal likelihood: see Remarks 3.1 and 3.2. Therefore combining the estimates of $\theta^{v,PS}$ obtained from the local models yields an estimate of the global parameter θ.

We then first show in Theorem 3.1 that whether, we deal with local conditional or marginal likelihoods, the local estimates of $\theta^{v,PS}$ are identical. It thus follows that we should use only local conditional estimates given their much simpler computational complexity. We call this estimate obtained through local conditional likelihoods the mcle. Second, we study the asymptotic properties of the mcle. Our result, Theorem 4.2, under the double asymptotic regime, p and N going to infinity, is new. It is stated under conditions similar to those imposed by Ravikumar et al. (2010) for local model selection through local conditional likelihoods. It indicates that for $\frac{N}{\log p}$ large enough, the mcle is close to the true value of the parameter with probability tending to 1. This behaviour compares well to the asymptotic behaviour of the global mle (see Theorem 4.3).

References

Besag, J., (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. Roy. Statist. Soc., Ser. B*, **36**, 192-236.

Fienberg, S. E. and Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist.*, **40**, 9961023.

Geyer, C.J. (1991). Markov chain Monte-Carlo maximum likelihood, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 156-163.

Hoeffding, W.(1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13-30.

Jordan, M. I. , Ghahramani, Z., Jaakkola, T. S. & Saul, L. K.(1999). An introduction to variational methods for graphical models. *Machine Learning*, **37**,183-233.

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

Letac, G. and Massam, H., (2012), Bayes regularization and the geometry of discrete hierarchical loglinear models, *Ann. Statist.*, **40**, 861-890.

Lindsay, B. G. (1988). Composite likelihood methods, *Contemp. Math.*, **80**, 221-239.

Liu, Q. and Ihler, A., (2012), Distributed parameter estimation via pseudo-likelihood, *International Conference on Machine Learning, (ICML).*

Meng, Z., Wei, D. Wiesel, A. and Hero, A.O. III, (2013), Distributed learning of Gaussian graphical models via marginal likelihood, *J. Mach. Learn. Res. W & CP*, **31**, 39-47.

Meng, D. Wei, A. Wiesel, A. Hero, (2014) "Distributed Learning of Gaussian Graphical Models via Marginal Likelihoods," *IEEE Trans. Signal Process.*, **62**, 5425-5438.

Mizrahi, Y.D., Denil, M. and de Freitas, N., (2014), Distributed Parameter Estimation in Probabilistic Graphical Models, *Advances in Neural Information Processing Systems (NIPS).*

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional graphical Ising model selection using l_1-regularized logistic regression, *Ann. Statist.*, **38**, 1287-1319.

Ravikumar, P., Wainwright, M. J., Rascutti, G. and Yu, B. (2011), High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence, *Electron. J. Statist.*, **5**, pp. 935-980.

Wang, N., Rauh, J. and Massam, H., (2016), Approximating faces of marginal polytopes in discrete hierarchical models, Arxiv 1603-04843v1.

Wainwright, M. and Jordan, M.I., (2008). Graphical Models, Exponential Families, and Variational Inference, *Found. Trends Mach. Learn.*, **1**, 1-305.

Wiesel, A. and Hero, A.O. III, (2012), Distributive covariance estimation in Gaussian graphical models, *IEEE Trans. signal process.*, **60**, 211-220.

6 Appendix

6.1 Proof of Lemma 3.1

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 (2.2), for $j \in J, S(j) \subset \mathcal{M}_v$, since $S(j)$ is complete,

$$\theta_j^{\mathcal{M}_v} = \sum_{j' \in J,\, j' \lhd j} (-1)^{|S(j)|-|S(j')|} \log \frac{p^{\mathcal{M}_v}(j')}{p^{\mathcal{M}_v}(0)}, \tag{6.1}$$

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,

$$\begin{aligned} p^{\mathcal{M}_v}(j) &= \sum_{i \in \mathcal{I}:\, i_{\mathcal{M}_v}=j} p(i) = \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v}=j} \exp\{ \sum_{j' \mid j' \lhd_0 j} \theta_{j'} + \sum_{\substack{j' \mid j' \lhd i \\ j' \ntriangleleft j \\ j'_{\mathcal{M}_v} \lhd_0 j}} \theta_{j'} \} \\ &= \Big(\exp \sum_{j' \mid j' \lhd_0 j} \theta_{j'} \Big) \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v}=j} \exp \sum_{\substack{j' \mid j' \lhd i \\ j' \ntriangleleft j \\ j'_{\mathcal{M}_v} \lhd_0 j}} \theta_{j'} \Big) . \end{aligned}$$

