Stochastic Processes and their Applications 180 (2025) 104523

Contents lists available at ScienceDirect

Stochastic Processes and their Applications

journal homepage: www.elsevier.com/locate/spa

Model selection for Markov random fields on graphs under a mixing condition☆

Florencia Leonardi *,[1], Magno T.F. Severino[2]

Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão 1010, 05508-090, São Paulo, Brazil

ARTICLE INFO

Keywords:
Model selection
Regularized estimator
Structure estimation
Mixing processes

ABSTRACT

We propose a global model selection criterion to estimate the graph of conditional dependencies of a random vector. By global criterion, we mean optimizing a function over the set of possible graphs, eliminating the need to estimate individual neighborhoods and subsequently combine them to estimate the graph. We prove the almost sure convergence of the graph estimator. This convergence holds, provided the data is a realization of a multivariate stochastic process that satisfies a polynomial mixing condition. These are the first results to show the consistency of a model selection criterion for Markov random fields on graphs under non-independent data.

1. Introduction

We consider a vector-valued stochastic process, denoted as $\{X^{(i)} : i \in \mathbb{N}\}$, taking values in A^d, where A is a finite alphabet. We assume the process is stationary and ergodic, with invariant distribution π. We let G^* denote the graph encoding the conditional dependencies in π. Our primary goal in this work is to estimate G^* and the associated conditional probability distributions.

In the case the process $X^{(1)}, X^{(2)}, \ldots$ is independent and identically distributed, we have the classical model selection problem for discrete graphical models or Markov random fields on graphs. Extensive research has been conducted on these models, including, but not limited to, Lauritzen [1], Koller and Friedman [2], Lerasle and Takahashi [3], Pensar et al. [4], Divino et al. [5], Leonardi et al. [6]. Furthermore, these models have found applications in various fields, including Biology [7], Social Sciences [8] or Neuroscience [9]. Up to this moment, the most studied model has been the binary graphical model with pairwise interactions where structure estimation can be addressed by using standard logistic regression techniques [8,10], distance-based approaches between conditional probabilities [11,12] and maximization of the ℓ_1-penalized pseudo-likelihood [13,14]; see also Santhanam and Wainwright [15]. In the case of bigger discrete alphabets or general types of interactions, to our knowledge, the only works addressing the structure estimation problem are Loh and Wainwright [16] and Leonardi et al. [6]. In Loh and Wainwright [16], the authors obtain a characterization of the edges in the graph with the zeros of 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. In the recent work Leonardi et al. [6], a penalized criterion is proposed to estimate the neighborhood of each vertex, and the results are combined to construct the model's graph. Markov random fields on graphs have also been proposed for

☆ Dedicated to the memory of Antonio Galves, who introduced us to the statistical analysis of random processes and whose ideas inspired this work. This article was produced as part of the activities of the Research, Innovation and Dissemination Center for Neuromathematics, Brazil (grant FAPESP 2013/07699-0). It was also supported by the FAPESP, Brazil project "Stochastic Systems Modeling, Brazil" (grant 2023/13453-5) and CNPq, Brazil international cooperation project *"Probabilistic and statistical properties of stochastic processes on graphs and networks"* (grant 441884/2023-7).

* Corresponding author.
E-mail addresses: florencia@usp.br (F. Leonardi), magnotfs@usp.br (M.T.F. Severino).

[1] Partially supported by a CNPq's research fellowship, grant 311763/2020-0.
[2] Supported by CAPES and CNPq Ph.D. fellowships.

https://doi.org/10.1016/j.spa.2024.104523
Received 31 October 2023; Received in revised form 7 October 2024; Accepted 3 November 2024

Available online 14 November 2024

continuous random variables, where the structure estimation problem has been addressed by ℓ_1-regularization for Gaussian Markov random fields [17] and also extended to non-parametric models [18,19] and general conditional distributions from the exponential family [20].

From another perspective, graphical models can be seen as non-homogeneous versions of general random fields or Gibbs distributions on lattices, classical models in stochastic processes and statistical mechanics theory [21]. In such a setting, despite having only one observation within the sample, the number of variables increases. Given the regularity of the graph (each node has the same neighborhood), inference and model selection can be done based on the unique observation. The statistical inference for Markov random fields and Gibbs distributions under this setting has been addressed in Francis Comets' works [22,23]. More recently, model selection criteria, such as the BIC proposed by Schwarz [24], have been proven consistent under this regular setting [25,26]; see also Tjelmeland and Besag [27] and Löcherbach and Orlandi [28].

From an applied point of view, the assumption of independence of the observations in the non-homogeneous Markov random fields setting is often too restrictive. Consider, for example, the task of estimating interaction graphs from EEG time series data [29], river stream flow data [30] or daily stock market indices [6]. In these scenarios, the independence assumption does not hold, and the methods commonly used for graphical models serve only as approximations to the true underlying distribution. While such approximations can be practical from an applied point of view, from a theoretical perspective it is interesting to consider the problem of estimation and model selection in a dependence scenario, as for example the case of mixing processes considered here.

Conventional model selection techniques for graphical models often involve estimating the neighborhoods of individual nodes and constructing the graph based on these neighborhoods, as exemplified by Ravikumar et al. [10]. But depending on the rule to combine the neighborhoods, the final estimated graph can drastically underestimate or overestimate the set of edges in the graph [6]. In this work, we adopt a global estimation perspective that overcomes this limitation. Our approach involves estimating the graph by optimizing the penalized pseudo-likelihood function over the set of all possible simple and undirected graphs. We prove in Theorem 5 that the estimator almost surely converges to the true underlying graph in cases of finite graphical models, provided a polynomial mixing condition holds for the generating process. The proof of this consistency is based on results about the rate of convergence of empirical probabilities, stated in Proposition 3. These rates are obtained from the Law of the Iterated Logarithm for this type of process, derived in Oodaira and Yoshihara [31]. From a computational point of view, as the estimator involves a search through all potential graph configurations, even moderate to big sets of nodes make the optimization prohibited. In these cases, iterative algorithms such as simulated annealing or stepwise greedy algorithms could be employed to facilitate an efficient computation.

The paper is organized as follows. In Section 2, we provide essential definitions and notations concerning classical graphical models. Section 3 is dedicated to introducing the vector-valued mixing process and presenting important auxiliary theoretical results. In Section 4, we introduce the penalized maximum pseudo-likelihood estimator of the graph of conditional dependencies and state and prove the main consistency result of the paper.

2. Markov random fields on graphs

A *graph* is defined as an ordered pair $G = (V, E)$, where V represents the set of vertices (or nodes), and $E \subseteq V \times V$ is the set of edges connecting pairs of vertices. We refer to a graph as *undirected* if $(v_i, v_j) \in E$ implies that $(v_j, v_i) \in E$ for all $(v_i, v_j) \in E$, where $v_i, v_j \in V$. Furthermore, a graph is considered *simple* if $(v, v) \notin E$ for all $v \in V$. For the purposes of this work, we concentrate exclusively on undirected simple graphs, which we will henceforth simply call a *graph*.

Consider a graph $G = (V, E)$, with $V = \{1, \dots, d\}$, for $d \in \mathbb{N}$, and assume we observe at each vertex $v \in V$ a random variable X_v, which is discrete and takes values in A, a finite alphabet. Moreover, let $X = (X_1, \dots, X_d)$ be the vector of all variables observed on the vertices of the graph. Denote by $\mathbb{P}$ the joint probability distribution of the vector X. For any $W \subset V$ and any configuration $a_W \in A^W$ we write

$$\pi(a_W) = \mathbb{P}(X_W = a_W).$$

Moreover, if $\pi(a_W) > 0$, for $a_U \in A^U$ we denote by

$$\pi(a_U | a_W) = \mathbb{P}(X_U = a_U | X_W = a_W)$$

the corresponding conditional probability.

