Received: 12 May 2022 | Revised: 9 March 2023 | Accepted: 1 June 2023

DOI: 10.1111/sjos.12674

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

# Structure recovery for partially observed discrete Markov random fields on graphs under not necessarily positive distributions

**Florencia Leonardi**[1] | **Rodrigo Carvalho**[1] | **Iara Frondana**[2]

[1]Statistics Department, Universidade de São Paulo, São Paulo, Brazil

[2]Pavia, Italy

**Correspondence**
Florencia Leonardi, Statistics Department, Universidade de São Paulo, São Paulo, Brazil.
Email: florencia@usp.br

**Funding information**
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Numbers: 311763/2020-0, 432310/2018-5; Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Numbers: 2013/07699-0, 2019/17734-3

## Abstract

We propose a penalized conditional likelihood criterion to estimate the basic neighborhood of each node in a discrete Markov random field that can be partially observed. We prove the convergence of the estimator in the case of a finite or countable infinite set of nodes. The estimated neighborhoods can be combined to estimate the underlying graph. In the finite case, the graph can be recovered with probability one. In contrast, we can recover any finite subgraph with probability one in the countable infinite case by allowing the candidate neighborhoods to grow as a function $o(\log n)$, with $n$ the sample size. Our method requires minimal assumptions on the probability distribution, and contrary to other approaches in the literature, the usual positivity condition is not needed. We evaluate the estimator's performance on simulated data and apply the methodology to a real dataset of stock index markets in different countries.

**KEYWORDS**

conditional likelihood, graphical model, model selection, structure estimation

# 1 | INTRODUCTION

Discrete Markov random fields on graphs, usually called graphical models in the statistical literature, have received much attention from researchers in recent years, mainly due to their flexibility to capture conditional dependence relations between variables (Divino et al., 2000; Koller & Friedman, 2009; Lauritzen, 1996; Lerasle & Takahashi, 2016; Pensar et al., 2017). They have been applied to many different problems in different fields such as Biology (Shojaie & Michailidis, 2010), Social Sciences (Strauss & Ikeda, 1990), or Neuroscience (Duarte et al., 2019). Graphical models are in some sense "finite" versions of general random fields or Gibbs distributions, classical models in stochastic processes and statistical mechanics theory (Georgii, 2011).

In this work, we focus on discrete Markov random field models (with a finite or countable infinite set of variables), where the set of random variables takes values on a finite alphabet. One of the main statistical questions for this model is how to recover the underlying graph; that is, the graph determined by the conditional dependence relationships between the variables. For the class of Markov random fields on lattices, some methods based on penalized pseudo-likelihood criteria like the Bayesian Information Criterion (BIC) of Schwarz (1978) have appeared in the literature (Csiszár & Talata, 2006; Ji & Seymour, 1996); see also Tjelmeland and Besag (1998) and Löcherbach and Orlandi (2011). In the case of Markov random fields defined on general graphs, the most studied model is the binary graphical model with pairwise interactions where structure estimation can be addressed by using standard logistic regression techniques (Ravikumar et al., 2010; Strauss & Ikeda, 1990), distance-based approaches between conditional probabilities (Bresler et al., 2018; Galves et al., 2015) and maximization of the $\ell_1$-penalized pseudo-likelihood (Atchade, 2014; Höfling & Tibshirani, 2009); see also Santhanam and Wainwright (2012). In the case of bigger discrete alphabets or general types of interactions, to our knowledge, the only work addressing the structure estimation problem is Loh and Wainwright (2013), where the authors obtain a characterization of the edges in the graph with the zeros in a generalized inverse covariance matrix. Then, this characterization is used to derive estimators for restricted classes of models, and the authors prove the consistency in probability of these estimators.

Markov random fields have also been proposed for continuous random variables, where the structure estimation problem has been addressed by $\ell_1$-regularization for Gaussian Markov random fields (Meinshausen & Bühlmann, 2006) and also extended to nonparametric models (Lafferty et al., 2012; Liu et al., 2012) and general conditional distributions from the exponential family Yang et al. (2015).

All these works, for discrete or continuous random variables, assume the model satisfies a usual "positivity" condition that states that the probability distributions of finite subsets of variables are strictly positive. The positivity condition guarantees a factorization property of the joint distribution, thanks to a classical result known as Hammersley–Clifford theorem Hammersley and Clifford (1971). But the positivity condition is strong for discrete distributions, where in many applications, some configurations are impossible to occur and then have zero probability. When this occurs, the joint distribution of the variables could be described by using fewer parameters than needed by the full model, a property usually referred to as *sparsity*. The sparsity property is especially appealing for high dimensional data, where the number of variables is high, and the number of relevant parameters in the model must be assumed relatively small to have efficient estimators. The notion of sparsity we described above does not coincide exactly with that usually assumed in the literature, namely that the graph of interaction is sparse, that is, has few edges. Still, the two go in the same direction of describing more parsimonious models. We observe that in our setting of not necessarily positive distributions, we can simultaneously have models satisfying

both types of sparsity, with the byproduct of having a conditional likelihood function with fewer factors.

This work addresses the structure estimation problem for discrete Markov random fields without assuming the positivity condition. We first introduce a penalized conditional likelihood criterion to estimate the neighborhoods of the nodes, which are later combined to obtain an estimator of the underlying graph. We prove that both estimators converge almost surely to the true underlying graph in the case of a finite graphical model when the sample size grows without imposing additional hypotheses on the model. In the countable infinite case, when the underlying graph is infinite, and the number of observed variables is allowed to grow with the sample size, we prove that the estimator restricted to a finite subgraph also converges almost surely to the corresponding subgraph. A preliminary version of these results can be found in Frondana (2016).

The paper is organized as follows. Section 2 presents the definition of the model, including some examples. Section 3 introduces the estimator for the neighborhood of a node and two different forms of combining the neighborhoods to estimate the dependence graph. In that section, we also state the main theoretical results of the paper. Finally, in Section 4, we evaluate the estimator's performance through simulations; Section 5 shows a real data application. The proofs of the theoretical results are included in the Appendix.

## 2 | DISCRETE MARKOV RANDOM FIELDS ON GRAPHS

A *graph* is a pair $G = (V, E)$, where $V$ is the set of vertices (or nodes) and $E$ is the set of edges, $E \subset V \times V$. A graph $G$ is said *simple* if for all $i \in V$, $(i, i) \notin E$ and it is said *undirected* if $(i, j) \in E$ implies $(j, i) \in E$ for every pair $(i, j) \in V \times V$. Given any set $S$, the symbol $|S|$ denotes its cardinality.

Let $A$ be a finite set. For $\Delta \subseteq V$, a subset of vertices, the set $a_\Delta = \{a_i \in A : i \in \Delta\}$ denotes a configuration on $\Delta$. $A^\Delta$ denotes the set of all configurations on $\Delta$.

A *random field* on $A^V$ is a family of random variables indexed by the elements of $V$, $\{X_v : v \in V\}$, where each $X_v$ is a random variable with values in $A$. For $\Delta \subseteq V$ we write $X_\Delta = \{X_i : i \in \Delta\}$. The law of the random variables $X_V$ is denoted by $\mathbb{P}$.

For any finite $\Delta \subset V$ we write

$$p(a_\Delta) = \mathbb{P}(X_\Delta = a_\Delta) \text{ with } a_\Delta \in A^\Delta \tag{1}$$

and if $p(a_\Delta) > 0$ we denote by

$$p(a_\Phi | a_\Delta) = \mathbb{P}(X_\Phi = a_\Phi | X_\Delta = a_\Delta) \text{ for } a_\Phi \in A^\Phi,\ a_\Delta \in A^\Delta \tag{2}$$

the corresponding conditional probability distributions.

Given $v \in V$, a *neighborhood* $W$ of $v$ is any finite set of vertices with $v \notin W$. If there is a neighborhood $W$ of $v$ satisfying

$$p(a_v | a_W) = p(a_v | a_\Delta) \tag{3}$$

for all finite $\Delta \supset W$, $v \notin \Delta$ and all $a_v \in A$, $a_\Delta \in A^\Delta$ with $p(a_\Delta) > 0$, then $W$ is called *Markov neighborhood* of $v$. The definition of a Markov neighborhood $W$ is equivalent to request that for all $\Phi$ finite (not containing $v$) with $\Phi \cap W = \emptyset$, $X_\Phi$ is conditionally independent of $X_v$, given $X_W$.

Formally,

$$X_v \perp\!\!\!\perp X_\Phi | X_W \ , \ \text{for all} \ \ \Phi \ \ \text{with} \ \ \Phi \cap W = \emptyset, \tag{4}$$

where $\perp\!\!\!\perp$ is the usual symbol denoting the independence of random variables. This conditional independence assumption defining the Markov neighborhoods corresponds to the property known as *local Markov* in finite graphical models. This is a weaker condition than the usually assumed *global Markov* property, see Lauritzen (1996) for details.

An essential fact that we can derive from the definition is that if $W$ is a Markov neighborhood of $v \in V$, then any finite set $\Delta \supset W$ is also a Markov neighborhood of $v$. On the other hand, if $W_1$ and $W_2$ are Markov neighborhoods of $v$, then it is only sometimes generally valid that $W_1 \cap W_2$ is a Markov neighborhood, as shown in the following example.

> **Example 1.** Let $V = \{1, 2, 3\}$ and consider the vector $(X_1, X_2, X_3)$ of Bernoulli random variables with $\mathbb{P}(X_i = 0) = 1/2, \mathbb{P}(X_i = 1) = 1/2$, for $i = 1, 2, 3$. Suppose that $X_1 = X_2 = X_3$ with probability 1. Then it is easy to check that both $\{2\}$ and $\{3\}$ are Markov neighborhoods of node 1, but the intersection is not a Markov neighborhood (which will imply $X_1$ being independent of $X_2$ and $X_3$).

This property, satisfied by some probability measures, defines the following Markov intersection condition, formally stated here.

**Markov intersection property**: For all $v \in V$ and all $W_1$ and $W_2$ Markov neighborhoods of $v$, the set $W_1 \cap W_2$ is also a Markov neighborhood of $v$.