Therefore $\log p^{\mathcal{M}_v}(j) = \sum_{j' \mid j' \lhd_0 j} \theta_{j'} + \log \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v}=j} \exp \sum_{\substack{j' \mid j' \lhd i \\ j' \ntriangleleft j}} \theta_{j'} \Big)$, which we can write

$$\sum_{j' \mid j' \lhd_0 j} \theta_{j'} = \log p^{\mathcal{M}_v}(j) - \log \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v}=j} \exp \sum_{\substack{k \mid k \lhd i \\ k \ntriangleleft j}} \theta_k \Big) . \tag{6.2}$$

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 $\lhd$ on J, we derive from (6.2) that for $j \in J^{\mathcal{M}_v} \subset J$

$$\begin{aligned} \theta_j &= \sum_{j' \mid j' \lhd_0 j} (-1)^{|S(j)-S(j')|} \log p^{\mathcal{M}_v}(j') \\ &\quad - \sum_{j' \mid j' \lhd_0 j} (-1)^{|S(j)-S(j')|} \log \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v}=j'} \exp \sum_{\substack{k \mid k \lhd i \\ k \ntriangleleft j'}} \theta_k \Big) \\ &= \theta_j^{\mathcal{M}_v} - \sum_{j' \mid j' \lhd_0 j} (-1)^{|S(j)-S(j')|} \log \Big(1 + \sum_{i \in \mathcal{I},\, i_{\mathcal{M}_v}=j'} \exp \sum_{\substack{k \mid k \lhd i \\ k \ntriangleleft j'}} \theta_k \Big) \end{aligned} \tag{6.3}$$

which we prefer to write as (3.9).

6.2 Proof of Lemma 3.2

Since (3.9) is already proved, *(2.)* holds. Let us prove that *(1.)* holds, i.e., that when $S(j) \not\subset \mathcal{B}_v$, the alternating sum on the right-hand side of (3.9) 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 $\{\theta_k, k \triangleleft i^{(1)}, k \ntriangleleft j'\}$ in l_0 for $i^{(1)}$ such that $i^{(1)}_{\mathcal{M}_v} = 0$ is the same as the set $\{\theta_k, k \triangleleft i^{(2)}, k \ntriangleleft j'\}$ 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 (3.9) 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 $\theta_k, k \triangleleft i^{(1)}, k \ntriangleleft j'$ in $l_{j'}$ and the set $\theta_k, k \triangleleft i^{(2)}, k \ntriangleleft j'_w$ 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 (2.2), we have, for $S(i) = E \subset \mathcal{M}_v$

$$\begin{aligned}
\theta_i^{\mathcal{M}_v} &= \sum_{F\subset E} (-1)^{|E\setminus F|} \log p^{\mathcal{M}_v}(i_F, 0_{\mathcal{M}_v\setminus F}) \\
&= \sum_{F\subset E} (-1)^{|E\setminus F|} \log \Big(p(i_F, 0_{V\setminus F}) + \sum_{L\subset V\setminus\mathcal{M}_v} \sum_{k_L\in I_L} p(i_F, 0_{\mathcal{M}_v\setminus F}, k_L, 0_{V\setminus(\mathcal{M}_v\cup L)}) \Big) \\
&= \sum_{F\subset E} (-1)^{|E\setminus F|} \log \Big(\exp(\sum_{j\in J, j\triangleleft i_F} \theta_j) + \sum_{L\subset V\setminus F} \sum_{k_L\in I_L} \exp(\sum_{j\in J, j\triangleleft i_F} \theta_j + \sum_{j\ntriangleleft i_F, j\triangleleft(i_F,k_L)} \theta_j) \Big) \\
&= \sum_{F\subset E} (-1)^{|E\setminus F|} \log \Big(\exp(\sum_{j\in J, j\triangleleft i_F} \theta_j) \Big) \qquad (6.4) \\
&\quad + \sum_{F\subset E} (-1)^{|E\setminus F|} \log (1 + \sum_{L\subset V\setminus F} \sum_{k_L\in I_L} \exp(\sum_{j\ntriangleleft i_F, j\triangleleft(i_F,k_L)} \theta_j) \Big) \\
&= \theta_i + \sum_{F\subset E} (-1)^{|E\setminus F|} \log (1 + \sum_{L\subset V\setminus F} \sum_{k_L\in I_L} \exp(\sum_{j\ntriangleleft i_F, j\triangleleft(i_F,k_L)} \theta_j) \Big) \qquad (6.5)
\end{aligned}$$