Given a given vertex $v \in V$, any set $W \subset V$, with $v \notin W$, is called a *neighborhood* of v. Furthermore, a neighborhood W of v is called *Markov neighborhood* if

$$\pi(a_v | a_U) = \pi(a_v | a_W)$$

for all $U \supset W$, $v \notin U$ and all $a_U \in A^U$, with $\pi(a_U) > 0$. The definition of a Markov neighborhood W of v is equivalent to request that for all $U' \subset V \setminus \{v\}$ such that $U' \cap W = \emptyset$, $X_{U'}$ is conditionally independent of X_v, given X_W. That is,

$$X_v \perp\!\!\!\perp X_{U'} | X_W$$

for all U' with $U' \cap W = \emptyset$, where $\perp\!\!\!\perp$ is the usual symbol denoting independence of random variables.

As discussed in Leonardi et al. [6], if W is a Markov neighborhood of $v \in V$, then any finite set $U \supset W$ is also a Markov neighborhood of v. In contrast, W_1 and W_2 being Markov neighborhoods of v does not imply in general that $W_1 \cap W_2$ is a Markov neighborhood of v, however this property is satisfied by some probability measures. This fact leads to the following definition.

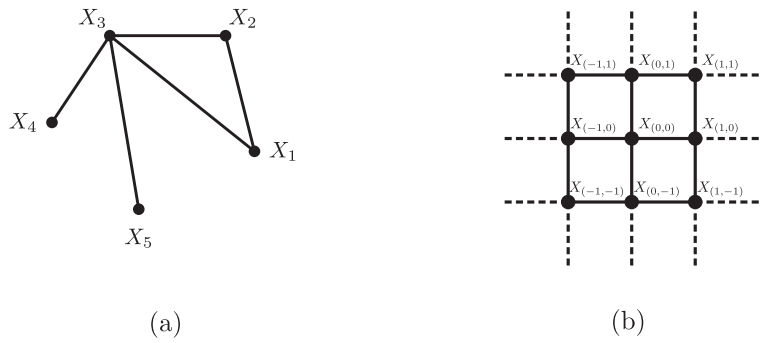

Fig. 1. Different graph structures for Markov random fields under different settings. The figure on the left is a finite graphical model or Markov random field on a general graph, and the figure on the right represents the interaction graph in a classical Markov random field or Gibbs distribution on a regular lattice.

Definition 1 (*Markov Intersection Property*)**.** We say that Markov intersection property holds if 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 property is desirable in this context to define the smallest Markov neighborhood of a node. This property is guaranteed under the usually assumed positivity condition (see Lauritzen [1]). But the positivity assumption is not necessary to obtain consistent estimators, and then it is enough to assume the Markov intersection property, see Leonardi et al. [6] for details.

Definition 2 (*Basic Neighborhood*)**.** For $v \in V$, let $\mathcal{W}(v)$ be the set of all subsets of V that are Markov neighborhoods of v. The *basic neighborhood* of v is defined as

$$\mathrm{ne}(v) = \bigcap_{W \in \mathcal{W}(v)} W. \tag{1}$$

By the Markov intersection property, $\mathrm{ne}(v)$ is the smallest Markov neighborhood of $v \in V$. Based on these basic neighborhoods, define the graph $G^* = (V, E^*)$ as

$$(v, w) \in E^* \text{ if and only if } w \in \mathrm{ne}(v), \tag{2}$$

where $E^* \subseteq V \times V$. The graph G^* with edges defined in (2) is *undirected*, as proved by Leonardi et al. [6]. Fig. 1 shows two examples of graphs for Markov random fields under different settings: the finite non-homogeneous graphical model case in (a) and the interaction graph in a classical Markov random field, or Gibbs distribution, on a regular lattice [23,26] in (b).

3. Vector-valued mixing processes

In this paper we consider a stationary and ergodic vector-valued stochastic process $X^{(1)}, X^{(2)}, \ldots$, where each variable $X^{(i)}$ is a vector of d components, belonging to the set A^d, with A a finite alphabet. We denote by $\left((A^d)^{\mathbb{N}}, \mathcal{F}, \mathbb{P}\right)$ the probability space for the process $\{X^{(i)} : i \in \mathbb{N}\}$ and by π its stationary distribution. Sometimes we need to consider "slices" of the entire realization $X^{(1)}, \ldots, X^{(n)}$ on both dimensions. To avoid misleading notations, we use superscripts to denote the indexes in "time" (ranging from 1 to n) and subscripts to denote indexes on "space" (a subset of $V = \{1, \ldots, d\}$). For any set $U \subset V$ and any integer interval $i : j$ we denote by $X_U^{(i:j)}$ the sequence $X_U^{(i)}, \ldots, X_U^{(j)}$ with $X_U^{(k)} = (X_u^{(k)} : u \in U)$, $k = i, \ldots, j$. When $U = V$ we avoid the subscript and simply write $X^{(i:j)}$. The same notation is used for "realizations" of the process, denoted in lower case $x_U^{(i:j)}$ instead of the notation for the random variables $X_U^{(i:j)}$. See an example in Fig. 2.

We say the process $\{X^{(i)} : i \in \mathbb{N}\}$ satisfies a mixing condition with rate $\{\psi(\ell)\}_{\ell \in \mathbb{N}}$, where $\psi(\ell) \to 0$ as $\ell \to \infty$, if for each k, m and each $x^{(1:k)}$, $x^{(1:m)}$ with $\mathbb{P}(X^{(1:m)} = x^{(1:m)}) > 0$ we have that

$$\begin{aligned} \left| \mathbb{P}(X^{(n:(n+k-1))} = x^{(1:k)} \mid X^{(1:m)} = x^{(1:m)}) - \mathbb{P}(X^{(n:(n+k-1))} = x^{(1:k)}) \right| \\ \leq \psi(n-m)\, \mathbb{P}(X^{(n:(n+k-1))} = x^{(1:k)}) \end{aligned} \tag{3}$$

for $n > m$.

Assume we observe a sample of size n of the process, denoted by $\{x^{(i)} : i = 1, \ldots, n\}$. Since the stationary distribution of the process π is not known, we must estimate it from the data. For any $W \subset V$ and any $a_W \in A^W$ denote by

$$\hat{\pi}(a_W) = \frac{N(a_W)}{n},$$

where $N(a_W)$ denotes the number of times the configuration a_W appears in the sample $x^{(1)}, \ldots, x^{(n)}$. If $\hat{\pi}(a_W) > 0$, we can also define the conditional probabilities

$$\hat{\pi}(a_W | a_{W'}) = \frac{\hat{\pi}(a_{W \cup W'})}{\hat{\pi}(a_{W'})}, \tag{4}$$

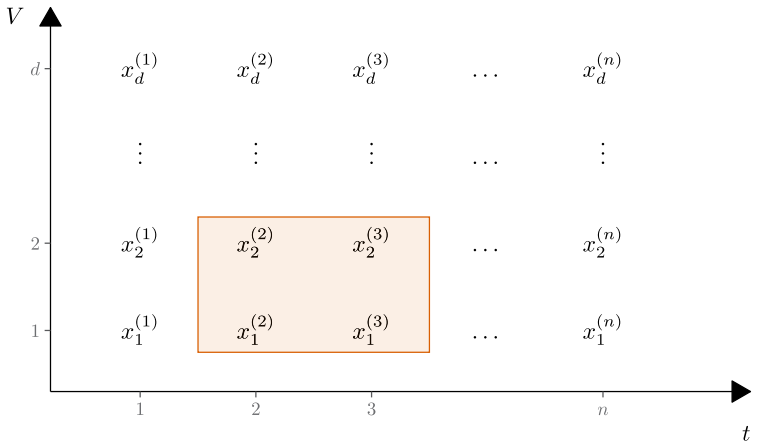

Fig. 2. Representation of a realization of the process **X**, with set of vertices $V = \{1, \dots, d\}$ observed from time 1 to n. Subscript indicates the vertex and superscript indicates the time at which the observation was taken. The highlighted rectangle indicates the observed slice $x^{(2:3)}_{\{1,2\}}$.

for two disjoint subsets $W, W' \subset V$ and configurations $a_W \in A^W, a_{W'} \in A^{W'}$.