The Markov intersection condition is desirable in this context to define the smallest Markov neighborhood of a node and to enable the structure estimation problem to be well defined. This property is guaranteed under the following usual condition assumed in the literature:

**Positivity condition**: For all finite $W \subset V$ and all $a_W \in A^W$ we have $p(a_W) > 0$.

It can be seen that the positivity condition implies the Markov intersection property, see Lauritzen (1996) for details. For this reason, the literature on Markov random fields generally assumes that the positivity condition holds. But there are distributions satisfying the Markov intersection property that are not strictly positive. An example is a typical realization of a Markov chain with some zeros in the transition matrix. The following basic example shows that the positivity condition and Markov intersection property are generally not equivalent.

> **Example 2.** Let $V = \mathbb{Z}$ and take a stationary Markov chain of order one assuming values in $A = \{0, 1\}$ with transition matrix
>
> $$P = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}$$
>
> The distribution $\mathbb{P}$ of the Markov chain does not satisfy the positivity condition (any configuration with two subsequent ones has zero probability). But the distribution satisfies the Markov intersection property because any Markov neighborhood of node $i$ necessarily contains nodes $i - 1$ and $i + 1$, corresponding to the *minimal* Markov neighborhood of node $i$.

From now on, we assume the distribution $\mathbb{P}$ satisfies the Markov intersection property defined before. For $v \in V$, let $\Theta(v)$ be the set of all subsets of $V$ that are Markov neighborhoods of $v$. The

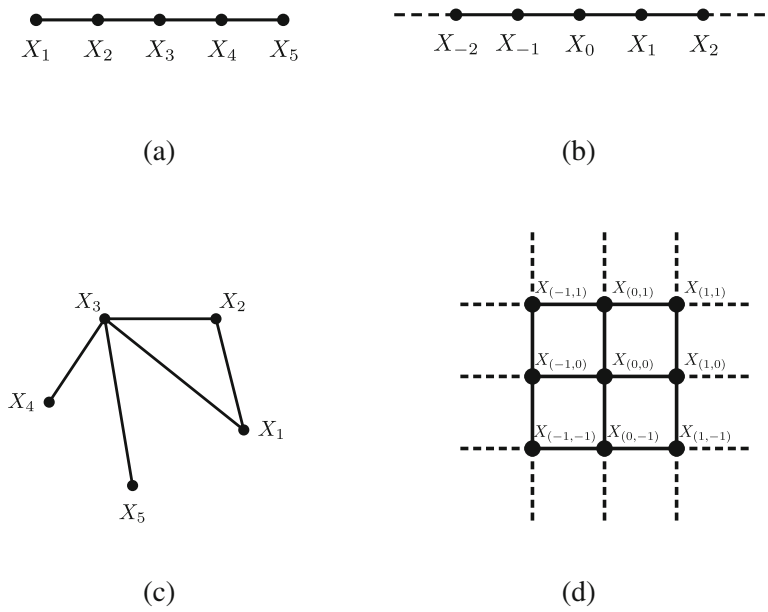



**FIGURE 1** Different graph structures of Markov random fields, with finite (left) and countable infinite (right) sets of variables. Examples (a) and (b) are obtained in particular by the distribution in Example 2. Case (a) is the projection on $\{0,1\}^{\{1,\dots,5\}}$ and case (b) is a representation of the joint distribution on $\{0,1\}^{\mathbb{Z}}$. (c) is a finite graphical model defined on a general graph that we will further use in the simulations in Section 4 see Example 3 below and (d) represents the interaction graph in a classical Ising model, see for example Csiszár and Talata (2006) and Georgii (2011).

*basic neighborhood* of $v$ is defined as

$$\mathrm{ne}(v) = \bigcap_{W \in \Theta(v)} W. \tag{5}$$

By the Markov intersection property, $\mathrm{ne}(v)$ is the smallest Markov neighborhood of $v \in V$. The set $\mathrm{ne}(v)$ is often called *Markov blanket* of node $v$. Based on these special neighborhoods, define the graph $G = (V, E)$ by

$$(v, w) \in E \text{ if and only if } w \in \mathrm{ne}(v). \tag{6}$$

We can state the following fundamental result.

**Lemma 1.** *The graph $G$, defined by (6), is undirected, that is, if $(v, w) \in E \Rightarrow (w, v) \in E$.*

The proof of Lemma 1 can be found in the Appendix.

As an illustration, we show in Figure 1 different Markov random field models with finite as well as infinite undirected graphs. Besides, the graphs with an infinite set of nodes in this example are regular in the sense that the neighborhoods have the same structure for each node; in our setting, we allow for different neighborhood structures for different nodes, not imposing a model

**TABLE 1** Conditional probabilities used to define a joint distribution on $\{0,1,2\}^5$ by the factorization $p(x_1,x_2,x_3,x_4,x_5) = p(x_3)p(x_2|x_1,x_3)p(x_1|x_3)p(x_4|x_3)p(x_5|x_3)$.

| $x_3$ | 0 | 1 | 2 |
|---|---|---|---|
| $p(x_3)$ | 0.3 | 0.2 | 0.5 |
| $x_2$ | 0 | 1 | 2 |
| $p(x_2\|x_1=0,x_3=0)$ | 0.5 | 0.5 | 0 |
| $p(x_2\|x_1=1,x_3=0)$ | 0.5 | 0.25 | 0.25 |
| $p(x_2\|x_1=2,x_3=0)$ | 0.25 | 0.25 | 0.5 |
| $p(x_2\|x_1=0,x_3=1)$ | 0.3 | 0 | 0.7 |
| $p(x_2\|x_1=1,x_3=1)$ | 0.25 | 0.25 | 0.5 |
| $p(x_2\|x_1=2,x_3=1)$ | 0.3 | 0.7 | 0 |
| $p(x_2\|x_1=0,x_3=2)$ | 0 | 0.75 | 0.25 |
| $p(x_2\|x_1=1,x_3=2)$ | 0.3 | 0.3 | 0.4 |
| $p(x_2\|x_1=2,x_3=2)$ | 0.4 | 0.3 | 0.3 |
| $x_1$ | 0 | 1 | 2 |
| $p(x_1\|x_3=0)$ | 0.2 | 0.4 | 0.4 |
| $p(x_1\|x_3=1)$ | 0.3 | 0.4 | 0.3 |
| $p(x_1\|x_3=2)$ | 0.4 | 0.3 | 0.3 |
| $x_4$ | 0 | 1 | 2 |
| $p(x_4\|x_3=0)$ | 0.1 | 0.4 | 0.5 |
| $p(x_4\|x_3=1)$ | 0.2 | 0.7 | 0.1 |
| $p(x_4\|x_3=2)$ | 0.3 | 0.6 | 0.1 |
| $x_5$ | 0 | 1 | 2 |
| $p(x_5\|x_3=0)$ | 0.2 | 0.6 | 0.2 |
| $p(x_5\|x_3=1)$ | 0.3 | 0.1 | 0.6 |
| $p(x_5\|x_3=2)$ | 0.4 | 0.3 | 0.3 |

*Note*: As some conditional probabilities are 0, the joint distribution does not satisfy the positivity condition. The graph of conditional dependencies of the vector $(X_1,X_2,\ldots,X_5)$ is given by Figure 1c.

over a regular lattice as the approach considered in Csiszár and Talata (2006). For example, we present a joint distribution on five nodes with the graph in Figure 1c as the graph of conditional dependencies between nodes.

**Example 3.** Consider the alphabet $A = \{0,1,2\}$ and define the joint probability distribution of the vector $(X_1,X_2,\ldots,X_5)$ using the factorization

$$p(x_1,x_2,x_3,x_4,x_5) = p(x_3)p(x_2|x_1,x_3)p(x_1|x_3)p(x_4|x_3)p(x_5|x_3) \tag{7}$$

with the conditional distributions given in Table 1. This distribution does not satisfy the positivity condition, but the basic neighborhoods are well-defined for each node, and its graph of conditional dependencies is given by Figure 1c. We consider this particular distribution later in the simulations in Section 4.

## 3 | ESTIMATION AND MODEL SELECTION

We will consider first the problem of estimating the basic neighborhood $\mathrm{ne}(v)$ of a given node $v \in V$. Then, the estimated neighborhoods will be combined to obtain the estimated graph.

Suppose we observe an independent sample with size $n$ of a subset of nodes of the random field $\{X_v : v \in V\}$, specified as follows. Let $\{V_n\}_{n\in\mathbb{N}}$ denote a sequence of finite subsets of $V$. We assume either that

(a) $V_n$ is constant, that is, $V_n = U \subset V$ for all $n \in \mathbb{N}$ and $U$ is such that $\{v\} \cup \{\mathrm{ne}(v)\} \subset U$.
(b) $V_n \nearrow V$ when $n \to \infty$, then $V_n$ will eventually contain the set $\{v\} \cup \{\mathrm{ne}(v)\}$.

From the theoretical point of view, case (a) is easier to tackle because the number of candidate neighborhoods of $v$ is finite. In case (b), this number grows with $n$, and if $V$ is infinite, the error must be controlled on an increasing set of "bad" subsets of nodes. To prove our theoretical results, we assume the size of $V_n$ grows at most as a function $o(\log n)$, so the number of subsets of $V_n$ is tractable. This condition has previously appeared in the literature; for example, (Csiszár & Talata, 2006, Theorem 2.1). For practical purposes, we will consider only (b), as (a) can be derived straightforward from (b).

Denote by $x_v^{(i)}$ the value observed at the vertex $v$ on the $i$th observation of the sample. The structure estimation problem consists of determining the set of neighbors $\mathrm{ne}(v)$ for some target nodes $v$ belonging to a finite set, based on the partial sample $\{x_v^{(1:n)} : v \in V_n\}$. Recovering the neighbors of a set of nodes enables us to recover the induced subgraph of $G$ over this set, as we prove in Corollary 1 below.