Now, following an argument similar to that of *(1.)* above, we can show that the second component of the sum in (6.5) is equal to zero. It follows that when $\theta_i = 0$ then $\theta_i^{\mathcal{M}_v} = 0$. This completes the proof of Lemma 3.2.

6.3 Proof of Theorem 3.1

The local relaxed marginal loglikelihood is

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

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

$$t(j) - p^{\mathcal{M}_{l,v}}(j_{S(j)}) = 0, \tag{6.6}$$

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

For the argument to follow is essentially the same for the one-hop or two-hop neighbourhood. We present it for the more general case of the two hop neighbourhood. The local conditional log likelihood is

$$\begin{aligned} l^{v,2PS}(\theta^{v,2PS}) &= \sum_{i_{\mathcal{M}_v}\in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v}) \log \frac{p(X_v = i_v, X_{\mathcal{N}_v} = i_{\mathcal{N}_v}, X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})}{p(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})} \\ &= \sum_{i_{\mathcal{M}_v}\in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v}) \log \frac{p^{\mathcal{M}^{2,v}}(X_{\mathcal{M}_v} = i_{\mathcal{M}_v})}{p^{\mathcal{M}^{2,v}}(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}})} \\ &= \sum_{i_{\mathcal{M}_v}\in I_{\mathcal{M}_v}} n(i_{\mathcal{M}_v}) \log p^{\mathcal{M}^{2,v}}(X_{\mathcal{M}_v} = i_{\mathcal{N}_v}) - \sum_{i_{\mathcal{N}_{2v}}\in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \log p^{\mathcal{M}^{2,v}}(X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}) \\ &= l^{\mathcal{M}_{2,v}}(\theta^{\mathcal{M}_{2,v}}) - \sum_{i_{\mathcal{N}_{2v}}\in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \log \sum_{x_{v\cup\mathcal{N}_v}\in I_{v\cup\mathcal{N}_v}} p^{\mathcal{M}^{2,v}}(X_{v\cup\mathcal{N}_v} = x_{v\cup\mathcal{N}_v}, X_{\mathcal{N}_{2v}} = i_{\mathcal{N}_{2v}}) \\ &= l^{\mathcal{M}_{2,v}}(\theta^{\mathcal{M}_{2,v}}) - Q \end{aligned} \tag{6.7}$$

where

$$Q = \sum_{i_{\mathcal{N}_{2v}}\in I_{\mathcal{N}_{2v}}} n(i_{\mathcal{N}_{2v}}) \log \sum_{x_{v\cup\mathcal{N}_v}\in I_{v\cup\mathcal{N}_v}} \exp\Big(\theta_0 + \sum_{\substack{k\lhd(x_{v\cup\mathcal{N}_v}, i_{\mathcal{N}_{2v}})\\ k\in J^{\mathcal{M}_{2,v}}}} \theta_k\Big) \tag{6.8}$$

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 (3.3) 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 at 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 3.2, that remains unchanged when we change from the global model to the $\mathcal{M}_{2,v}$-marginal models.