Based on the Law of the Iterated Logarithm for stationary polynomial mixing processes proved in Oodaira and Yoshihara [31], we can derive the rate of convergence of the empirical probabilities to the true probabilities of the process. From now on, the phrase *eventually almost surely* means with probability one, for all n large enough.

Proposition 3. *Assume the process $\{X^{(i)} : i \in \mathbb{N}\}$ is stationary and satisfies the mixing condition (3) with mixing rate $\psi(\ell) = O(1/\ell^{1+\epsilon})$, for some $\epsilon > 0$. Then for any $\delta > 0$, any $W \subset V$ and any $a_W \in A^W$ we have that*

$$|\widehat{\pi}(a_W) - \pi(a_W)| < \sqrt{\frac{\delta \log n}{n}}$$

eventually almost surely as $n \to \infty$. Moreover, for any disjoint sets $W, W' \subset V$ and any $a_W \in A^W$, $a_{W'} \in A^{W'}$ we have that

$$\left|\widehat{\pi}(a_W | a_{W'}) - \pi(a_W | a_{W'})\right| < \sqrt{\frac{\delta \log n}{N(a_{W'})}},$$

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

The proof of Proposition 3 is postponed to the Appendix.

Remark 4. For simplicity in the presentation, in this paper, we assume stationarity of the process $\{X^{(i)} : i \in \mathbb{N}\}$. This hypothesis is also required in Oodaira and Yoshihara [31], whose results we use to prove Proposition 3. But the stationarity hypothesis could be relaxed provided we can obtain sufficient rates of convergence of the empirical probabilities to their limits, as proved in Proposition 3. These rates can also be obtained for non-stationary processes, see for example the results in Csiszár [32] where an exponential mixing rate is assumed.

4. The graph's estimator and its consistency

In this paper we take a regularized pseudo maximum likelihood approach to estimate the graph G^*, given a sample $x^{(1)}, \dots, x^{(n)}$ of the stochastic process. Instead of estimating each neighborhood and then combining the results, as is proposed in several works, we globally estimate the graph G^* by optimizing a function over the set of all simple graphs over V.

Given any graph G defined on the same set of vertices V, the pseudo-likelihood function is defined by

$$L(G) = \prod_{i=1}^{n} \prod_{v \in V} \pi(x_v^{(i)} | x_{G(v)}^{(i)}),$$

where $G(v)$ denotes the neighborhood of node v in the graph G, that is $G(v) = \{u \in V : (u, v) \in E\}$. As the conditional probabilities of π are not known, we can estimate them from the data, obtaining the maximum pseudo-likelihood given by

$$\widehat{L}(G) = \prod_{i=1}^{n} \prod_{v \in V} \widehat{\pi}(x_v^{(i)} | x_{G(v)}^{(i)}),$$

with $\widehat{\pi}(x_v^{(i)} | x_{G(v)}^{(i)})$ defined as in (4), taking $W = \{v\}$ and $W' = G(v)$.

Applying the logarithm and taking into account the number of occurrences of each configuration in the sample, we can write the log pseudo likelihood function as

$$\log \widehat{L}(G) = \sum_{v\in V} \sum_{a_v, a_{G(v)}} \widehat{\pi}(a_v|a_{G(v)})^{N(a_v, a_{G(v)})}, \tag{5}$$

where the sum is taken over all $v \in V$ and all configurations $a_v \in A$, $a_{G(v)} \in A^{G(v)}$ such that $N(a_v, a_{G(v)}) > 0$.

We then define the graph estimator by

$$\widehat{G} = \arg\max_G \Big\{ \log \widehat{L}(G) - \lambda_n \sum_{v\in V} |A|^{|G(v)|} \Big\}, \tag{6}$$

with $|G(v)|$ denoting the cardinal of the set $G(v)$ and λ_n being a non-negative decreasing sequence.

The main result in this paper is the following consistency result for the graph estimator $\widehat{G}$.

Theorem 5. *Assume the process $\{X^{(i)} : i \in \mathbb{N}\}$ is stationary and satisfies the mixing condition (3) with rate $\psi(\ell) = O(1/\ell^{1+\epsilon})$ for some $\epsilon > 0$. Then, taking $\lambda_n = c \log n$, with $c > 0$, we have that $\widehat{G}$ defined in (6) satisfies $\widehat{G} = G^*$ eventually almost surely when $n \to \infty$.*

Before proving Theorem 5, we recall the definition of the Küllback–Leibler divergence between two probability distributions p and q over A. It is given by

$$D(p; q) = \sum_{a\in A} p(a) \log \frac{p(a)}{q(a)} \tag{7}$$

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 Küllback–Leibler divergence is that $D(p; q) = 0$ if and only if $p(a) = q(a)$ for all $a \in A$.

Denote by $G_{\max}$ the complete graph over V, that is $G_{\max} = (V, E_{\max})$ with

$$E_{\max} = \{(u, v) \in V \times V : u \neq v\}.$$

Observe that we have $G_{\max}(v) = V \setminus \{v\}$ for all $v \in V$. For any $v \in V$, denote by $\mathcal{S}(v)$ the set of neighborhoods of v not containing $G^*(v)$, that is

$$\mathcal{S}(v) = \{S \subset V \setminus \{v\} : G^*(v) \not\subset S\},$$

where $G^*(v) \not\subset S$ means S does not contain $G^*(v)$. Over this set, we take the minimal "distance", measured as the mean Kullback–Leibler divergence, between the transition probabilities given $G^*(v)$ and given $S \in \mathcal{S}(v)$, defined as

$$\alpha(v) = \min_{S\in\mathcal{S}(v)} \left\{ \sum_{a_{G_{\max}(v)}} \pi(a_{G_{\max}(v)}) D\big(\pi(\cdot_v|a_{G^*(v)}); \pi(\cdot_v|a_S)\big) \right\}. \tag{8}$$

By Definition 2 we must have $\alpha(v) > 0$, for a proof see Leonardi et al. [6].

Proof of Theorem 5. First observe that we can decompose the event $\{\widehat{G} \neq G^*\}$ as the union of the two events $\{G^* \subsetneq \widehat{G}\} \cup \{G^* \not\subset \widehat{G}\}$. The event $\{G^* \subsetneq \widehat{G}\}$ represents overfitting, occurring when $\widehat{G}$ is a strict superset of G^*; that is, $\widehat{G}$ has at least one more edge than G^*. On the other hand, the event $\{G^* \not\subset \widehat{G}\}$ represents underfitting, where $\widehat{G}$ has at least one fewer edge than G^*. We will consider these two events separately, proving that eventually almost surely as $n \to \infty$ neither of them can happen, implying that $\widehat{G} = G^*$.