Given a vertex $v \in V_n$ and a set $W \subset V_n$ not containing $v$, the operator $N(a_v, a_W)$ will denote the number of occurrences of the event

$$\{X_v = a_v\} \cap \{X_W = a_W\}$$

in the sample. That is

$$N(a_v, a_W) = \sum_{i=1}^{n} \mathbf{1}\left\{x_v^{(i)} = a_v \ , \ x_W^{(i)} = a_W\right\}.$$

The conditional likelihood function of $X_v$ given $X_W = a_W$, for a set of parameters $\{q_a : a \in A\}$, is then

$$L\left((q_a)_{a\in A} | x_W^{(1:n)}\right) = \prod_{a\in A} q_a^{N(a,a_W)} \tag{8}$$

and the maximum likelihood estimators of the parameters are given by

$$q_a = \hat{p}(a_v|a_W) = \frac{N(a_v, a_W)}{N(a_W)}, \qquad a \in A, \tag{9}$$

for all $a_W$ with $N(a_W) > 0$, where $N(a_W) = \sum_{a_v\in A} N(a_v, a_W)$.

By multiplying all the maximum likelihoods of the conditional distribution of $X_v$ given $X_W = a_W$ for the different $a_W \in A^W$, we can compute a maximal conditional likelihood function

for vertex $v \in V$, given by

$$\hat{\mathbb{P}}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) = \prod_{a_W \in A^W} \prod_{a_v \in A} \hat{p}(a_v|a_W)^{N(a_v,a_W)}, \tag{10}$$

where the product is overall $a_W \in A^W$ with $N(a_W) > 0$ and all $a_v \in A$ with $\hat{p}(a_v|a_W) > 0$. For our purpose of estimating the basic neighborhood of $v$, this maximal conditional likelihood function suffices, but in the case of Ising models or Gibbs distributions on lattices, these functions are multiplied over the different nodes leading to what is known as *pseudo-likelihood* function, see for example Csiszár and Talata (2006) and references therein.

Before presenting the main definitions and results of this section, we state a proposition of independent interest, that shows a nonasymptotic upper bound for the rate of convergence of $\hat{p}(a_v|a_W)$ to $p(a_v|a_W)$. This proposition is related to a result obtained in Garivier and Leonardi (2011) for the estimation of the context tree of a stationary and ergodic process, and its proof is given in the appendix.

**Proposition 1.** *For all $\delta > 0$, $n \geq \exp(\delta^{-1})$, $v \in V_n$, $W \subset V_n \setminus \{v\}$ and all $a_W \in A^W$ we have*

$$\mathbb{P}\left(N(a_W) \sup_{a_v \in A} |\hat{p}(a_v|a_W) - p(a_v|a_W)|^2 > \delta \log n\right) \leq \frac{2|A|\delta \log^2 n}{n^\delta}.$$

We are ready to introduce the following neighborhood estimator for the set $\mathrm{ne}(v)$, for $v \in V$.

**Definition 1.** Given a partial sample $\{x_v^{(1:n)} : v \in V_n\}$ and a constant $c > 0$, the empirical neighborhood of $v \in V_n$ is the set of vertices $\widehat{\mathrm{ne}}(v)$ defined by

$$\widehat{\mathrm{ne}}(v) = \underset{W \subset V_n \setminus \{v\}}{\arg\max} \left\{ \log \hat{\mathbb{P}}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) - c\ |A|^{|W|} \log n \right\}. \tag{11}$$

To state our main results, we recall the definition of the Kullback–Leibler divergence between two probability distributions $p$ and $q$ over $A$. It is given by

$$D(p;q) = \sum_{a \in A} p(a) \log \frac{p(a)}{q(a)} \tag{12}$$

where, by convention, $p(a) \log \frac{p(a)}{q(a)} = 0$ if $p(a) = 0$ and $p(a) \log \frac{p(a)}{q(a)} = +\infty$ if $p(a) > q(a) = 0$. An important property of the Kullback–Leibler divergence is that $D(p;q) = 0$ if and only if $p(a) = q(a)$ for all $a \in A$.

For any $v \in V$ denote by

$$p_n(v) = \min_{\substack{a_v, a_W \\ W \subset V_n \setminus \{v\}}} \{\ p(a_v|a_W) : p(a_v|a_W) > 0\ \}$$

and

$$\alpha_n(v) = \min_{\substack{W \subset V_n \setminus \{v\} \\ \mathrm{ne}(v) \not\subset W}} \left\{ \sum_{a_{W \cup \mathrm{ne}(v)}} p(a_{W \cup \mathrm{ne}(v)}) D(p(\cdot_v|a_{\mathrm{ne}(v)})\ ;\ p(\cdot_v|a_W)) \right\}$$

where $p(\cdot_v|a_{\mathrm{ne}(v)})$ denotes the probability distribution over $A$ given by $\{p(a_v|a_{\mathrm{ne}(v)}\}_{a_v\in A}$ and similarly for $p(\cdot_v|a_W)$. We note that for any vertex $v \in V$ and any $n \in \mathbb{N}$, we must have $p_n(v) > 0$ and also by the definition of the basic neighborhood and Lemma 3, we must have $\alpha_n(v) > 0$. For simplicity in the proofs and as is usual in the literature, we will assume that these quantities are uniformly bounded from below by positive constants; that is, we assume that

$$p_* = \inf_v \inf_n \{\ p_n(v)\ \} \quad \text{and} \quad \alpha_* = \inf_v \inf_n \{\ \alpha_n(v)\ \}$$

are positive constants. Observe that $p_* > 0$ is not equivalent to the positivity condition, where *all* the conditional probabilities are assumed to be positive. Here we assume that those positive probabilities in the model are bounded from below, but some can be zero.

We can now state the following consistency result for the neighborhood estimator given in (11).

**Theorem 1.** *Let $v \in V$. Assume $V_n \nearrow V$ with $|V_n| = o(\log n)$. Assume also that $p_* > 0$ and $\alpha_* > 0$. Then for any $c > 0$, the estimator given by (11) satisfies $\widehat{\mathrm{ne}}(v) = \mathrm{ne}(v)$ with probability converging to 1 as $n \to \infty$. Moreover, if $c > |A|^2[p_*(|A| - 1)]^{-1}$ then $\widehat{\mathrm{ne}}(v) = \mathrm{ne}(v)$ eventually almost surely as $n \to \infty$.*

Once we have guarantees that we can consistently estimate the neighborhood of a node $v \in V$, we can consider the estimation of a finite subgraph of $G$. To do this, we can estimate the neighborhood of each node and reconstruct the subgraph based on the set of estimated neighborhoods. Given a set $V' \subset V$, we denote by $G_{V'}$ the induced subgraph; that is, the graph given by the pair $(V', E')$, where $E' = \{(v, w) \in E : v, w \in V'\}$. Based on the neighborhood estimator (11), we can construct an estimator of the subgraph $G_{V'}$ by defining the set of edges

$$\widehat{E}'_\wedge = \{(v, w) \in V' \times V' : v \in \widehat{\mathrm{ne}}(w) \ \text{ and } \ w \in \widehat{\mathrm{ne}}(v)\} \tag{13}$$

where this is refereed as the *conservative approach* or we can take

$$\widehat{E}'_\vee = \{(v, w) \in V' \times V' : v \in \widehat{\mathrm{ne}}(w) \ \text{ or } \ w \in \widehat{\mathrm{ne}}(v)\}\ , \tag{14}$$

a *nonconservative* approach.

Theorem 1 then implies the following strong consistency result for any $G'$ with a finite set of vertices $V'$.

**Corollary 1.** *Let $G' = (V', E')$ be an induced subgraph of $G$ with a finite set of vertices $V'$, and assume the hypotheses of Theorem 1 hold. Then, for $c > 0$ (respectively $c > |A|^2[p_*(|A| - 1)]^{-1}$), if $|V_n| = o(\log n)$ we have $\widehat{E}'_\wedge = \widehat{E}'_\vee = E'$ with probability converging to 1 as $n \to \infty$ (respectively eventually almost surely as $n \to \infty$).*

The proofs of all the theoretical results in this section and some auxiliary results are presented in the Appendix.

## 4 | SIMULATIONS

In this section, we show the results of a simulation study to evaluate the performance of the graph estimators (13) and (14) on different models, sample sizes and different values of the penalizing constant $c$.

We first simulated a probability distribution on five vertices with the alphabet $A = \{0, 1, 2\}$ and with a graph of conditional dependencies given by Figure 1c. The joint distribution is assumed to factorize as in (7), having conditional probabilities given by Table 1. Figure 2 shows the results of both approaches for values of the penalizing constant $c$ in the set $\{0.25, 0.5, 1, 1.5, 2\}$ and sample sizes $n$ in the set $\{100,200, 500, 1000, 2500\}$. In this example, the nonconservative approach seems to converge to the underlying graph faster than the conservative approach. To evaluate this difference in a quantitative form, we compute a numeric value for the underestimation error (*ue*), overestimation error (*oe*) and total error (*te*), given by

$$ue = \frac{\sum_{(v,w)} \mathbf{1}\left\{(v,w) \in E \text{ and } (v,w) \notin \widehat{E}\right\}}{\sum_{(v,w)} \mathbf{1}\{(v,w) \in E\}} \tag{15}$$

$$oe = \frac{\sum_{(v,w)} \mathbf{1}\left\{(v,w) \notin E \text{ and } (v,w) \in \widehat{E}\right\}}{\sum_{(v,w)} \mathbf{1}\{(v,w) \notin E\}} \tag{16}$$

and

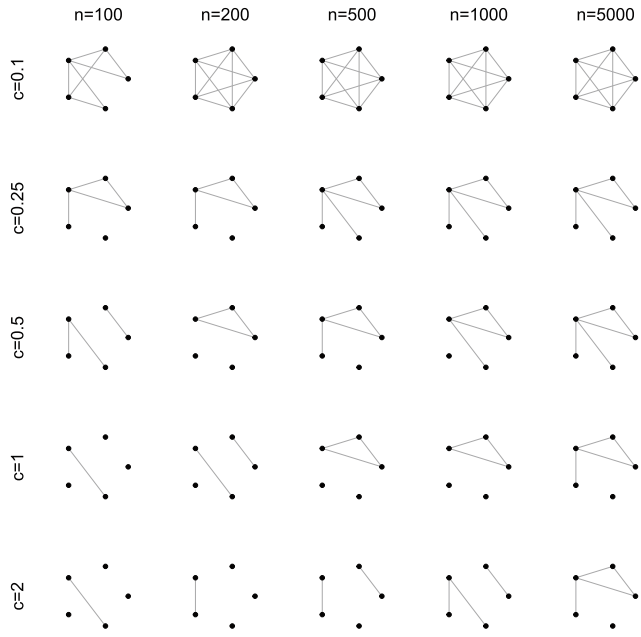
$$te = \frac{ue \sum_{(v,w)} \mathbf{1}\{(v,w) \in E\} + oe \sum_{(v,w)} \mathbf{1}\{(v,w) \notin E\}}{|V|(|V| - 1)}. \tag{17}$$

Figure 3 shows an evaluation of *ue*, *oe* and *te* for both methods, run with penalizing constant $c = 1$, and with sample sizes ranging from 100 to 10,000. As before, the nonconservative approach (PCL-NC) performs better than the conservative approach (PCL-C). In the Supplementary Material to this article, we present similar results for other graph structures, as the line graph shown in Figure 1a and the complete graph with five vertices.