We will now differentiate the expression of $l^{v,2PS}$ in (6.8) 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

$$\mathbf{1}_{j \triangleleft (x_{v \cup \mathcal{N}_v}, i_{\mathcal{N}_{2v}})} = \left\{ \begin{array}{ll} 1 & \text{if } j \triangleleft (x_{v \cup \mathcal{N}_v}, i_{\mathcal{N}_{2v}}) \\ 0 & \qquad \text{otherwise} \end{array} \right. \quad ,$$

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

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

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

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

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

$$\frac{\partial Q}{\partial \theta_j} = 0, \ \ j \in J^{\mathcal{M}_{2,v}}, \ S(j) \subset \mathcal{N}_{2v}. \tag{6.9}$$

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

$$\begin{aligned}\frac{\partial Q}{\partial \theta_j} &= -Np^{\mathcal{M}^{2,v}}(j_{S(j)}) + N \sum_{i_{\mathcal{N}_{2v}} \in I_{\mathcal{N}_{2v}}} p^{\mathcal{M}^{2,v}}(j_{S(j)\cap(v\cup\mathcal{N}_v)}, i_{\mathcal{N}_{2v}})\mathbf{1}_{j_{\mathcal{N}_{2v}} \lhd i_{\mathcal{N}_{2v}}} \\ &= -Np^{\mathcal{M}^{2,v}}(j_{S(j)}) + Np^{\mathcal{M}^{2,v}}(j_{S(j)}) = 0 \qquad (6.10)\end{aligned}$$

It follows from (6.9) and (6.10) 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

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

6.4 Proof of Theorem 4.1

Given the definition of $\bar{\theta}$, to show (4.1), 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 (4.2). It is immediate to show that G is finite. By the central limit theorem we then 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 (4.3) is also an immediate consequence of this asymptotic distribution.

6.5 Proof of Theorem 4.2

To prove Theorem 4.2, we need two preliminary results.

Lemma 6.1 *Let $\theta^{v,*} = (\theta^*)^{v,PS}$ be the true value of the parameter for the conditional model of X_v given X_{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 (4.5), if there exists $\epsilon > 0$ such that*

$$\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^{v,*})\|_\infty \leq \epsilon \leq \frac{C_{min}^2}{10 D_{max} d_v} \tag{6.11}$$

then

$$\|\hat{\theta}^{v,PS} - \theta^{v,*}\|_F \leq \frac{5\sqrt{d_v}\epsilon}{C_{min}} \tag{6.12}$$

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 $C(\delta) = \{\Delta \mid \quad \|\Delta\|_F = \delta\}$. Since $Q(\Delta)$ is a convex function of Δ, if we can prove

$$\inf_{\Delta \in C(\delta)} Q(\Delta) > 0, \tag{6.13}$$

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

$$
\begin{aligned}
Q(\Delta) &= l(\theta^*) - l(\theta^* + \Delta) = \theta^{*t} t - k(\theta^*) - ((\theta^* + \Delta)^t t - k(\theta^* + \Delta)) \\
&= k(\theta^* + \Delta) - k(\theta^*) - \Delta^t t = \Delta^t k^{'}(\theta^*) + \tfrac{1}{2}\Delta^t k^{''}(\theta^* + \alpha\Delta)\Delta - \Delta^t t, \quad \alpha \in [0,1] \\
&= \underbrace{\Delta^t [k^{'}(\theta^*) - t]}_{Q_1} + \underbrace{\frac{1}{2}\Delta^t k^{''}(\theta^* + \alpha\Delta)\Delta}_{Q_2}
\end{aligned}
\tag{6.14}
$$

By Hölder's and Cauchy' s inequality, we have the following bound for Q_1.

$$
|Q_1| = |\Delta^t [k^{'}(\theta^*) - t]| \;\; \leq \;\; \|k^{'}(\theta^*) - t\|_\infty ||\Delta||_1 \leq \epsilon\sqrt{d}\|\Delta\|_F = \epsilon\sqrt{d}\delta \tag{6.15}
$$

For Q_2, we have

$$
Q_2 \geq \frac{1}{2}\|\Delta\|_F^2 \min_{\alpha\in[0,1]} \lambda_{min} k^{''}(\theta^* + \alpha\Delta) = \frac{1}{2}\delta^2 \min_{\alpha\in[0,1]} \lambda_{min} k^{''}(\theta^* + \alpha\Delta) \tag{6.16}
$$

