# Identifiability and Consistency of Bayesian Network Structure Learning from Incomplete Data

**Tjebbe Bodewes** TJEBBE.BODEWES@LINACRE.OX.AC.UK
*Zivver, Rotterdam, The Netherlands; formerly Departments of Statistics, University of Oxford, UK*

**Marco Scutari** SCUTARI@IDSIA.CH
*Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland*

## Abstract

Bayesian network (BN) structure learning from complete data has been extensively studied in the literature. However, fewer theoretical results are available for incomplete data, and most are based on the use of the Expectation-Maximisation (EM) algorithm. Balov (2013) proposed an alternative approach called Node-Average Likelihood (NAL) that is competitive with EM but computationally more efficient; and proved its consistency and model identifiability for discrete BNs.

In this paper, we give general sufficient conditions for the consistency of NAL; and we prove consistency and identifiability for conditional Gaussian BNs, which include discrete and Gaussian BNs as special cases. Hence NAL has a wider applicability than originally stated in Balov (2013).

**Keywords:** Bayesian networks; score-based structure learning; incomplete data.

## 1. Introduction

Bayesian Networks (BNs; Pearl, 1988) are a class of graphical models in which the nodes of a directed acyclic graph (DAG) $\mathcal{G}$ represent a set $\mathbf{X} = \{X_1, \ldots, X_N\}$ of random variables describing some quantities of interest. The arcs connecting those nodes express direct dependence relationships, with graphical separation in $\mathcal{G}$ (called *d-separation*) implying conditional independence in probability. Multiple DAGs can represent the same set of independencies, and can be grouped into *equivalence classes* (Chickering, 1995) whose elements are probabilistically indistinguishable without additional information. $\mathcal{G}$ induces the factorisation

$$\mathrm{P}_{\mathcal{G}}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}\left(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}\right), \tag{1}$$

in which the joint distribution of $\mathbf{X}$ decomposes into one *local distribution* for each $X_i$ (with parameters $\Theta_{X_i}$, $\bigcup_{X_i \in \mathbf{X}} \Theta_{X_i} = \Theta$) conditional on its parents $\Pi_{X_i}$. Thus BNs provide a compact and modular representation of high-dimensional problems.

As for the probability distribution of $\mathbf{X}$, the literature has mostly focused on three cases for analytical and computational reasons: *discrete BNs* (Heckerman et al., 1995), in which both $\mathbf{X}$ and the $X_i$ are multinomial random variables; *Gaussian BNs* (GBNs; Geiger and Heckerman, 1994), in which $\mathbf{X}$ is multivariate normal and the $X_i$ are univariate normals linked by linear dependencies; and *conditional Gaussian BNs* (CGBNs; Heckerman and Geiger, 1995) which combine discrete and continuous random variables in a mixture-of-Gaussians model. CGBNs are defined as follows:

- Discrete $X_i$ are only allowed to have discrete parents (denoted $\Delta_{X_i}$), and are assumed to follow a multinomial distribution. Their parameters $\Theta_{X_i}$ are the conditional probabilities $\pi_{ik\,|\,j} = \mathrm{P}(X_i = k \,|\, \Delta_{X_i} = j)$.

- Continuous $X_i$ are allowed to have both discrete and continuous parents (denoted $\Gamma_{X_i}$, with $\Delta_{X_i} \cup \Gamma_{X_i} = \Pi_{X_i}$), and their local distributions are

$$X_i = \mu_{ij} + \Gamma_{X_i}\boldsymbol{\beta}_{ij} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2_{ij} I) \tag{2}$$

  which define a mixture of linear regressions against the continuous parents with one component for each configuration $j$ of the discrete parents $\Delta_{X_i}$. Hence $\Theta_{X_i} = \{\mu_{ij}, \boldsymbol{\beta}_{ij}, \sigma^2_{ij}\}$. If $X_i$ has no discrete parents, the mixture simplifies to a single linear regression.

We denote discrete nodes with $\boldsymbol{\Delta}$ and Gaussian nodes as $\boldsymbol{\Gamma}$, with $\boldsymbol{\Delta} \cup \boldsymbol{\Gamma} = \mathbf{X}$. If $\boldsymbol{\Delta} = \varnothing$ or $\boldsymbol{\Gamma} = \varnothing$, a CGBN reduces respectively to a GBN or to a discrete BN. For this reason, we will consider only CGBNs but our results will hold for both discrete BNs and GBNs as particular cases.