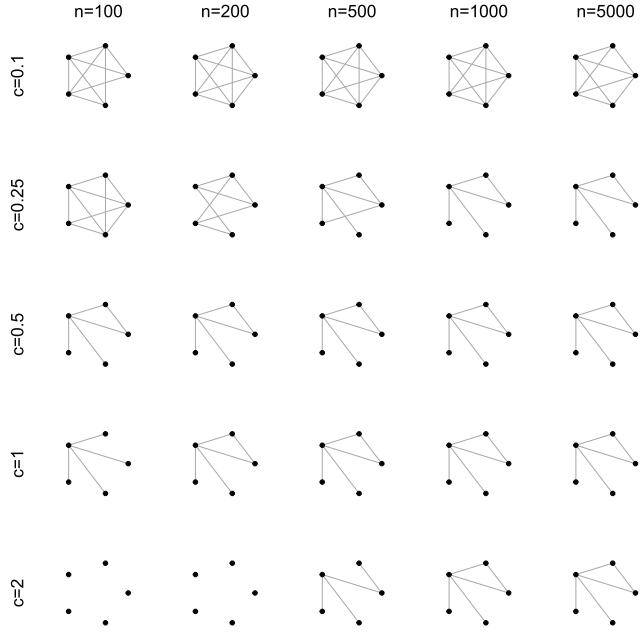
To evaluate the methods on more general graphs, we implemented a simulation selecting different graphs and distributions and comparing the performance of our methods with the $\ell_1$-regularized logistic regression approach for binary models proposed by Ravikumar et al. (2010).

In Figure 4, we show the proportion of correct reconstruction of the graph on 20 replications of the simulations, as a function of sample size, for the nonconservative approach, PCL-NC (left) and the conservative approach, PCL-C (right). In both figures, we compare the performance of our methods with different penalizing constants to the performance of the $\ell_1$-regularized logistic regression approach (LASSO). On each replication, a different graph with ten nodes and ten edges was randomly drawn, and a sample of the model was generated using the `R` package `IsingSampler`. The LASSO estimator was computed with the `R` package `rIsing`. In the figures, we see that both methods, PCL-NC and PCL-C, with different constants, have an increased recovery proportion as the sample size increases, contrary to what happens with LASSO, which seems to maintain a nonzero error even for large sample sizes.

We also evaluated what happens when we consider graphs with different numbers of edges (from more sparse to more dense graphs). As before, in each replication, we randomly chose a graph with a specified number of edges and then generated a sample using the `R` package `IsingSampler`. In Figure 5, we see that, for all except for $c = 0.25$, the nonconservative approach has excellent performance for sparse graphs. Then the performance decreases drastically as the number of edges increases. This is more evident for small sample sizes such as $n = 2500$ or $n = 5000$ (top figures). This shows, in particular, that selecting the penalizing constant is




(a) Conservative approach defined by (13).




(b) Non-conservative approach defined by (14).

**FIGURE 2** Estimated graphs for different values of the constants $c$ in the penalty term and different sample sizes $n$, in the case of the conservative approach (a) and the nonconservative approach (b).

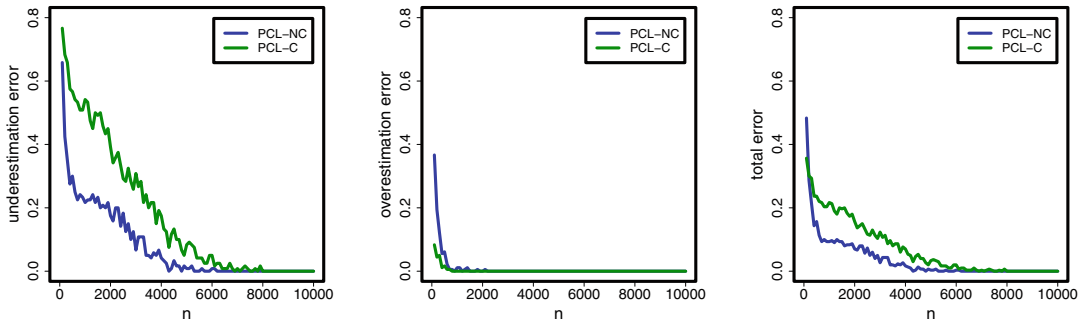



**FIGURE 3** Mean of underestimation error (left), overestimation error (center) and total error (right) defined by (15)–(17), computed on 30 runs of the simulation for both conservative and nonconservative approaches and with penalizing constant $c = 1$.

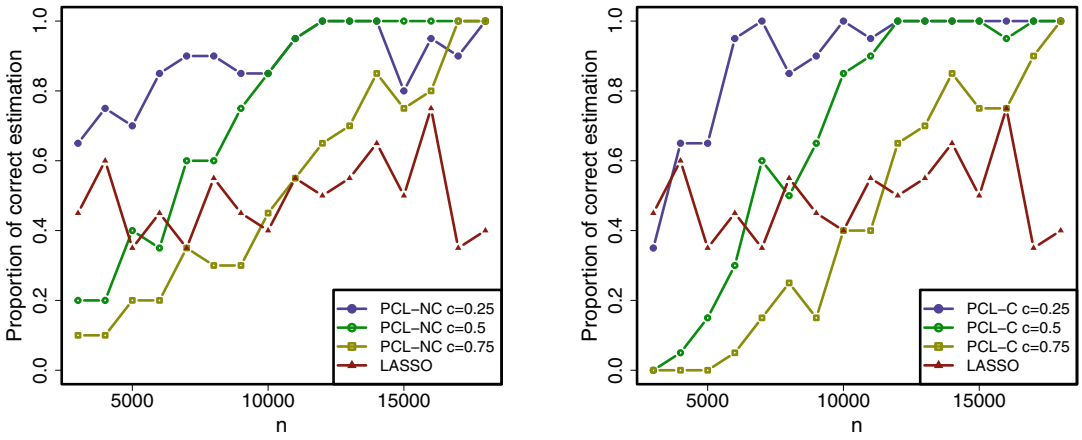



**FIGURE 4** The proportion of correct estimation on 20 runs of the simulation, as a function of sample size $n$, comparing the nonconservative approach (PCL-NC) and the conservative approach (PCL-C) with different penalizing constants $c$, to LASSO. We simulated different distributions with 10 nodes and 10 edges for each sample size and replication.

a crucial step for this type of method and that the subjacent model influences the algorithms' performance. On the other hand, even though the LASSO implementation has an optimization step to select the penalizing constant, its performance is almost constant for different sample sizes and also decreases as the number of edges in the graph increases.

The algorithms implementing the graph estimators given by (13), PCL-C (conservative approach) and (14), PCL-NC (nonconservative approach) are available as the `R` package called `mrfse`, and can be installed from the `R` repository.

## 5 | APPLICATION ON REAL DATA

To illustrate the performance of the estimator on real data we analyzed a stock index from fifteen countries at different times taken from the site https://br.investing.com/indices/world-indices.