We now want to bound the term $q = \min_{\alpha\in[0,1]} \lambda_{min}[k^{''}(\theta^* + \alpha\Delta)]$ from below. Following (4.7), we can write $z_{y_v}(\theta + \alpha\Delta) = \sum_{j\in J; v\in S(j)} (\theta_j + \alpha\Delta_j) f_j(y_v, x_{N_v}^{(n)})$, then we can rewrite the entries of H in (4.8) as

$$
\eta_{k,l}^{n,v}(\theta^* + \alpha\Delta, x_{N_v}^{(n)}) = \begin{cases} \frac{\exp z_{k_v}(\theta^*+\alpha\Delta)}{1+\sum_{y_v\in I_v\setminus\{0\}} \exp z_{y_v}(\theta^*+\alpha\Delta)} - (\frac{\exp z_{k_v}(\theta^*+\alpha\Delta)}{1+\sum_{y_v\in I_v\setminus\{0\}} \exp z_{k_v}(\theta^*+\alpha\Delta)})^2, & \text{if } k_v = l_v \\ -\frac{\exp z_{k_v}(\theta^*+\alpha\Delta) \exp z_{l_v}(\theta^*+\alpha\Delta)}{(1+\sum_{y_v\in I_v\setminus\{0\}} \exp z_{y_v}(\theta^*+\alpha\Delta))^2}, & \text{if } k_v \neq l_v \end{cases}
$$

then

$$
\frac{\partial \eta_{k,l}^{n,v}(\theta^* + \alpha\Delta, x_{N_v}^{(n)})}{\partial\alpha} = \sum_{y_v\in I_v\setminus\{0\}} (\eta_{k,l}^{n,v})'_{y_v}(\theta^* + \alpha\Delta, x_{N_v}^{(n)}) \frac{\partial z_{y_v}}{\partial\alpha},
$$

where $(\eta_{k,l}^{n,v})'_{y_v}(\theta^* + \alpha\Delta, x_{N_v}^{(n)}) = \frac{\partial\eta_{k,l}^{n,v}(\theta^*+\alpha\Delta, x_{N_v}^{(n)})}{\partial z_{y_v}}$. It is easy to see that these derivatives can all be expressed in terms of probabilities of the type (4.6) and that they are always less than 1 in absolute value. Therefore, since $\frac{\partial z_{y_v}(\theta+\alpha\Delta)}{\partial\alpha} = \sum_{j\in J; v\in S(j)} \Delta_j f_j(y_v, x_{N_v}^n)$, we have

$$
\begin{aligned}
|\tfrac{\partial\eta_{k,l}^{n,v}(\theta^*+\alpha\Delta, x_{N_v}^{(n)})}{\partial\alpha}| \;\; &\leq \;\; \textstyle\sum_{y_v\in I_v\setminus\{0\}} \frac{\partial z_{y_v}}{\partial\alpha} = \sum_{y_v\in I_v\setminus\{0\}} \sum_{j\in J; v\in S(j)} \Delta_j f_j(y_v, x_{N_v}^n) \\
&= \;\; \textstyle\sum_{j\in J; v\in S(j)} \Delta_j \sum_{y_v\in I_v\setminus\{0\}} f_j(y_v, x_{N_v}^n) = \langle\Delta, W^n\rangle \;,
\end{aligned}
\tag{6.17}
$$

since for each $j \in J^{v,PS}$, $\sum_{y_v\in I_v\setminus\{0\}} f_j(y_v, x_{N_v}^n) = f_j(j_v, x_{N_v}^n) = W_j^n$.

The Taylor series expansion of $\eta^{n,v}_{k,l}(\theta^* + \alpha\Delta, x^{(n)}_{N_v})$ yields

$$\eta^{n,v}_{k,l}(\theta^* + \alpha\Delta, x^{(n)}_{N_v}) = \eta^{n,v}_{k,l}(\theta^*, x^{(n)}_{N_v}) + \alpha\frac{\partial \eta^{n,v}_{k,l}(\theta^* + \alpha'\Delta, x^{(n)}_{N_v})}{\partial\alpha}, \ \alpha^{'} \in [0, \alpha] \ .$$

Let $K(\theta^* + \alpha^{'}\Delta, x^{(n)}_{N_v})$ denote the $d_v \times d_v$ matrix with entry $\frac{\partial\eta^{n,v}_{k,l}(\theta^*+\alpha\Delta, x^{(n)}_{N_v})}{\partial\alpha}$. Coming back to (6.16), we have

$$\begin{aligned} k^{''}(\theta^* + \alpha\Delta) &= \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\left[H(\theta^* + \alpha\Delta, x^{(n)}_{N_v}) \circ [W^n(W^n)^t]\right] \\ &= \tfrac{1}{N}\textstyle\sum_{n=1}^{N} H(\theta^*, x^{(n)}_{N_v}) \circ [W^n(W^n)^t] + \alpha\tfrac{1}{N}\textstyle\sum_{n=1}^{N} K(\theta^* + \alpha^{'}\Delta, x^{(n)}_{N_v}) \circ [W^n(W^n)^t] \ . \end{aligned}$$