Let first prove that $\mathbb{P}(G^* \subsetneq \widehat{G}) = 0$ e.a.s as $n \to \infty$ (non-overfitting). To prove that this event cannot happen eventually almost surely, we will prove that for all graphs $G \supsetneq G^*$ we have that

$$\log \widehat{L}(G) - \lambda_n \sum_{v\in V} |A|^{|G(v)|} < \log \widehat{L}(G^*) - \lambda_n \sum_{v\in V} |A|^{|G^*(v)|} \tag{9}$$

or equivalently that

$$\log \widehat{L}(G) - \log \widehat{L}(G^*) < \lambda_n \Big(\sum_{v\in V} |A|^{|G(v)|} - \sum_{v\in V} |A|^{|G^*(v)|} \Big) \tag{10}$$

eventually almost surely as $n \to \infty$, proving that $\widehat{G} \neq G$ for all $G \supsetneq G^*$. Observe that

$$\log \widehat{L}(G) = \sum_{v\in V} \sum_{a_v, a_{G(v)}} N(a_v, a_{G(v)}) \log \widehat{\pi}(a_v|a_{G(v)})$$

and similarly for $\log \widehat{L}(G^*)$. Then

$$\begin{aligned} \log \widehat{L}(G) - \log \widehat{L}(G^*) \;=\; & \sum_{v\in V} \sum_{a_v, a_{G(v)}} N(a_v, a_{G(v)}) \log \widehat{\pi}(a_v|a_{G(v)}) \\ & - \sum_{v\in V} \sum_{a_v, a_{G^*(v)}} N(a_v, a_{G^*(v)}) \log \widehat{\pi}(a_v|a_{G^*(v)}). \end{aligned} \tag{11}$$

Fix $v \in V$. By the definition of the maximum likelihood estimators and as $G \supset G^*$ we have that

$$\begin{aligned} \sum_{a_v, a_{G^*(v)}} N(a_v, a_{G^*(v)}) \log \widehat{\pi}(a_v|a_{G^*(v)}) \;\geq\; & \sum_{a_v, a_{G^*(v)}} N(a_v, a_{G^*(v)}) \log \pi(a_v|a_{G^*(v)}) \\ = \; & \sum_{a_v, a_{G(v)}} N(a_v, a_{G(v)}) \log \pi(a_v|a_{G^*(v)}). \end{aligned} \tag{12}$$

Therefore, using (12), the difference in (11) can be upper-bounded by

$$\sum_{v\in V}\sum_{a_v,a_{G(v)}} N(a_v,a_{G(v)})\log\frac{\widehat{\pi}(a_v|a_{G(v)})}{\pi(a_v|a_{G(v)})} = \sum_{v\in V}\sum_{a_{G(v)}} N(a_{G(v)})D(\widehat{\pi}(\cdot_v|a_{G(v)});\pi(\cdot_v|a_{G(v)})),$$

where D denotes the Küllback–Leibler divergence, see (7). Therefore we have, by Lemma 10, that

$$\begin{aligned}&\sum_{v\in V}\sum_{a_{G(v)}} N(a_{G(v)})D(\widehat{\pi}(\cdot_v|a_{G(v)});\pi(\cdot_v|a_{G(v)}))\\ &\quad\le \sum_{v\in V}\sum_{a_{G(v)}} N(a_{G(v)})\sum_{a_v\in A}\frac{[\,\widehat{\pi}(a_v|a_{G(v)})-\pi(a_v|a_{G(v)})\,]^2}{\pi(a_v|a_{G(v)})}.\end{aligned} \tag{13}$$

Then, by Proposition 3 with $\delta>0$ and (13) we have, with probability 1, for n sufficiently large that

$$\sum_{v\in V}\sum_{a_{G(v)}} N(a_{G(v)})\,D(\widehat{\pi}(\cdot_v|a_{G(v)});\pi(\cdot_v|a_{G(v)}))\le\frac{\delta|A|\log n}{\pi_{\min}}\sum_{v\in V}|A|^{|G(v)|},$$

where

$$\pi_{\min}=\min_{v\in V}\left\{\pi(a_v|a_{G(v)}):\pi(a_v|a_{G(v)})>0,a_v\in A,a_{G(v)}\in A^{G(v)}\right\}.$$

On the other hand, we have that

$$\begin{aligned}\sum_{v\in V}|A|^{|G(v)|}-\sum_{v\in V}|A|^{|G^*(v)|} &= \sum_{v\in V}|A|^{|G(v)|}\left(1-\frac{\sum_{v\in V}|A|^{|G^*(v)|}}{\sum_{v\in V}|A|^{|G(v)|}}\right)\\ &\ge c'\sum_{v\in V}|A|^{G(v)},\end{aligned} \tag{14}$$

for $0<c'<1$. This inequality holds because

$$\frac{\sum_{v\in V}|A|^{|G^*(v)|}}{\sum_{v\in V}|A|^{|G(v)|}}<1.$$

Thus if we chose δ such that

$$\frac{\delta|A|}{c'\pi_{\min}}<c,$$

that is, we take $\delta<cc'\pi_{\min}\frac{1}{|A|}$, we obtain that

$$\max_{G\supsetneq G^*}\{\log\widehat{L}(G)-\lambda_n\sum_{v\in V}|A|^{|G(v)|}\}<\log\widehat{L}(G^*)-\lambda_n\sum_{v\in V}|A|^{|G^*(v)|}$$

with probability 1 for n sufficiently large. This concludes the proof for the event $\{G^*\subsetneq\widehat{G}\}$.

We now prove that $\mathbb{P}(G^*\not\subset\widehat{G})=0$ e.a.s as $n\to\infty$ (non-underfitting). In order to prove this case we need to show that for any graph G, such that $G\not\supset G^*$ we have that

$$\log\widehat{L}(G)-\lambda_n\sum_{v\in V}|A|^{G(v)}<\log\widehat{L}(G^*)-\lambda_n\sum_{v\in V}|A|^{G^*(v)} \tag{15}$$

eventually almost surely as $n\to\infty$. In order to prove that (15) holds, first we prove that

$$\log\widehat{L}(G)-\lambda_n\sum_{v\in V}|A|^{G(v)}<\log\widehat{L}(G_{\max})-\lambda_n\sum_{v\in V}|A|^{G_{\max}(v)},$$

with $G_{\max}$ denoting the complete graph in V. Then, this inequality and (9) proved for the case $G^*\subsetneq\widehat{G}$ show that

$$\log\widehat{L}(G_{\max})-\lambda_n\sum_{v\in V}|A|^{G_{\max}(v)}<\log\widehat{L}(G^*)-\lambda_n\sum_{v\in V}|A|^{G^*(v)}$$

eventually almost surely as $n\to\infty$, what implies the desired result.

Note that we have

$$\begin{aligned}\log \hat{L}(G_{\max}) - \lambda_n \sum_{v\in V} |A|^{G_{\max}(v)} - \log \hat{L}(G) + \lambda_n \sum_{v\in V} |A|^{G(v)} \\ = \sum_{v\in V} \sum_{a_v, a_{G_{\max}(v)}} N(a_v, a_{G_{\max}(v)}) \log \frac{\hat{\pi}(a_v|a_{G_{\max}(v)})}{\hat{\pi}(a_v|a_{G(v)})} \\ - \lambda_n \sum_{v\in V} \left(|A|^{G_{\max}(v)} - |A|^{G(v)}\right) \\ = n\Big[\sum_{v\in V} \sum_{a_v, a_{G_{\max}(v)}} \frac{N(a_v, a_{G_{\max}(v)})}{n} \log \frac{\hat{\pi}(a_v|a_{G_{\max}(v)})}{\hat{\pi}(a_v|a_{G(v)})} \\ - \frac{\lambda_n}{n} \sum_{v\in V} \left(|A|^{G_{\max}(v)} - |A|^{G(v)}\right)\Big].\end{aligned} \tag{16}$$