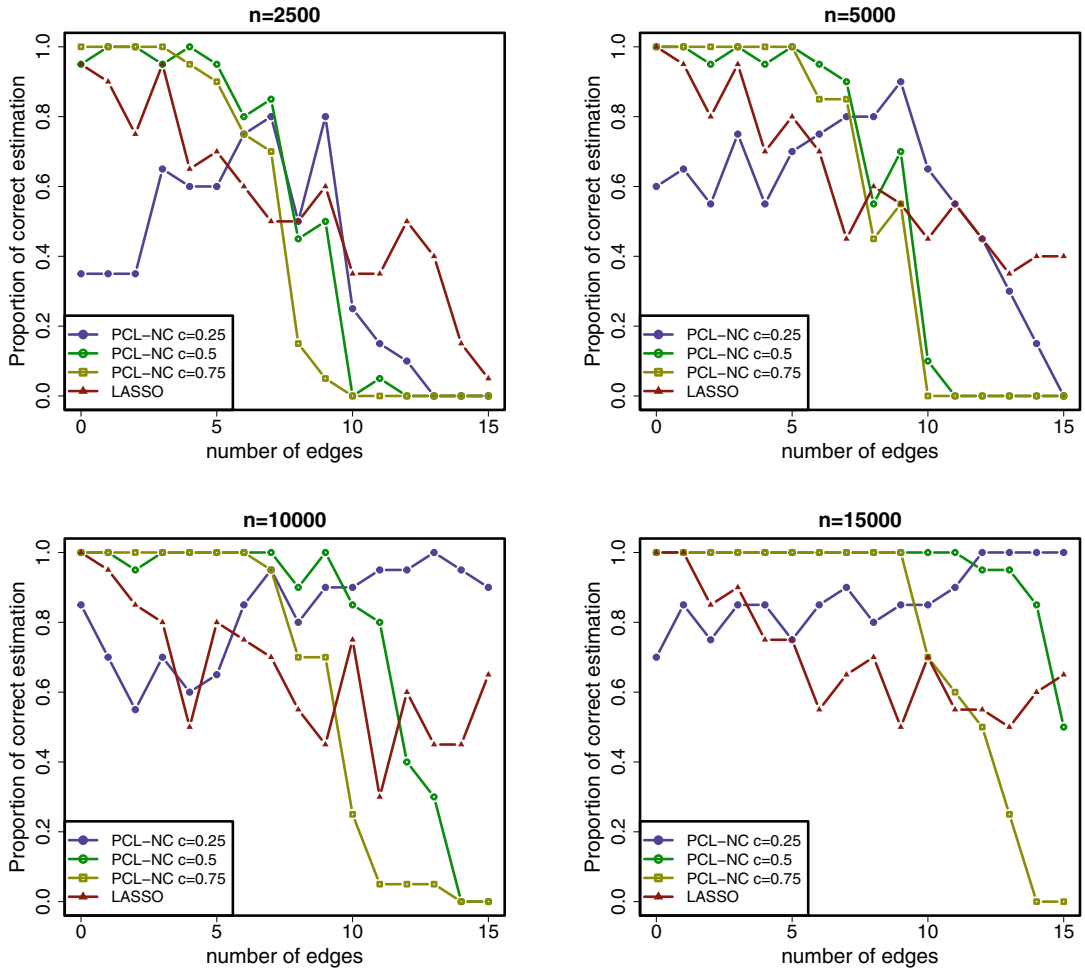
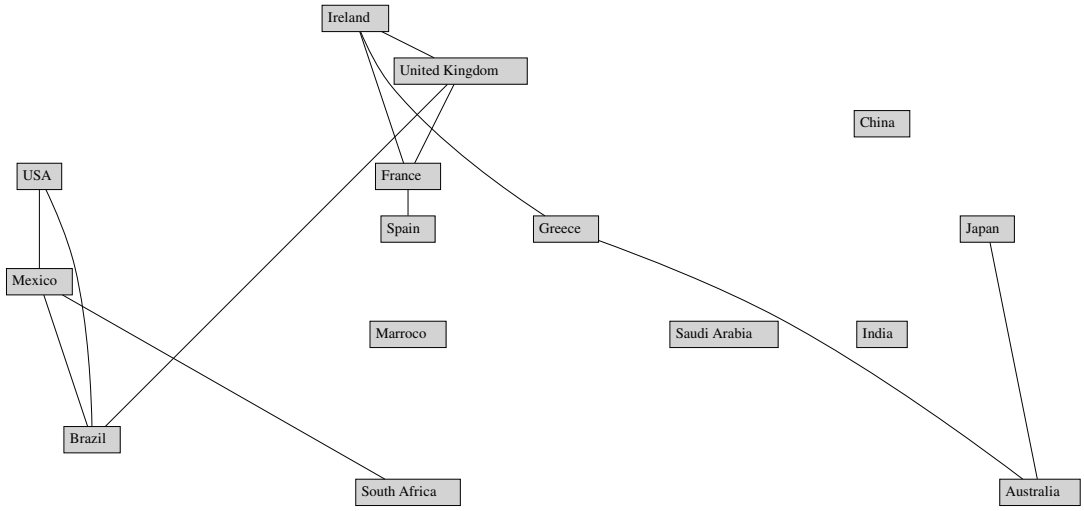
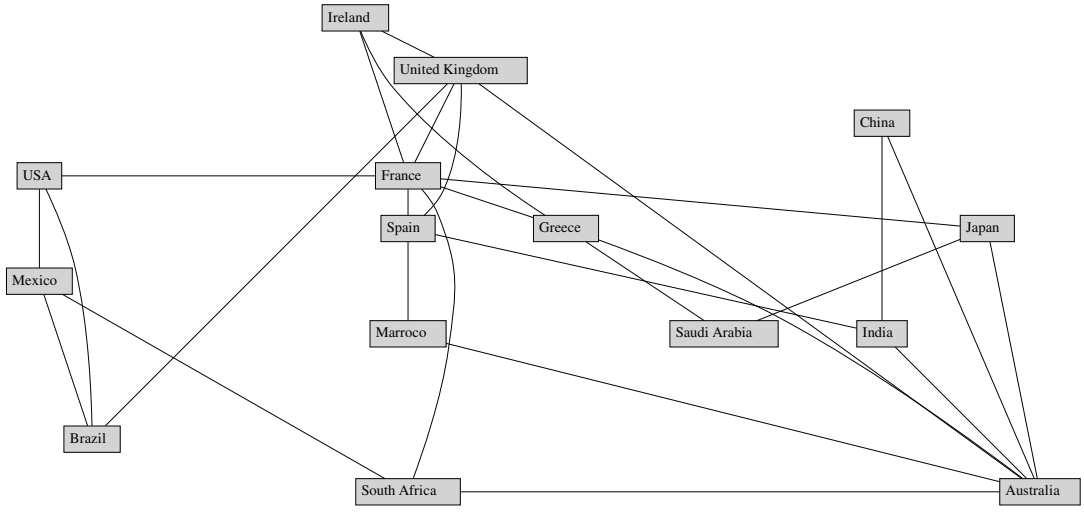



**FIGURE 5** The proportion of correct estimation on 20 runs of the simulation, as a function of the number of edges in the graph, comparing the nonconservative approach (PCL-NC) with different penalizing constants $c$, to LASSO, for four different sample sizes. We simulated different distributions with ten nodes and the specified number of edges in each replication.

The countries are Brazil, USA, UK, France, India, Japan, Greece, Ireland, South Africa, Spain, Marroco, Australia, Mexico, China and Saudi Arabia, with stock markets Bovespa, NASDAQ, FTSE 100, CAC 40, Nifty 50, Nikkei 225, FTSE ATHEX Large Cap, FTSE Ireland, FTSE South Africa, IBEX 35, Moroccan All Shares, S&P ASX 200, S&P BMV IPC, Shanghai Composite and Tadawul All Share, respectively. We collected 2120 entries where each entry contains the indicator function of an increasing variation in the stock index for a given day with respect to the previous day, for each one of the fifteen stock markets. That is, a stock market for a given day $d$ is codified as 1 if the stock index at day $d$ is greater than the stock index at day $d-1$, and 0 otherwise. The main goal is to estimate the conditional dependence graph between the codified stock markets corresponding to the fifteen countries. The dataset of stock index variation corresponds to subsequent time points (days) in the period from December 29, 2010 to October 22, 2018. We selected data points with a difference of 4 days to reduce sample correlation between subsequent




(a) Estimated graph with the conservative approach defined by (13).




(b) Estimated graph with the non-conservative approach defined by (13).

**FIGURE 6** Estimated graphs with the conservative (a) and nonconservative (b) approaches. The constant $c$ in (1) was selected by cross-validation, as the text explains.

observations. The final sample has a total of 530 time points. The dataset is available at https://github.com/rodrigorsdc/ic/tree/master/stock_data.

We applied the two approaches given by (13) and (14) to estimate the conditional dependence graph. To select the penalizing constant, we use 10-Fold Cross-validation in the following sense. The observations are randomly divided into approximately equal-sized 10 groups or folds. At each step, one of the 10 groups is treated as a validation set, and the remaining 9 groups are the training set used to estimate the model. For each value of $c$ in a grid, the graph and the conditional

probabilities are estimated on the training set, and the log conditional likelihood is computed over the validation set, that is, we compute

$$\ell(c) = \sum_{v \in V_n} \sum_{a_v, a_{\widehat{\mathrm{ne}}(v)}} N(a_v, a_{\widehat{\mathrm{ne}}(v)}) \log \widehat{p}(a_v | a_{\widehat{\mathrm{ne}}(v)}) \,, \tag{18}$$

where $N(a_v, a_{\widehat{\mathrm{ne}}(v)})$ is computed over the validation set, and $\widehat{\mathrm{ne}}(v)$ and $\widehat{p}(a_v | a_{\widehat{\mathrm{ne}}(v)})$ are computed on the training set. This procedure is repeated 10 times; at the end, we associate to each value of $c$ the mean of the log conditional likelihoods over the ten different folds and choose the value of $c$ maximizing the mean over the ten folds.

For the stock index data, the procedure described above was used for each node and values of $c$ in the interval $[0.01, 2]$, with step 0.01. The resulting graphs with the conservative and nonconservative approaches are shown in Figure 6. In both cases, the obtained graphs connect geographically near countries, as could be somehow expected. Also, the conservative approach underestimates a few edges compared to the nonconservative one.

## 6 | DISCUSSION

In this paper, we introduced an estimator for the basic neighborhood of a node in a general discrete Markov random field defined on a graph. We showed that the estimator is consistent for any value of the penalizing constant and is strongly consistent for a sufficiently large value of the constant. This result implies that any finite subgraph can be recovered with probability one when the sample size diverges, provided the set of observed nodes increases not too fast or contains all the target nodes and their neighborhoods. The proof is based on some deviation inequalities given in Proposition 1, a result derived from a martingale approach appearing in Garivier and Leonardi (2011) but new in this context of Markov random fields. One advantage of our results is that we do not need to assume a positivity condition, namely that all conditional probabilities in the model are strictly positive. This allows us to consider sparse models with many parameters equal to zero and then a few significant parameters. This property is appealing in high-dimensional contexts. We consider the Markov random field samples to be independent and identically distributed. Still, one important question to address in future work is if this method can be generalized to dependent data, for example, the case of mixing processes as considered in Leonardi et al. (2021). Another open problem is the necessity of the condition on $c$ for almost sure convergence in Theorem 1. We conjecture that the estimator of the basic neighborhood must be strongly, and not only weakly, consistent for any $c > 0$, but the proof of this fact would undoubtedly need other techniques different from that developed in this work and is out of the scope of this paper.

### ACKNOWLEDGMENTS

This work was produced as part of the activities of the *Research, Innovation and Dissemination Center for Neuromathematics* (grant FAPESP 2013/07699-0). It was also supported by FAPESP project (grant 2017/10555-0) *"Stochastic Modeling of Interacting Systems"* and CNPq Universal project (grant 432310/2018-5) *"Statistics, stochastic processes and discrete structures"*. FL is partially supported by a CNPq's research fellowship, grant 311763/2020-0. During the development of this work, RC and IF were supported by a CAPES fellowship.

## ORCID

*Florencia Leonardi* https://orcid.org/0000-0002-0299-0680

## REFERENCES

Atchade, Y. F. (2014). Estimation of high-dimensional partially-observed discrete Markov random fields. *Electronic Journal of Statistics*, *8*(2), 2242–2263.