The task of learning a BN from a data set $\mathcal{D}$ containing $n$ independent observations is performed in two steps: structure learning and parameter learning. *Structure learning* consists in finding the DAG $\mathcal{G}$ that encodes the dependence structure of the data in the space $\mathbb{G}$ of all possible DAGs, thus maximising $\mathrm{P}(\mathcal{G} \mid \mathcal{D})$ or some alternative goodness-of-fit measure. *Parameter learning* consists in estimating the parameters $\Theta$ given the $\mathcal{G}$ obtained from structure learning, that is $\mathrm{argmax}_{\Theta} \mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})$. If the data contain no missing values (Heckerman et al., 1995), parameter learning is straightforward since it can be implemented independently for each $X_i$; Scutari (2020) provides a review of suitable approaches for incomplete data.

On the other hand, structure learning is NP-hard (Chickering and Heckerman, 1994). Many algorithms have been proposed for this problem; a comprehensive review and comparison can be found in Scutari et al. (2019) for complete data and in Scutari (2020) for incomplete data. In the latter case, the Expectation-Maximisation algorithm (EM; Dempster et al., 1977) is commonly embedded within structure learning to reuse algorithms originally proposed for complete data. However, this choice comes at a significant computational cost. To address this issue, Balov (2013) proposed an alternative approach based on the Node-Averaged Likelihood (NAL) that is competitive with EM-based approaches in terms of structural accuracy at a much lower computational cost. He then proved both identifiability and consistency for NAL for discrete BNs.

In this paper we will establish general conditions for both properties and we will show that they hold for CGBNs. In Section 2 we will briefly review score-based learning from complete data, moving to incomplete data in Section 3. Our novel results and the required regularity conditions will be introduced in Section 4 for both identifiability (Section 4.1) and consistency (Section 4.2). Proofs for all but the main consistency result are included in Appendix A.

## 2. Structure Learning from Complete Data

BN structure learning consists of two components: a score function $S(\mathcal{G} \mid \mathcal{D})$ and an algorithm that determines how we explore the space DAGs.[1] Each candidate DAG is assigned a score $S(\mathcal{G} \mid \mathcal{D})$ reflecting its goodness of fit, which the algorithm then attempts to maximise to obtain a (possibly local) optimum DAG as $\mathrm{argmax}_{\mathcal{G} \in \mathbb{G}} S(\mathcal{G} \mid \mathcal{D})$. Heuristic algorithms such as tabu search (Glover and Laguna, 1998) are more common in practical applications, but exact algorithms have been proposed

1. Other approaches using conditional independence tests (constraint-based algorithms) alone or in combination with score functions (hybrid algorithms) are beyond the scope of this paper; we refer the interested reader to Scutari et al. (2019).

as well (for instance Cussens, 2012). As for the scoring metric, we can say $\mathrm{P}(\mathcal{G} \mid \mathcal{D}) \propto \mathrm{P}(\mathcal{D} \mid \mathcal{G})$ and use the *marginal likelihood* $\mathrm{P}(\mathcal{D} \mid \mathcal{G})$ to define

$$S_{\mathrm{ML}}(\mathcal{G} \mid \mathcal{D}) = \int \mathrm{P}(\mathcal{D} \mid \mathcal{G}, \Theta)\, \mathrm{P}(\Theta \mid \mathcal{G})\, d\Theta = \prod_{i=1}^{N} \int \mathrm{P}(X_i \mid \Pi_{X_i}, \Theta_{X_i})\, \mathrm{P}(\Theta_{X_i} \mid \Pi_{X_i})\, d\Theta_{X_i}, \quad (3)$$

which decomposes into one term for each node due to (1). This speeds up structure learning considerably because (3) is available in closed form for discrete BNs, GBNs and CGBNs (Heckerman and Geiger, 1995) and because we only recompute differing portions of $\mathrm{P}(\mathcal{D} \mid \mathcal{G})$ as we score and compare DAGs. It is also consistent for complete data.