We write $||X||_2 = \lambda_{max}(X)$ for the operator norm of a matrix X. By Lemma 7.4 of the Supplementary file,

$$\lambda_{min}\Big(k^{''}(\theta^*+\alpha\Delta)\Big) \geq \lambda_{min}\Big(\frac{1}{N}\sum_{n=1}^{N} H(\theta^*, x^{(n)}_{N_v})\circ[W^n(W^n)^t]\Big) - \|\alpha\frac{1}{N}\sum_{n=1}^{N} K(\theta^*+\alpha^{'}\Delta, x^{(n)}_{N_v})\circ[W^n(W^n)^t]\|_2$$

Then since $|\alpha| < 1$, we have

$$\begin{aligned} q &= \min\nolimits_{\alpha\in[0,1]} \lambda_{min}[\tfrac{1}{N}\textstyle\sum_{n=1}^{N} H(\theta^* + \alpha\Delta, x^{(n)}_{N_v}) W^n(W^n)^t] \\ &\geq \lambda_{min}(\tfrac{1}{N}\textstyle\sum_{n=1}^{N}\left[H(\theta^*, x^{(n)}_{N_v}) \circ (W^n(W^n)^t)\right]) \\ &\qquad\qquad - \max\nolimits_{\alpha\in[0,1]} \|\alpha\tfrac{1}{N}[\textstyle\sum_{n=1}^{N} K(\theta^* + \alpha\Delta, x^{(n)}_{N_v}) \circ (W^n(W^n)^t)]\|_2 \\ &\geq C_{min} - \max\nolimits_{\alpha\in[0,1]} \underbrace{\|\frac{1}{N}\sum_{n=1}^{N}\Delta^t W^n(W^n(W^n)^t)\|_2}_{A} \\ &= C_{min} - \max\nolimits_{\alpha\in[0,1]} ||A||_2 \ , \end{aligned} \tag{6.18}$$

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

$$\begin{aligned} \langle y, Ay\rangle &= \frac{1}{N}\sum_{n=1}^{N}(\Delta^t W^n)(y^t W^n)^2 \leq \frac{1}{N}\sum_{n=1}^{N}|\Delta^t W^n|(y^t W^n)^2, \\ |\Delta^t W^n| &\leqslant \sqrt{d}||\Delta||_F = \sqrt{d}\delta \ . \end{aligned} \tag{6.19}$$

and, by definition of the operator norm and from Assumption (B),

$$\frac{1}{N}\sum_{n=1}^{N}(y^t W^n)^2 \leq ||\frac{1}{N}\sum_{n=1}^{N} W^n(W^n)^t||_2 < D_{max} \ . \tag{6.20}$$

From (6.18), (6.19) and (6.20), 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 (6.16), we get

$$Q_2 \geq \frac{1}{2}\delta^2(C_{min} - D_{max}\sqrt{d}\delta). \tag{6.21}$$

From the two inequalities (6.15) and (6.21), it follows that

$$Q(\Delta) \geq Q_2 - |Q_1| \geq \frac{1}{2}\delta^2(C_{min} - D_{max}\sqrt{d}\delta) - \epsilon\sqrt{d}\delta. \tag{6.22}$$

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 (6.22) 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 (6.13) holds for $\delta = \frac{5\sqrt{d}\epsilon}{C_{min}}$ and the theorem is proved. □

In the next lemma, we will make use of Hoeffding inequality (see Hoeffding (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 $\epsilon > 0$

$$p(|\bar{X} - \mu| \geq \epsilon) \leq 2\exp\Big(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big) \ . \tag{6.23}$$