Bresler, G., Gamarnik, D., & Shah, D. (2018). Learning graphical models from the Glauber dynamics. *IEEE Transactions on Information Theory*, *64*(6), 4072–4080. https://doi.org/10.1109/TIT.2017.2713828

Csiszár, I., & Talata, Z. (2006). Consistent estimation of the basic neighborhood of Markov random fields. *The Annals of Statistics*, *34*(1), 123–145. https://doi.org/10.1214/009053605000000912

Divino, F., Frigessi, A., & Green, P. J. (2000). Penalized pseudolikelihood inference in spatial interaction models with covariates. *Scandinavian Journal of Statistics*, *27*(3), 445–458. https://doi.org/10.1111/1467-9469.00200

Duarte, A., Galves, A., Löcherbach, E., & Ost, G. (2019). Estimating the interaction graph of stochastic neural dynamics. *Bernoulli*, *25*(1), 771–792. https://doi.org/10.3150/17-bej1006

Frondana, I. M. (2016). *Model selection for discrete Markov random fields on graphs* [PhD thesis]. Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Galves, A., Orlandi, E., & Takahashi, D. Y. (2015). Identifying interacting pairs of sites in Ising models on a countable set. *Brazilian Journal of Probability and Statistics*, *29*(2), 443–459.

Garivier, A., & Leonardi, F. (2011). Context tree selection: A unifying view. *Stochastic Processes and their Applications*, *121*(11), 2488–2506. https://doi.org/10.1016/j.spa.2011.06.012

Georgii, H.-O. (2011). *Gibbs measures and phase transitions*. In *de Gruyter studies in mathematics* (Vol. *9*, 2nd ed.). Walter de Gruyter & Co.

Hammersley, J. M., & Clifford, P. (1971). *Markov fields on finite graphs and lattices* [unpublished manuscript]. http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf

Höfling, H., & Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, *10*(32), 883–906. http://jmlr.org/papers/v10/hoefling09a.html

Ji, C., & Seymour, L. (1996). A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *The Annals of Applied Probability*, *6*(2), 423–443.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques—Adaptive computation and machine learning*. The MIT Press ISBN 0262013193, 9780262013192.

Lafferty, J., Liu, H., & Wasserman, L. (2012). Sparse nonparametric graphical models. *Statistical Science*, *27*(4), 519–537. https://doi.org/10.1214/12-STS391

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Leonardi, F., Lopez-Rosenfeldz, M., Rodriguez, D., Severino, M. T. F., & Sued, M. (2021). Independent block identification in multivariate time series. *Journal of Time Series Analysis*, *42*(1), 19–33.

Lerasle, M., & Takahashi, D. Y. (2016). Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields. *Bernoulli*, *22*(1), 325–344.

Liu, H., Han, M., Yuan, F., Lafferty, L., & Wasserman, J. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, *40*(4), 2293–2326. https://doi.org/10.1214/12-AOS1037

Löcherbach, E., & Orlandi, E. (2011). Neighborhood radius estimation for variable-neighborhood random fields. *Stochastic Processes and their Applications*, *121*(9), 2151–2185.

Loh, P.-L., & Wainwright, M. J. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, *41*(6), 3022–3049.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436–1462.

Pensar, J., Nyman, H., & Corander, J. (2017). Structure learning of contextual Markov networks using marginal pseudo-likelihood. *Scandinavian Journal of Statistics*, *44*(2), 455–479. https://doi.org/10.1111/sjos.12260

Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional Ising model selection using $l_1$-regularized logistic regression. *The Annals of Statistics*, *38*(3), 3022–1319.

Santhanam, N. P., & Wainwright, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, *58*(7), 4117–4134.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.
Shojaie, A., & Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, *97*(3), 519–538.
Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, *85*(409), 204–212.
Tjelmeland, H., & Besag, J. (1998). Markov random fields with higher-order interactions. *Scandinavian Journal of Statistics*, *25*(3), 415–433.
Yang, E., Ravikumar, P., Allen, G. I., & Liu, Z. (2015). Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, *16*(115), 3813–3847.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Leonardi, F., Carvalho, R., & Frondana, I. (2024). Structure recovery for partially observed discrete Markov random fields on graphs under not necessarily positive distributions. *Scandinavian Journal of Statistics*, *51*(1), 64–88. https://doi.org/10.1111/sjos.12674

## APPENDIX. PROOF OF THEORETICAL RESULTS

Here we present the proofs of the main results in the paper, namely Lemma 1, Proposition 1, Theorem 1 and Corollary 1. We also prove some auxiliary results needed to demonstrate the main theorem in the article.

*Proof of Lemma* 1. Suppose $w \notin \text{ne}(v)$. Take any $\Delta$ such that $\text{ne}(v) \subseteq \Delta, v \notin \Delta$. Then

$$p(a_v|a_\Delta) = p(a_v|a_{\Delta\setminus\{w\}}) \quad \text{for all } a_\Delta \in A^\Delta \text{ with } p(a_\Delta) > 0. \tag{A1}$$

By the definition of conditional probability and (A1) we have that

$$\begin{aligned} p(a_v, a_w|a_{\Delta\setminus\{w\}}) &= p(a_v|a_\Delta, a_w)p(a_w|a_{\Delta\setminus\{w\}}) \\ &= p(a_v|a_{\Delta\setminus\{w\}})p(a_w|a_{\Delta\setminus\{w\}}) \end{aligned} \tag{A2}$$

for all $a_\Delta$ with $p(a_\Delta) > 0$. But we also have that

$$p(a_v, a_w|a_{\Delta\setminus\{w\}}) = p(a_w|a_{\Delta\setminus\{w\}}, a_v)p(a_v|a_{\Delta\setminus\{w\}}). \tag{A3}$$

Therefore, by (A2) and (A3), if $p(a_v|a_{\Delta\setminus\{w\}}) > 0$ we obtain

$$p(a_w|a_{\Delta\setminus\{w\}}, a_v) = p(a_w|a_{\Delta\setminus\{w\}}).$$

As the equality holds for all $\Delta$ and all $(a_v, a_{\Delta\setminus\{w\}})$ with $p(a_v, a_{\Delta\setminus\{w\}}) > 0$ then we conclude that $v \notin \text{ne}(w)$. ∎

*Proof of Proposition* 1. First, observe that

$$\mathbb{P}\Big(N(a_W)\sup_{a_v\in A}|\hat{p}(a_v|a_W)-p(a_v|a_W)|^2>\delta\log n\Big)\\ \leq\sum_{a_v\in A}\mathbb{P}\left(N(a_W)\;|\hat{p}(a_v|a_W)-p(a_v|a_W)|^2>\delta\log n\right) \tag{A4}$$

then we will fix $a_v\in A$ and bound above each term in the right hand side separately. For simplifying the notation we write $\hat{p}_n=\hat{p}(a_v|a_W)$, $p=p(a_v|a_W)$, $O_n=N(a_v,a_W)$ and $N_n=N(a_W)$. Observe that $\hat{p}_n=O_n/N_n$. For $\lambda>0$ define $\phi(\lambda)=\log(1-p+2^\lambda p)$. Let $W_0^\lambda=1$ and for $n\geq 1$ define

$$W_n^\lambda=2^{\lambda O_n-N_n\phi(\lambda)}.$$

Observe that $W_n^\lambda$ is a martingale with respect to $\mathcal{F}_n=\sigma(X_{v,W}^{(1:n)},X_W^{(n+1)})$ and that $\mathbb{E}[\,W_n^\lambda\,]=1$. In fact, conditioned on $\mathcal{F}_n$ we have that

$$O_{n+1}-O_n=\begin{cases}1, & \text{if } x_v^{(n+1)}=a_v,\ x_W^{(n+1)}=a_W;\\ 0, & \text{c.c.}\end{cases}$$

and similarly

$$N_{n+1}-N_n=\begin{cases}1, & \text{if } x_W^{(n+1)}=a_W;\\ 0, & \text{c.c.}\end{cases}$$

Observe that if $x_W^{(n+1)}=a_W$ then

$$\mathbb{E}\left[2^{\lambda(O_{n+1}-O_n)}\mid\mathcal{F}_n\right]=\mathbb{E}\left[2^{\lambda\mathbf{1}\{x_v^{(n+1)}=a_v\}}\mid\mathcal{F}_n\right]\\ =2^{\phi(\lambda)}\\ =2^{(N_{n+1}-N_n)\phi(\lambda)}. \tag{A5}$$

On the other hand, if $x_W^{(n+1)}\neq a_W$ the equality trivially holds. Then rearranging the terms in (A5) we conclude that

$$\mathbb{E}\left[2^{\lambda O_{n+1}-N_{n+1}\phi(\lambda)}\mid\mathcal{F}_n\right]=2^{\lambda O_n-N_n\phi(\lambda)}$$

and $W_n^\lambda$ is a martingale with respect to $\mathcal{F}_n$. Now divide the interval $\{1,\dots,n\}$ of possible values of $N_n$ into "slices" $\{t_{k-1}+1,\dots,t_k\}$ of geometrically increasing size, and treat the slices independently. We take $\alpha=\delta\log n$, assuming that $n$ is sufficiently large so that $\alpha>1$. Take $\eta=1/(\alpha-1)$, $t_0=0$ and for $k\geq 1$, $t_k=\lfloor(1+\eta)^k\rfloor$. Let $m$ be the first integer such that $t_m\geq n$, that is

$$m=\left\lceil\frac{\log n}{\log(1+\eta)}\right\rceil.$$

Define the events $B_k = \{t_{k-1} < N_n \le t_k\} \cap \left\{N_n\ |\hat{p}_n - p|^2 > \alpha\right\}$. We have

$$\mathbb{P}\left(N_n\ |\hat{p}_n - p|^2 > \alpha\right) \le \mathbb{P}\left(\bigcup_{k=1}^{m} B_k\right) \le \sum_{k=1}^{m} \mathbb{P}(B_k). \tag{A6}$$