One can see that for $\lambda_n = c \log n$, with $c > 0$, the second term in the brackets in the last expression of (16) vanishes when $n \to \infty$, i.e.,

$$\frac{\lambda_n}{n} \sum_{v\in V} \left(|A|^{G_{\max}(v)} - |A|^{G(v)}\right) \xrightarrow[n\to\infty]{} 0.$$

Now, by adding

$$\frac{N(a_v, a_{G_{\max}(v)})}{n} \log \frac{\pi(a_v|a_{G(v)})}{\pi(a_v|a_{G(v)})} = 0$$

into the first term of the sum in (16), we can write it as

$$\begin{aligned}\sum_{v\in V} \sum_{a_v, a_{G_{\max}(v)}} \frac{N(a_v, a_{G_{\max}(v)})}{n} \log \frac{\hat{\pi}(a_v|a_{G_{\max}(v)})}{\hat{\pi}(a_v|a_{G(v)})} \\ = \sum_{v\in V} \sum_{a_v, a_{G_{\max}(v)}} \frac{N(a_v, a_{G_{\max}(v)})}{n} \left(\log \frac{\hat{\pi}(a_v|a_{G_{\max}(v)})}{\hat{\pi}(a_v|a_{G(v)})} + \log \frac{\pi(a_v|a_{G(v)})}{\pi(a_v|a_{G(v)})}\right) \\ = \sum_{v\in V} \sum_{a_v, a_{G_{\max}(v)}} \frac{N(a_v, a_{G_{\max}(v)})}{n} \left(\log \frac{\hat{\pi}(a_v|a_{G_{\max}(v)})}{\pi(a_v|a_{G(v)})} - \log \frac{\hat{\pi}(a_v|a_{G(v)})}{\pi(a_v|a_{G(v)})}\right) \\ = \underbrace{\sum_{v\in V} \sum_{a_v, a_{G_{\max}(v)}} \frac{N(a_v, a_{G_{\max}(v)})}{n} \log \frac{\hat{\pi}(a_v|a_{G_{\max}(v)})}{\pi(a_v|a_{G(v)})}}_{(1)} \\ \underbrace{- \sum_{v\in V} \sum_{a_v, a_{G_{\max}(v)}} \frac{N(a_v, a_{G_{\max}(v)})}{n} \log \frac{\hat{\pi}(a_v|a_{G(v)})}{\pi(a_v|a_{G(v)})}}_{(2)}.\end{aligned} \tag{17}$$

As highlighted in (17), we analyze this expression in two parts. The second term in the right-hand side of (17) can be written as

$$\begin{aligned}\sum_{v\in V} \sum_{a_v, a_{G(v)}} \frac{N(a_{G(v)})}{n} \hat{\pi}(a_v|a_{G(v)}) \log \frac{\hat{\pi}(a_v|a_{G(v)})}{\pi(a_v|a_{G(v)})} \\ = \sum_{v\in V} \sum_{a_{G(v)}} \frac{N(a_{G(v)})}{n} \sum_{a_v} \hat{\pi}(a_v|a_{G(v)}) \log \frac{\hat{\pi}(a_v|a_{G(v)})}{\pi(a_v|a_{G(v)})} \\ = \sum_{v\in V} \sum_{a_{G(v)}} \frac{N(a_{G(v)})}{n} D\big(\hat{\pi}(\cdot_v|a_{G(v)}); \pi(\cdot_v|a_{G(v)})\big).\end{aligned}$$

Observe that, by Lemma 10 and Proposition 3, for $\delta > 0$, we have

$$\begin{aligned}\sum_{v\in V} \sum_{a_{G(v)}} \frac{N(a_{G(v)})}{n} D(\hat{\pi}(\cdot_v|a_{G(v)}); \pi(\cdot_v|a_{G(v)})) \\ \le \sum_{v\in V} \sum_{a_{G(v)}} \frac{N(a_{G(v)})}{n} \sum_{a_v\in A} \frac{[\,\hat{\pi}(a_v|a_{G(v)}) - \pi(a_v|a_{G(v)})\,]^2}{\pi(a_v|a_{G(v)})} \\ \le \sum_{v\in V} \sum_{a_{G(v)}} \frac{N(a_{G(v)})}{n} \sum_{a_v\in A} \frac{\delta \log n}{N(a_{G(v)})\pi(a_v|a_{G(v)})} \\ \le |V||A|^{|V|} \frac{\delta}{\pi_{\min}} \frac{\log n}{n} \longrightarrow 0\end{aligned}$$

eventually almost surely as $n \to \infty$. On the other hand, since $\hat{\pi}(a_v|a_{G_{\max}(v)})$ are the maximum likelihood estimators of $\pi(a_v|a_{G_{\max}(v)})$ and $G^* \subseteq G_{\max}$, the first term in the right-hand side of (17) can be lower-bounded by

$$\begin{aligned}
&\sum_{v\in V}\sum_{a_v,a_{G_{\max}(v)}} \frac{N(a_v,a_{G_{\max}(v)})}{n}\log\frac{\hat{\pi}(a_v|a_{G_{\max}(v)})}{\pi(a_v|a_{G(v)})}\\
&\qquad\geq \sum_{v\in V}\sum_{a_v,a_{G_{\max}(v)}} \frac{N(a_v,a_{G_{\max}(v)})}{n}\log\frac{\pi(a_v|a_{G_{\max}(v)})}{\pi(a_v|a_{G(v)})}\\
&\qquad= \sum_{v\in V}\sum_{a_v,a_{G_{\max}(v)}} \frac{N(a_v,a_{G_{\max}(v)})}{n}\log\frac{\pi(a_v|a_{G^*(v)})}{\pi(a_v|a_{G(v)})}
\end{aligned} \tag{18}$$

By Proposition 3

$$\frac{N(a_v,a_{G_{\max}(v)})}{n} \;=\; \hat{\pi}(a_v,a_{G_{\max}(v)}) > \pi(a_v,a_{G_{\max}(v)}) - \sqrt{\frac{\delta\log n}{n}},$$

eventually almost surely as $n \to \infty$. Then, one can see that (18) can be lower-bounded by

$$\begin{aligned}
&\sum_{v\in V}\sum_{a_v,a_{G_{\max}(v)}} \left[\pi(a_v,a_{G_{\max}(v)}) - \sqrt{\frac{\delta\log n}{n}}\right]\log\frac{\pi(a_v|a_{G^*(v)})}{\pi(a_v|a_{G(v)})}\\
&\qquad= \sum_{v\in V}\sum_{a_{G_{\max}(v)}}\pi(a_{G_{\max}(v)})\sum_{a_v}\pi(a_v|a_{G_{\max}(v)})\log\frac{\pi(a_v|a_{G^*(v)})}{\pi(a_v|a_{G(v)})}\\
&\qquad\qquad - \sqrt{\frac{\delta\log n}{n}}\sum_{v\in V}\sum_{a_v,a_{G_{\max}(v)}}\log\frac{\pi(a_v|a_{G^*(v)})}{\pi(a_v|a_{G(v)})}\\
&\qquad\geq \sum_{v\in V}\left(\sum_{a_{G_{\max}(v)}}\pi(a_{G_{\max}(v)})D\big(\pi(\cdot_v|a_{G^*(v)});\pi(\cdot_v|a_{G(v)})\big)\right.\\
&\qquad\qquad \left. + \sqrt{\frac{\delta\log n}{n}}|A|^{|V|}\log\pi_{\min}\right)\\
&\qquad\geq \frac{1}{2}\sum_{v\in V}\alpha(v) > 0
\end{aligned} \tag{19}$$