Lemma 6.2 *Let $t_{J^{v,PS}}, k^{v,PS}$ and d_v be as defined above. For any $\epsilon > 0$, we have*

$$p(\{\max_{v\in V} \|t_{J^{v,PS}} - (k^{v,PS})'(\theta^{v,*})\|_\infty \geq \epsilon\}) \leq 2|J|\exp(-2N\epsilon^2) \ . \tag{6.24}$$

Proof. For $j \in J^{v,PS}$, we clearly have

$$E_{\theta^*}\Big(\frac{\partial l(\theta)}{\partial \theta_j}\Big) = E_{\theta^*}\Big(t_j - \frac{\partial k(\theta)}{\partial \theta_j}\Big) = E_{\theta^*}\Big(\frac{1}{N}\sum_{n=1}^{N} f_j(x_v^{(n)}, x_{N_v}^{(n)}) - p(x_v = j_v | x_{N_v}^{n}) f_j(x_v = j_v, x_{N_v}^{(n)})\Big) = 0$$

We note that since $x_{N_v}^{(n)}$ is given and $f_j(x_v^{(n)}, x_{N_v}^{(n)})$ takes values 0 or 1, we have $E(f_j(x_v^{(n)}, x_{N_v}^{(n)})) = p(x_v = j_v | x_{N_v}^{n}) f_j(x_v = j_v, x_{N_v}^{(n)})$ and by Hoeffding's inequality (6.23), 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

$$\begin{aligned} P(\max_{v\in V}\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\|_\infty \leq \epsilon) &= 1 - P(\cup_{j\in\cup J^{v,PS}}\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\| \geq \epsilon) \\ &\geq 1 - \sum_{j\in\cup J^{v,PS}} P(\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\| \geq \epsilon), \\ &\geq 1 - 2|J|\exp(-2N\epsilon^2) \end{aligned}$$

which proves the lemma. □

Proof of Theorem 4.2 Let $\epsilon = C\sqrt{\frac{\log p}{N}}$, where C is a constant that we will choose later in this proof. From Lemma 6.2, we have

$$p(\max_{v\in V}\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\|_\infty \geq C\sqrt{\frac{\log p}{N}}) \leq 2|J|\exp(-2C^2\log p) = \frac{2|J|}{p^{2C^2}} \qquad (6.25)$$

From Lemma 6.1, for $\epsilon = C\sqrt{\frac{\log p}{N}} \leq \frac{C_{min}^2}{10D_{max}d_v}$, i.e. for $N \geq (\frac{10CD_{max}d_v}{C_{min}^2})^2\log p$, we have

$$\|t_{J^{v,PS}} - (k^{v,PS})'(\theta^*)\|_\infty \leq \epsilon \leq \frac{C_{min}^2}{10D_{max}d_v} \Rightarrow \|\hat{\theta}^{v,PS} - \theta^{v,*}\|_F \leq \frac{5\sqrt{d_v}\epsilon}{C_{min}}.$$

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:

$$\begin{aligned} \|\bar{\theta} - \theta^*\|_F &\leq \Big(\textstyle\sum_{v\in V}\|\hat{\theta}^{v,PS} - \theta^{v,*}\|_F^2\Big)^{\frac{1}{2}} \\ &\leq (\textstyle\sum_{v\in V}(\frac{5\sqrt{d_v}C\sqrt{\frac{\log p}{N}}}{C_{min}})^2)^{\frac{1}{2}} = \frac{5C}{C_{min}}\sqrt{\frac{\sum_{v\in V} d_v\log p}{N}} \end{aligned}$$

Therefore under the condition $N \geq \max_{v \in V} \, (\frac{10CD_{max}d_v}{C_{min}^2})^2 \log p$, we have

$$p(\|\bar{\theta}-\theta^*\|_F \leq \frac{5C}{C_{min}}\sqrt{\frac{\sum_{v\in V} d_v \log p}{N}}) \geq p(\max_{v\in V}\|t_{J^{v,PS}}-k'^{v,PS}(\theta^*)\|_\infty \leq C\sqrt{\frac{\log p}{N}}) \geq 1-\frac{2|J|}{p^{2C^2}}$$

with the last inequality due to (6.25).

The theorem would make no sense if probability of the convergence rate was negative and thus C must satisfy

$$1-\frac{2|J|}{p^{2C^2}} > 0 \Rightarrow C \geq \sqrt{\frac{\log(2|J|)}{2\log p}}\ .$$

□