Without loss of generality, we can assume that $\hat{p} \ge p$ (the case $\hat{p} \le p$ holds by symmetry). Observe that $|x-p|^2$ is a continuous increasing function for $x \in [p;1]$, with $0 \le |x-p|^2 \le |1-p|^2$. Let $x$ be such that $|x-p|^2 = \alpha/(1+\eta)^k$, that is we take

$$x = \sqrt{\frac{\alpha}{(1+\eta)^k}} + p.$$

Observe that $x \in [p,1]$ unless $\alpha/(1+\eta)^k > |1-p|^2$. But in this case, we have that if $N_n \le (1+\eta)^k$ then

$$\alpha > (1+\eta)^k |1-p|^2 \ge N_n |\hat{p}_n - p|^2$$

so $\mathbb{P}(B_k) = 0$. So we may assume that such an $x$ always exists over the nonempty events $B_k$. Moreover, on $B_k$ we have that $|\hat{p}_n - p|^2 \ge \alpha/N_n \ge \alpha/(1+\eta)^k$ then we must have $\hat{p}_n \ge x$. Now take $\lambda$ such that $\lambda x - \phi(\lambda) \ge |x-p|^2$. Then on $B_k$ we have that

$$\lambda \hat{p}_n - \phi(\lambda) \ge \lambda x - \phi(\lambda) \ge |x-p|^2 = \frac{\alpha}{(1+\eta)^k} \ge \frac{\alpha}{(1+\eta)N_n}$$

therefore

$$\begin{aligned} B_k &\subset \left\{\lambda \hat{p}_n - \phi(\lambda) > \frac{\alpha}{(1+\eta)N_n}\right\} \\ &\subset \left\{W_n^{\lambda} > 2^{\alpha/(1+\eta)}\right\}. \end{aligned}$$

As $\mathbb{E}\left[W_n^{\lambda}\right] = 1$, Markov's inequality implies that

$$\begin{aligned} \mathbb{P}(B_k) &\le \mathbb{P}\left(W_n^{\lambda} > 2^{\alpha/(1+\eta)}\right) \\ &\le 2^{-\alpha/(1+\eta)}. \end{aligned} \tag{A7}$$

Finally, by (A6) we have that

$$\mathbb{P}\left(N_n |\hat{p}_n - p|^2 > \alpha\right) \le m\ 2^{-\alpha/(1+\eta)}.$$

But as $\eta = 1/(\alpha-1)$, $m = \left\lceil \frac{\log n}{\log(1+\eta)} \right\rceil$ and $\log(1+1/(\alpha-1)) \ge 1/\alpha$ we obtain

$$\mathbb{P}\left(N_n\ |\hat{p}_n - p|^2 > \alpha\right) \le 2\alpha \log(n) 2^{-\alpha} = \frac{2\delta \log^2(n)}{n^{\delta}}.$$

Finally, by (A4) we obtain that

$$\mathbb{P}\left(N(a_W) \sup_{a_v \in A} |\hat{p}(a_v|a_W) - p(a_v|a_W)|^2 > \delta \log n\right) \le \frac{2|A|\delta \log^2 n}{n^{\delta}}.$$

■

Now we state a result controlling the probability in Proposition 1 for all possible neighborhoods simultaneously.

**Proposition 2.** *For all $\delta > 0$, all $v \in V_n$ and $|V_n| = o(\log n)$ we have*

$$\mathbb{P}\left( \sup_{W \subset V_n \setminus \{v\}} \sup_{a_W \in A^W} \sup_{a_v \in A} N(a_W)\ |\hat{p}(a_v|a_W) - p(a_v|a_W)|^2 < \delta \log n \right) \to 1$$

*when $n \to \infty$. Moreover, if $\delta > 1$, the probability equals one for all sufficiently large $n$.*

*Proof.* Assume $|V_n| = \epsilon_n \log n$, with $\epsilon_n \to 0$ when $n \to \infty$. By Proposition 1 and a union bound, we have that

$$\begin{aligned} \mathbb{P}( \sup_{W \subset V_n \setminus \{v\}} \sup_{a_W \in A^W} \sup_{a_v \in A} N(a_W)\ |\hat{p}(a_v|a_W) - p(a_v|a_W)|^2 > \delta \log n) & \\ \leq 2^{|V_n|} |A|^{|V_n|} \frac{2|A|\delta \log^2 n}{n^\delta} & \\ \leq \frac{c\delta \log^2 n}{n^{\delta - 2\epsilon_n}}\,. & \end{aligned}$$

For $\delta > 0$, the bound on the right-hand side converges to 0 and it is summable in $n$ for any $\delta > 1$. Then the almost sure convergence follows by the Borel–Cantelli lemma. ∎

The following primary result about the Kullback–Leibler divergence corresponds to (Csiszár & Talata, 2006, Lemma 6.3). We omit its proof here.

**Lemma 2.** *For any $P$ and $Q$ we have*

$$D(P;Q) \leq \sum_{a \in A : Q(a) > 0} \frac{[P(a) - Q(a)]^2}{Q(a)}.$$

The following lemma was proved in (Csiszár & Talata, 2006, Lemma A.2) for translation invariant Markov random fields. As our setting is different, we include its proof here.

**Lemma 3.** *If a neighborhood $W$ of $v \in V$ satisfies*

$$p(a_v|a_W) = p(a_v|a_{\mathrm{ne}(v)})$$

*for all $a_v \in A$, and all $a_{W \cup \mathrm{ne}(v)} \in A^{W \cup \mathrm{ne}(v)}$ with $p(a_{W \cup \mathrm{ne}(v)}) > 0$ then $W$ is a Markov neighborhood.*

*Proof.* We have to show that for any $\Delta \subset V$ finite, with $\Delta \supset W$,

$$p(a_v|a_\Delta) = p(a_v|a_W) \tag{A8}$$

for all $a_v \in A$ and all $a_\Delta \in A^\Delta$ with $p(a_\Delta) > 0$. As $\mathrm{ne}(v)$ is a Markov neighborhood, the lemma's condition implies

$$p(a_v|a_W) = p(a_v|a_{\mathrm{ne}(v)}) = p(a_v|a_{\mathrm{ne}(v) \cup \Delta})$$

or all $a_v \in A$ and all $a_{\mathrm{ne}(v)\cup\Delta} \in A^{\mathrm{ne}(v)\cup\Delta}$ with $p(a_{\mathrm{ne}(v)\cup\Delta}) > 0$. So (A8) follows, because $W \subseteq \Delta \subseteq \mathrm{ne}(v) \cup \Delta$. ∎

The following proposition guarantees uniform control of all empirical marginal probabilities for subsets of variables in $V_n$.

**Proposition 3.** *Let $\{V_n\}_{n\in\mathbb{N}}$ be such that $|V_n| = o(\log n)$. Then for all $\delta > 2$ we have*

$$|\hat{p}(a_W) - p(a_W)| < \sqrt{\frac{\delta \log n}{n}}$$

*simultaneously for all $W \subseteq V_n$ and $a_W \in A^W$, eventually almost surely as $n \to \infty$.*

*Proof.* For $W \subset V_n$ and $a_W \in A^W$ define

$$Y_i(a_W) = \mathbf{1}\{x_W^{(i)} = a_W\} - p(a_W)\,, \qquad i = 1, 2, \ldots, n.$$

Note that $\mathbb{E}(Y_i(a_W)) = 0$ and $|Y_i(a_W)| \le 1$ for all $i = 1, 2, \ldots, n$. Then by Hoeffding's Inequality, we have that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_W) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i(a_W)\right]\right| \ge t\right) \le 2\exp\left(-\frac{nt^2}{2}\right).$$

Observe that

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(a_W) = \frac{N(a_W)}{n} - p(a_W)$$

and

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i(a_W)\right] = 0.$$

Therefore

$$\mathbb{P}\left(\left|\frac{N(a_W)}{n} - p(a_W)\right| \ge t\right) \le 2\exp\left(-\frac{nt^2}{2}\right).$$

Taking $t = \sqrt{\frac{\delta \log n}{n}}$ we have that

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{p}(a_W) - p(a_W)\right| \ge \sqrt{\frac{\delta \log n}{n}} \text{ for some } W \subset V_n \text{ and } a_W \in A^W\right) \\
&\le \sum_{W\subset V_n}\sum_{a_W\in A^W} \mathbb{P}\left(\left|\hat{p}(a_W) - p(a_W)\right| \ge \sqrt{\frac{\delta \log n}{n}}\right) \\
&\le 2^{|V_n|}|A|^{|V_n|}\, 2\exp\left(-\frac{\delta \log n}{2}\right)
\end{aligned}$$

which is summable in $n$ for $\delta > 2$. This completes the proof. ∎

*Proof of Theorem* 1. Denote by

$$\mathrm{PML}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) = \log \hat{\mathbb{P}}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) - c|A|^{|W|}\log n\ ,$$

where $\hat{\mathbb{P}}(x_v^{(1:n)}|x_W^{(1:n)})$ is given by (10). If $V_n \setminus \{v\}$ is the bounding set for the candidate neighborhoods of vertex $v$ and $\mathrm{ne}(v)$ is the basic neighborhood of $v$, we need to prove that for any $c > 0$

$$\max_{W\subset V_n\setminus\{v\}, W\neq \mathrm{ne}(v)} \mathrm{PML}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) < \mathrm{PML}\left(x_v^{(1:n)}|x_{\mathrm{ne}(v)}^{(1:n)}\right) \tag{A9}$$

with probability converging to 1 when $n \to \infty$. Moreover, for $c > [p_{\min}(v)(|A|-1)]^{-1}$ we need to prove that (A9) holds eventually almost surely as $n \to \infty$. We divide the proof into two cases, showing that

$$\max_{W\in\mathcal{B}_i} \mathrm{PML}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) < \mathrm{PML}\left(x_v^{(1:n)}|x_{\mathrm{ne}(v)}^{(1:n)}\right) \tag{A10}$$

with high probability or almost surely as $n \to \infty$, depending on the value of $c$, for $i = 1, 2$ where

(a) $\mathcal{B}_1 = \{W \subset V_n \setminus \{v\} : \mathrm{ne}(v) \subsetneq W\}$
(b) $\mathcal{B}_2 = \{W \subset V_n \setminus \{v\} : \mathrm{ne}(v) \not\subset W\}$

For case (a), observe that for all $W \in \mathcal{B}_1$

$$\mathrm{PML}\left(x_v^{(1:n)}|x_{\mathrm{ne}(v)}^{(1:n)}\right) - \mathrm{PML}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) = \tag{A11}$$
$$c\left(|A|^{|W|} - |A|^{|\mathrm{ne}(v)|}\right)\log n\ - \sum_{a_v,a_W\in A^{|W|+1}} N(a_v,a_W)\log\frac{\hat{p}(a_v|a_W)}{\hat{p}(a_v|a_{\mathrm{ne}(v)})}\ .$$