eventually as $n \to \infty$, since $\alpha(v) > 0$ for all v. Therefore, we obtain from (16) and the difference in (17) that

$$\log\hat{L}(G) - \lambda_n\sum_{v\in V}|A|^{G(v)} \;<\; \log\hat{L}(G_{\max}) - \lambda_n\sum_{v\in V}|A|^{G_{\max}(v)},$$

eventually almost surely as $n \to \infty$. Now, since $G^* \subseteq G_{\max}$, if $G^* \neq G_{\max}$ we have by the arguments used in the proof of non-overfitting that

$$\log\hat{L}(G_{\max}) - \lambda_n\sum_{v\in V}|A|^{G_{\max}(v)} \;\leq\; \log\hat{L}(G^*) - \lambda_n\sum_{v\in V}|A|^{G^*(v)}$$

eventually almost surely as $n \to \infty$, and this shows that $\mathbb{P}(G^* \not\subset \widehat{G}) = 0$ eventually almost surely as $n \to \infty$. Combining the two cases, as $\{\widehat{G} \neq G^*\} = \{G^* \subsetneq \widehat{G}\} \cup \{G^* \not\subset \widehat{G}\}$ we obtain that $\mathbb{P}(G^* \neq \widehat{G}) = 0$ eventually almost surely as $n \to \infty$ and this concludes the proof of Theorem 5. □

Discussion

In this paper, we introduced a model selection approach to estimate the underlying graph of conditional dependencies in a multivariate stochastic process. Our method relies on a penalized pseudo-likelihood and employs a global estimation approach. We have established the almost sure convergence of this estimator to the true underlying graph, specifically in the context of finite graphical models, provided a certain mixing condition is satisfied. While the case of independent and identically distributed processes has been extensively explored in the literature, this assumption often proves too restrictive for real-world applications where independence does not hold.

Our approach distinguishes itself by considering the estimation of the entire graph at once, diverging from the usual practice found in the literature, which typically estimates individual neighborhoods for each vertex and subsequently combines them to form the graph. In practical terms, computing the proposed estimator poses a significant computational challenge, as it involves searching through all potential graph configurations. Exploring efficient algorithms to approximate the estimator and derive theoretical properties remains an important problem for future work. Moreover, the estimator's definition is based on a penalization constant that must be stated before the analysis. The choice of this constant is also a challenging task in regularized approaches and could be addressed with methods like cross-validation.

There are also other promising directions for extending this work. One such direction involves adapting the theoretical framework to accommodate continuous multivariate stochastic processes, thereby broadening the range of potential applications of our methodology. Another line of research is the generalization to infinite vertex sets and unbounded estimators, where the size of the estimated graph is allowed to grow with the sample size.

Acknowledgments

We thank the anonymous reviewer for the excellent comments and suggestions, which helped us to significantly improve the manuscript.

Appendix

In this section, we state and prove auxiliary results that were used throughout the paper.

Lemma 6. *If the process $\{X^{(i)} : i \in \mathbb{N}\}$ satisfies the mixing condition (3) with rate $\{\psi(\ell)\}_{\ell\in\mathbb{N}}$, then $\{X_W^{(i)} : i \in \mathbb{N}\}$ with $W \subset V$ also satisfies the mixing condition with rate $\{\psi(\ell)\}_{\ell\in\mathbb{N}}$.*

Proof. Observe that for any $x_W^{(1:k)}, x_W^{(1:m)} \in A^W$ we have that

$$\begin{aligned}
&\left|\mathbb{P}(X_W^{(n:(n+k-1))} = x_W^{(1:k)} \mid X_W^{(1:m)} = x_W^{(1:m)}) - \mathbb{P}(X_W^{(n:(n+k-1))} = x_W^{(1:k)})\right| \\
&= \Big| \sum_{x_{W^c}^{(1:k)}} \mathbb{P}(X_{W\cup W^c}^{(n:(n+k-1))} = x_{W\cup W^c}^{(1:k)} \mid X_W^{(1:m)} = x_W^{(1:m)}) \\
&\qquad\qquad - \sum_{x_{W^c}^{(1:k)}} \mathbb{P}(X_{W\cup W^c}^{(n:(n+k-1))} = x_{W\cup W^c}^{(1:k)})\Big| \\
&\le \sum_{x_{W^c}^{(1:k)}} \psi(n-m)\mathbb{P}(X_{W\cup W^c}^{(n:(n+k-1))} = x_{W\cup W^c}^{(1:k)}) \\
&\le \psi(n-m)\mathbb{P}(X_W^{(n:(n+k-1))} = x_W^{(1:k)}).
\end{aligned}$$

Then the process $\{X_W^{(i)} : i \in \mathbb{N}\}$ with $W \subset \{1,\dots,d\}$ is mixing with rate $\psi(\ell)$. □

Lemma 7. *Let the process $\{X^{(i)} : i \in \mathbb{N}\}$ satisfy the mixing condition (3) with rate $\{\psi(\ell)\}_{\ell\in\mathbb{N}}$ and let $f : A^d \to \mathbb{R}$ be a function. Then, the process $\{f(X^{(i)}) : i \in \mathbb{N}\}$ is also mixing with rate $\{\psi(\ell)\}_{\ell\in\mathbb{N}}$.*

Proof. Denote by $Y^{(i)} = f(X^{(i)})$ and let $\mathbb{P}_Y$ denote the distribution of the process $\{Y^{(i)} : i \in \mathbb{N}\}$. Now, for $i < j$, define the set

$$C(y^{(i:j)}) = \left\{(x^{(i)},\dots,x^{(j)}) \in A^{d\times(j-i+1)} : Y^{(i:j)} = y^{(i:j)}\right\},$$

which denotes all configurations $(x^{(i)},\dots,x^{(j)})$ such that $\{Y^{(i:j)} = y^{(i:j)}\}$ holds. Then

$$\begin{aligned}
&\mathbb{P}_Y\left(Y^{(n:(n+k-1))} = y^{(1:k)} \,\middle|\, Y^{(1:m)} = y^{(1:m)}\right) \\
&= \sum_{x^{(1:k)}\in C(y^{(1:k)})} \mathbb{P}\left(X^{(n:(n+k-1))} = x^{(1:k)} \,\middle|\, \cup_{x^{(1:m)}\in C(y^{(1:m)})} \left\{X^{(1:m)} = x^{(1:m)}\right\}\right) \\
&= \sum_{x^{(1:k)}\in C(y^{(1:k)})} \frac{\sum_{x^{(1:m)}\in C(y^{(1:m)})} \mathbb{P}\left(X^{(n:(n+k-1))} = x^{(1:k)}, X^{(1:m)} = x^{(1:m)}\right)}{\sum_{x^{(1:m)}\in C(y^{(1:m)})} \mathbb{P}\left(X^{(1:m)} = x^{(1:m)}\right)}
\end{aligned} \tag{20}$$

and similarly

$$\mathbb{P}_Y\left(Y^{(n:(n+k-1))} = y^{(1:k)}\right) = \sum_{x^{(1:k)}\in C(y^{(1:k)})} \mathbb{P}\left(X^{(n:(n+k-1))} = x^{(1:k)}\right).$$

Observe that by the mixing property (3) we obtain, for each $x^{(1:m)} \in C(y^{(1:m)})$, that

$$\begin{aligned}
&[1-\psi(n-m)]\,\mathbb{P}(X^{(n:(n+k-1))} = x^{(1:k)}) \le \\
&\qquad \frac{\mathbb{P}\left(X^{(n:(n+k-1))} = x^{(1:k)}, X^{(1:m)} = x^{(1:m)}\right)}{\mathbb{P}\left(X^{(1:m)} = x^{(1:m)}\right)} \\
&\qquad\qquad \le\ [1+\psi(n-m)]\,\mathbb{P}(X^{(n:(n+k-1))} = x^{(1:k)}).
\end{aligned} \tag{21}$$

Then, substituting the inequalities (21) in (20) we obtain that

$$
\begin{aligned}
&\Big| \mathbb{P}_Y\big(Y^{(n:(n+k-1))} = y^{(1:k)} \;\Big|\; Y^{(1:m)} = y^{(1:m)}\big) - \mathbb{P}_Y\big(Y^{(n:(n+k-1))} = y^{(1:k)}\big) \Big| \\
&\quad \le \sum_{x^{(1:k)} \in C(y^{(1:k)})} \Bigg| \frac{\sum_{x^{(1:m)} \in C(y^{(1:m)})} \mathbb{P}\big(X^{(n:(n+k-1))} = x^{(1:k)}, X^{(1:m)} = x^{(1:m)}\big)}{\sum_{x^{(1:m)} \in C(y^{(1:m)})} \mathbb{P}\big(X^{(1:m)} = x^{(1:m)}\big)} \\
&\qquad\qquad - \mathbb{P}\big(X^{(n:(n+k-1))} = x^{(1:k)}\big) \Bigg| \\
&\quad \le \sum_{x^{(1:k)} \in C(y^{(1:k)})} \psi(n-m) \mathbb{P}\big(X^{(n:(n+k-1))} = x^{(1:k)}\big) \\
&\quad = \psi(n-m)\, \mathbb{P}_Y\big(Y^{(n:(n+k-1))} = y^{(1:k)}\big)
\end{aligned}
$$

what proves that the process $\{Y^{(i)}, i \in \mathbb{N}\}$ is mixing with rate $\{\psi(\ell)\}_{\ell \in \mathbb{N}}$. □

The following result states a Law of the Iterated Logarithm for stationary stochastic processes satisfying the mixing condition (3), and the proof is based on the classical result by [31]. This result is essential to prove the rate of convergence of the empirical probabilities in Proposition 3.

Theorem 8. *Let $\{Y^{(i)} : i \in \mathbb{N}\}$ be stationary and satisfy the mixing condition (3) with rate $\psi(\ell) = O(1/\ell^{1+\epsilon})$, for some $\epsilon > 0$. Define $Z_n = \sum_{i=1}^n Y^{(i)}$. Then*

$$|Z_n| < 2\sqrt{2\sigma^2 n \log\log n}$$

eventually almost surely as $n \to \infty$, where

$$\sigma^2 = \mathbb{E}\big[\big(Y^{(1)}\big)^2\big] + 2\sum_{j=2}^{\infty} \mathbb{E}(Y^{(1)}Y^{(j)}). \tag{22}$$

Proof. The proof follows by [31, Theorem 4], which states that for a stationary process satisfying a strong mixing condition and additionally

1. $|Y^{(i)}| < c$ with probability one,
2. $\psi(\ell) = O(1/\ell^{1+\epsilon})$ for some $\epsilon > 0$,

we have

$$|Z_n| < 2\sqrt{2\sigma n \log\log n}$$

eventually almost surely as $n \to \infty$, where σ^2 is given by (22).

The first condition is automatically satisfied, as $Y^{(i)}$ is a random variable assuming a finite number of values. The second condition follows by the hypotheses of the theorem. This implies the desired result. □

Lemma 9. *Let $\{Y^{(i)} : i \in \mathbb{N}\}$ be a stationary process satisfying the mixing condition (3) with rate $\psi(\ell) = O(1/\ell^{1+\epsilon})$, for some $\epsilon > 0$, and assume $\mathbb{E}(Y^{(j)}) = 0$ and $Var(Y^{(j)}) \le \sigma_0^2$ for all j. Then*

$$\sigma^2 \le (1+\psi)\sigma_0^2$$

where σ^2 is defined by (22) and $\psi = \sum_{j=1}^{\infty} \psi(j)$.

Proof. Observe that we can write

$$
\begin{aligned}
\mathbb{E}\big(Y^{(1)}Y^{(j)}\big) &= \sum_{y_1, y_j} y_1 y_j \mathbb{P}\big(Y^{(1)} = y_1, Y^{(j)} = y_j\big) \\
&= \sum_{y_1, y_j} y_1 y_j \mathbb{P}\big(Y^{(j)} = y_j | Y^{(1)} = y_1\big) \mathbb{P}\big(Y^{(1)} = y_1\big).
\end{aligned}
$$

Then, by the mixing condition, for $j \ge 2$ we obtain

$$(1-\psi(j-1))\mathbb{P}\big(Y^{(j)} = y_j\big) \le \mathbb{P}\big(Y^{(j)} = y_j | Y^{(1)} = y_1\big) \le (1+\psi(j-1))\mathbb{P}\big(Y^{(j)} = y_j\big)$$

therefore

$$
\begin{aligned}
\mathbb{E}\big(Y^{(1)}Y^{(j)}\big) &\le \mathbb{E}(Y^{(1)})\mathbb{E}(Y^{(j)}) + \psi(j-1)\mathbb{E}\big(|Y^{(1)}|\big)\mathbb{E}\big(|Y^{(j)}|\big) \\
&\le \psi(j-1)\sigma_0^2
\end{aligned}
$$

since $\mathbb{E}\big(Y^{(j)}\big) = 0$ and $\mathbb{E}(|Y^{(j)}|) \le \sigma_0$ for all j. Then

$$\sigma^2 \le \sigma_0^2 + \sum_{j=2}^{\infty} \psi(j-1)\sigma_0^2 = (1+\psi)\sigma_0^2. \quad \square$$

Proof of Proposition 3. Fix $a_W \in A^W$ and define

$$Y^{(i)} = \mathbb{1}\{X_W^{(i)} = a_W\} - \pi(a_W), \tag{23}$$

for $i = 1, 2, \dots, n$. As $\{X^{(i)} : i \in \mathbb{N}\}$ is mixing with rate $\{\psi(\ell)\}_{\ell\in\mathbb{N}}$, then by Lemmas 6 and 7 the process $\{Y^{(i)} : i \in \mathbb{N}\}$ is also mixing with the same rate. Then by Theorem 8 and Lemma 9, for $Z_n = \sum_{i=1}^n Y^{(i)}$ we have that

$$|Z_n| < 2\sqrt{2(1+\psi)\sigma_0^2 n \log\log n},$$

eventually almost surely as $n \to \infty$, where $\psi = \sum_{j=1}^{\infty} \psi(j)$ and $\sigma_0^2 = \mathbb{E}((Y^{(i)})^2)$.