As these empirical probabilities are the maximum likelihood estimators of the conditional probabilities and $\mathrm{ne}(v) \subset W$, we have that

$$\begin{aligned}\sum_{a_v,a_W\in A^{W+1}} N(a_v,a_W)\log\hat{p}(a_v|a_{\mathrm{ne}(v)}) &\geq \sum_{a_v,a_W\in A^{W+1}} N(a_v,a_W)\log p(a_v|a_{\mathrm{ne}(v)})\\ &= \sum_{a_v,a_W\in A^{W+1}} N(a_v,a_W)\log p(a_v|a_W)\ .\end{aligned}$$

Therefore, (A11) can be lower-bounded by

$$c\left(1-\frac{1}{|A|}\right)|A|^{|W|}\log n\ - \sum_{a_v,a_W\in A^{W+1}} N(a_v,a_W)\log\frac{\hat{p}(a_v|a_W)}{p(a_v|a_W)}. \tag{A12}$$

Note that

$$\sum_{a_v,a_W\in A^{W+1}} N(a_v,a_W)\log\frac{\hat{p}(a_v|a_W)}{p(a_v|a_W)} = \sum_{a_W\in A^W} N(a_W)D(\hat{p}(\cdot_v|a_W)\ ;\ p(\cdot_v|a_W))\ ,$$

where $D$ denotes the Kullback–Leibler divergence, see (12). Therefore we have, by Lemma 2, that

$$\sum_{a_W \in A^W} N(a_W) D(\hat{p}(\cdot_v|a_W) \; ; \; p(\cdot_v|a_W)) \leq \sum_{a_W \in A^W} N(a_W) \sum_{a_v \in A} \frac{\left[\hat{p}(a_v|a_W) - p(a_v|a_W)\right]^2}{p(a_v|a_W)}. \tag{A13}$$

Then, by Proposition 2 with $\delta > 0$ and (A13) we have, with probability converging to 1 that

$$\sup_{W \in \mathcal{B}_1} \sum_{a_W \in A^W} N(a_W) D(\hat{p}(\cdot_v|a_W) \; ; \; p(\cdot_v|a_W)) \leq \frac{\delta |A|^{|W|+1} \log n}{p_*}$$

and this holds eventually almost surely if $\delta > 1$. Then the difference (A12) can be lower bounded by

$$c\left(1 - \frac{1}{|A|}\right)|A|^{|W|} \log n - \frac{\delta |A|^{|W|+1} \log n}{p_*} > 0$$

if $\delta < c(|A| - 1)|A|^{-2} p_*$. Then, for any $c > 0$ there exists a sufficiently small $\delta > 0$ such that

$$\max_{W \in \mathcal{B}_1} \text{PML}(x_v^{(1:n)}|x_W^{(1:n)}) < \text{PML}(x_v^{(1:n)}|x_{\text{ne}(v)}^{(1:n)})$$

with probability converging to one. Moreover, if $c > |A|^2[p_*(|A| - 1)]^{-1}$ we can take $\delta > 1$ in Proposition 2, and we have that this inequality holds eventually almost surely as $n \to \infty$. This completes the proof of (A10) for case (a).

Finally, to prove (A10) in case (b) we will first prove that

$$\max_{W \in \mathcal{B}_2} \text{PML}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) \leq \text{PML}\left(x_v^{(1:n)}|x_{V_n \setminus \{v\}}^{(1:n)}\right)$$

eventually almost surely as $n \to \infty$. This inequality and case (a) will imply (A10) for $i = 2$. Note that we have

$$\begin{aligned}
&\text{PML}\left(x_v^{(1:n)}|x_{V_n \setminus \{v\}}^{(1:n)}\right) - \text{PML}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) \\
&\quad = \sum_{a_{V_n} \in A^{V_n}} N(a_v, a_{V_n \setminus \{v\}}) \log \frac{\hat{p}(a_v|a_{V_n \setminus \{v\}})}{\hat{p}(a_v|a_W)} - c\left(|A|^{|V_n|-1} - |A|^{|W|}\right) \log n \\
&\quad = n\left[\sum_{a_{V_n} \in A^{V_n}} \frac{N(a_{V_n})}{n} \log \frac{\hat{p}(a_v|a_{V_n \setminus \{v\}})}{\hat{p}(a_v|a_W)} - c\left(|A|^{|V_n|-1} - |A|^{|W|}\right) \frac{\log n}{n}\right].
\end{aligned}$$

Observe that for the second term in de brackets, we have

$$c\left(|A|^{|V_n|-1}-|A|^{|W|}\right)\frac{\log n}{n}\ \to\ 0$$

when $n\to\infty$, because we are assuming $|V_n|=o(\log n)$. For the first term by summing and subtracting $\frac{N(a_{V_n})}{n}\log p(a_v|a_W)$ inside the sum, we have that

$$\sum_{a_{V_n}}\frac{N(a_{V_n})}{n}\log\frac{\hat{p}(a_v|a_{V_n\setminus\{v\}})}{\hat{p}(a_v|a_W)}$$
$$=\sum_{a_{V_n}}\left[\frac{N(a_{V_n})}{n}\log\frac{\hat{p}(a_v|a_{V_n\setminus\{v\}})}{p(a_v|a_W)}-\frac{N(a_{V_n})}{n}\log\frac{\hat{p}(a_v|a_W)}{p(a_v|a_W)}\right]. \tag{A14}$$

We divide the expression again into two parts. On the one hand, by looking at the second term of the sum in (A14), we have that

$$\begin{aligned}\sum_{a_{V_n}}\frac{N(a_{V_n})}{n}\log\frac{\hat{p}(a_v|a_W)}{p(a_v|a_W)}&=\sum_{(a_v,a_W)\in A^{1+W}}\frac{N(a_v,a_W)}{n}\log\frac{\hat{p}(a_v|a_W)}{p(a_v|a_W)}\\&=\sum_{(a_v,a_W)\in A^{1+W}}\frac{N(a_W)}{n}\hat{p}(a_v|a_W)\log\frac{\hat{p}(a_v|a_W)}{p(a_v|a_W)}\\&=\sum_{a_W\in A^W}\frac{N(a_W)}{n}D(\hat{p}(\cdot_v|a_W)\ ;\ p(\cdot_v|a_W)).\end{aligned}$$

By Lemma 2 and Proposition 2 we have that

$$\begin{aligned}\max_{W\in\mathcal{B}_2}\sum_{a_W\in A^W}\frac{N(a_W)}{n}D(\hat{p}(\cdot_v|a_W)\ ;\ p(\cdot_v|a_W))&\le\sum_{a_W\in A^W}\frac{N(a_W)}{n}\sum_{a_v\in A}\frac{\left[\hat{p}(a_v|a_W)-p(a_v|a_W)\right]^2}{p(a_v|a_W)}\\&\le\max_{W\in\mathcal{B}_2}\frac{|A|^{|W|+1}\delta\log n}{p_*\,n}\ \to\ 0\end{aligned} \tag{A15}$$

eventually almost surely as $n\to\infty$, for $\delta>1$. On the other hand, as $\hat{p}(a_v|a_{V_n\setminus\{v\}})$ are the maximum likelihood estimators of $p(a_v|a_{V_n\setminus\{v\}})$ and as $V_n$ will eventually contain $\mathrm{ne}(v)$, the first term in the sum (A14) can be lower-bounded by

$$\sum_{a_{V_n}}\frac{N(a_{V_n})}{n}\log\frac{p(a_v|a_{V_n\setminus\{v\}})}{p(a_v|a_W)}=\sum_{a_v}\sum_{a_{W\cup\mathrm{ne}(v)}}\frac{N(a_v,a_{W\cup\mathrm{ne}(v)})}{n}\log\frac{p(a_v|a_{\mathrm{ne}(v)})}{p(a_v|a_W)}. \tag{A16}$$

By Proposition 3 we have that

$$\frac{N(a_v,a_{W\cup\mathrm{ne}(v)})}{n}\ge p(a_v,a_{W\cup\mathrm{ne}(v)})-\sqrt{\frac{3\log n}{n}}$$

eventually almost surely, simultaneously for all $W$ and all $(a_v, a_{W\cup \mathrm{ne}(v)})$). Then, (A16) can be lower bounded by

$$\sum_{a_{W\cup \mathrm{ne}(v)}} p(a_{W\cup \mathrm{ne}(v)}) D(p(\cdot_v|a_{\mathrm{ne}(v)}) \; ; \; p(\cdot_v|a_W)) - \sqrt{\frac{3\log n}{n}} \sum_{a_{W\cup \mathrm{ne}(v)}} \log \frac{p(a_v|a_{\mathrm{ne}(v)})}{p(a_v|a_W)} \geq \frac{\alpha_*}{2} \tag{A17}$$

eventually almost surely as $n \to \infty$. Therefore

$$\max_{W\in\mathcal{B}_2} \mathrm{PML}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) \leq \mathrm{PML}\left(x_v^{(1:n)}|x_{V_n\setminus\{v\}}^{(1:n)}\right)$$

eventually almost surely as $n \to \infty$. As $V_n$ will contain $\mathrm{ne}(v)$ for $n$ sufficiently large, then $V_n \setminus \{v\} \in \mathcal{B}_1$ for such values of $n$. Then, by case (a), we have

$$\mathrm{PML}\left(x_v^{(1:n)}|x_{V_n\setminus\{v\}}^{(1:n)}\right) \leq \max_{W\in\mathcal{B}_1} \mathrm{PML}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) < \mathrm{PML}\left(x_v^{(1:n)}|x_{\mathrm{ne}(v)}^{(1:n)}\right)$$

eventually almost surely as $n \to \infty$, which finishes the proof of case (b). By combining the results of the two cases, we conclude that

$$\max_{W\subset V_n\setminus\{v\}, W\neq \mathrm{ne}(v)} \mathrm{PML}\left(x_v^{(1:n)}|x_W^{(1:n)}\right) < \mathrm{PML}\left(x_v^{(1:n)}|x_{\mathrm{ne}(v)}^{(1:n)}\right)$$

and that $\widehat{\mathrm{ne}}(v) = \mathrm{ne}(v)$, with probability converging to 1 for all $c > 0$, or eventually almost surely as $n \to \infty$, for $c > |A|^2[p_*(|A|-1)]^{-1}$. ∎

*Proof of Corollary 1.* The proof of the corollary follows from Theorem 1 and the fact that $V'$ is finite. ∎