Since, by definition, $Z_n = N(a_W) - n\pi(a_W)$, we obtain that

$$\left|n^{-1} Z_n\right| = \left|\widehat{\pi}(a_W) - \pi(a_W)\right| < 2\sqrt{\frac{2(1+\psi)\sigma_0^2 \log\log n}{n}},$$

eventually almost surely as $n \to \infty$. Now, for any $\delta > 0$ we have that

$$8(1+\psi)\sigma_0^2 \log\log n < \delta \log n$$

for all n sufficiently large. Therefore,

$$\left|\widehat{\pi}(a_W) - \pi(a_W)\right| < \sqrt{\frac{\delta \log n}{n}}$$

eventually almost surely as $n \to \infty$, and this finishes the proof of the first part of Proposition 3. To prove the second part, fix the two disjoint subsets $W, W' \subset V$ and the configurations $a_W \in A^W, a_{W'} \in A^{W'}$. Define the process $\{Y^{(i)} : i \in \mathbb{Z}\}$ by

$$Y^{(i)} = \mathbb{1}\{X_W^{(i)} = a_W, X_{W'}^{(i)} = a_{W'}\} - \pi(a_W|a_{W'})\mathbb{1}\{X_W^{(i)} = a_{W'}\}$$

and let

$$Z_n = \sum_{i=1}^{n} Y^{(i)}. \tag{24}$$

Analogously to the first part, since $\{Y^{(i)} : i \in \mathbb{N}\}$ is mixing with rate $\psi(\ell) = O(1/\ell^{1+\epsilon})$, for some $\epsilon > 0$, by Theorem 8 we obtain that

$$|Z_n| < 2\sqrt{2(1+\psi)\sigma_0^2 n \log\log n}, \tag{25}$$

eventually almost surely as $n \to \infty$, where $\sigma_0^2 = \mathbb{E}((Y^{(i)})^2)$. Note that Z_n defined in (24) can be written as

$$Z_n = N(a_{W\cup W'}) - \pi(a_W|a_{W'})N(a_{W'}).$$

If we divide Z_n by $N(a_{W'})$ and we use that $\sigma_0^2 \le \frac{1}{4}\pi(a_{W'})$, by (25) we get that

$$\left|\widehat{\pi}(a_W|a_{W'}) - \pi(a_W|a_{W'})\right| < 2\sqrt{\frac{\pi(a_{W'})(1+\psi)n\log\log n}{2N(a_{W'})^2}}$$

eventually almost surely as $n \to \infty$. By Proposition 3, we also have that

$$N(a_{W'}) \;>\; n\pi(a_{W'}) - \sqrt{\delta n \log n} > \frac{1}{2} n\pi(a_{W'})$$

eventually almost surely as $n \to \infty$. Then we obtain that

$$\left|\widehat{\pi}(a_W|a_{W'}) - \pi(a_W|a_{W'})\right| < 2\sqrt{\frac{(1+\psi)\log\log n}{N(a_{W'})}}.$$

As before, for any $\delta > 0$ we have that

$$4(1+\psi)\log\log n \;< \delta \log n$$

for sufficiently large n, and therefore for all $\delta > 0$

$$\left|\widehat{\pi}(a_W|a_{W'}) - \pi(a_W|a_{W'})\right| < \sqrt{\frac{\delta \log n}{N(a_{W'})}}$$

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

The following basic result about the Küllback–Leibler divergence corresponds to [26, Lemma 6.3]. We omit its proof here.

Lemma 10. *For any P and Q we have*

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

References

[1] S.L. Lauritzen, Graphical Models, vol. 17, Clarendon Press, 1996.
[2] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning, The MIT Press, 2009.
[3] M. Lerasle, D.Y. Takahashi, Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields, Bernoulli 22 (1) (2016) 325–344.
[4] J. Pensar, H. Nyman, J. Corander, Structure learning of contextual Markov networks using marginal pseudo-likelihood, Scand. J. Stat. 44 (2) (2017) 455–479.
[5] F. Divino, A. Frigessi, P.J. Green, Penalized pseudolikelihood inference in spatial interaction models with covariates, Scand. J. Stat. 27 (3) (2000) 445–458.
[6] F. Leonardi, R. Carvalho, I. Frondana, Structure recovery for partially observed discrete Markov random fields on graphs under not necessarily positive distributions, Scand. J. Stat. 51 (1) (2024) 64–88.
[7] A. Shojaie, G. Michailidis, Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs, Biometrika 97 (3) (2010) 519–538.
[8] D. Strauss, M. Ikeda, Pseudolikelihood estimation for social networks, J. Amer. Statist. Assoc. 85 (409) (1990) 204–212.
[9] A. Duarte, A. Galves, E. Löcherbach, G. Ost, Estimating the interaction graph of stochastic neural dynamics, Bernoulli 25 (1) (2019) 771–792.
[10] P. Ravikumar, M.J. Wainwright, J.D. Lafferty, High-dimensional Ising model selection using l_1-regularized logistic regression, Ann. Statist. 38 (3) (2010) 1319–3022.
[11] A. Galves, E. Orlandi, D.Y. Takahashi, Identifying interacting pairs of sites in Ising models on a countable set, Braz. J. Probab. Stat. 29 (2) (2015) 443–459.
[12] G. Bresler, D. Gamarnik, D. Shah, Learning graphical models from the Glauber dynamics, IEEE Trans. Inform. Theory 64 (6) (2018) 4072–4080.
[13] Y.F. Atchade, Estimation of high-dimensional partially-observed discrete Markov random fields, Electron. J. Statist. 8 (2) (2014) 2242–2263.
[14] H. Höfling, R. Tibshirani, Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods, J. Mach. Learn. Res. 10 (32) (2009) 883–906.
[15] N.P. Santhanam, M.J. Wainwright, Information-theoretic limits of selecting binary graphical models in high dimensions, IEEE Trans. Inform. Theory 58 (7) (2012) 4117–4134.
[16] P.-L. Loh, M.J. Wainwright, Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses, Ann. Statist. 41 (6) (2013) 3022–3049.
[17] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, Ann. Statist. 34 (3) (2006) 1436–1462.
[18] J. Lafferty, H. Liu, L. Wasserman, Sparse nonparametric graphical models, Statist. Sci. 27 (4) (2012) 519–537.
[19] H. Liu, F. Han, M. Yuan, J. Lafferty, L. Wasserman, High-dimensional semiparametric Gaussian copula graphical models, Ann. Statist. 40 (4) (2012) 2293–2326.
[20] E. Yang, P. Ravikumar, G.I. Allen, Z. Liu, Graphical models via univariate exponential family distributions, J. Mach. Learn. Res. 16 (115) (2015) 3813–3847.
[21] H.-O. Georgii, Gibbs measures and phase transitions, second ed., in: de Gruyter Studies in Mathematics, vol. 9, Walter de Gruyter & Co., Berlin, 2011.
[22] F. Comets, On consistency of a class of estimators for exponential families of Markov random fields on the lattice, Ann. Statist. 20 (1) (1992) 455–468.
[23] F. Comets, B. Gidas, Parameter estimation for gibbs distributions from partially observed data, Ann. Appl. Probab. 2 (1) (1992) 142–170.
[24] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[25] C. Ji, L. Seymour, A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood, Ann. Appl. Probab. 6 (2) (1996) 423–443.
[26] I. Csiszár, Z. Talata, Consistent estimation of the basic neighborhood of Markov random fields, Ann. Statist. 34 (1) (2006) 123–145.
[27] H. Tjelmeland, J. Besag, Markov random fields with higher-order interactions, Scand. J. Stat. 25 (3) (1998) 415–433.
[28] E. Löcherbach, E. Orlandi, Neighborhood radius estimation for variable-neighborhood random fields, Stochastic Process. Appl. 121 (9) (2011) 2151–2185.
[29] A. Cerqueira, D. Fraiman, C.D. Vargas, F. Leonardi, A test of hypotheses for random graph distributions built from EEG data, IEEE Trans. Netw. Sci. Eng. 4 (2) (2017) 75–82.
[30] F. Leonardi, M. Lopez-Rosenfeldz, D. Rodriguez, M.T.F. Severino, M. Sued, Independent block identification in multivariate time series, J. Time Series Anal. 42 (1) (2020) 19–33.
[31] H. Oodaira, K.-i. Yoshihara, The law of the iterated logarithm for stationary processes satisfying mixing conditions, Kodai Math. Semin. Rep. 23 (3) (1971) 311–334.
[32] I. Csiszár, Large-scale typicality of Markov sample paths and consistency of MDL order estimators, IEEE Trans. Inform. Theory 48 (6) (2002) 1616–1628, Special issue on Shannon theory: perspective, trends, and applications.