Due to the difficulty of choosing a prior over $\Theta$ and the resulting performance implications (Scutari, 2017), a common alternative is the *Bayesian Information Criterion* (BIC; Schwarz, 1978),

$$\ell(\mathcal{G}, \Theta \mid \mathcal{D}) = \frac{1}{n} \sum\nolimits_{X_i \in \mathbf{X}} \log \mathrm{P}(X_i \mid \Pi_{X_i}, \widehat{\Theta}_{X_i}), \quad (4)$$

$$S_{\mathrm{BIC}}(\mathcal{G} \mid \mathcal{D}) = \ell(\mathcal{G}, \Theta \mid \mathcal{D}) - \frac{\log(n)}{2n} |\Theta|\,; \quad (5)$$

where the $\widehat{\Theta}_{X_i}$ are the maximum likelihood estimates (MLEs) of the $\Theta_{X_i}$ for $\mathcal{D}$. In contrast, we will denote by $\ell(\mathcal{G}, \Theta)$ and $\ell(X_i \mid \Pi_{X_i}, \Theta_{X_i})$ the population log-likelihoods. (5) is a particular case of the penalised log-likelihood

$$S_{\mathrm{PL}}(\mathcal{G} \mid \mathcal{D}) = \ell(\mathcal{G}, \Theta \mid \mathcal{D}) - \lambda_n h(\mathcal{G}), \qquad \lambda_n \geq 0 \quad (6)$$

where $\lambda_n$ is a penalisation coefficient and $h : \mathbb{G} \to \mathbb{R}^+$ measures model complexity such that $\mathcal{G}_1 \subset \mathcal{G}_2 \Rightarrow h(\mathcal{G}_1) < h(\mathcal{G}_2)$.[2] $S_{\mathrm{BIC}}(\mathcal{G} \mid \mathcal{D})$ is decomposable, since $h(\mathcal{G}) = |\Theta| = \sum_{\mathbf{X}} |\Theta_{X_i}|$ which is the number of parameters of each $X_i \mid \Pi_{X_i}$ in $\mathcal{G}$; and it does not depend on any hyperparameter. Furthermore, it is equivalent to the minimum description length (Rissanen, 2007) of $\mathcal{G}$ and it is asymptotically equivalent to $S_{\mathrm{ML}}(\mathcal{G} \mid \mathcal{D})$. Hence, it is consistent for complete data. Setting $\lambda_n = 1/n$ instead of $\lambda_n = \log(n)/2n$ gives the *Akaike Information Criterion* (AIC; Akaike, 1974) which, on the other hand, is not consistent for complete data (Bozdogan, 1987).

## 3. Structure Learning from Incomplete Data

In the context of BNs, incomplete data are modelled using auxiliary nodes $Z_i \in \mathbf{Z}$ that encode whether the corresponding $X_i$ is observed for each observation. The patterns of missingness originally introduced in Little and Rubin (1987) can then be modelled graphically: missing completely at random (MCAR) implies $\mathbf{Z} \perp\!\!\!\perp \mathbf{X}$; missing at random (MAR) implies $\mathbf{Z} \perp\!\!\!\perp \mathbf{X}$ for the incomplete observations conditional on the complete observations; and missing not at random (MNAR) does not imply any independence constraint. Note that, however, in the following we will still consider $\mathcal{G}$ to be spanning just $\mathbf{X}$ because we will restrict ourselves to the MCAR case in which the dependencies between the $\mathbf{X}$ and the $\mathbf{Z}$ are completely determined.

When the data are incomplete, scoring $\mathcal{G}$ requires the missing values to be integrated out thus making $S_{\mathrm{ML}}(\mathcal{G} \mid \mathcal{D})$ and $S_{\mathrm{BIC}}(\mathcal{G} \mid \mathcal{D})$ no longer decomposable. This can be avoided by using Expectation-Maximisation (EM; Dempster et al., 1977), an iterative procedure consisting of an

2. We say $\mathcal{G}_1 \subseteq \mathcal{G}_2$ if the arc set of $\mathcal{G}_1$ is a subset of that of $\mathcal{G}_2$.

E(xpectation)-step and a M(aximisation)-step. In the E-step, we compute the expected sufficient statistics conditional on the observed data using belief propagation (Lauritzen, 1995; Pearl, 1988; Shafer and Shenoy, 1990). In the M-step, complete-data learning methods can be applied using the expected sufficient statistics instead of the (unobservable) empirical ones.

There are two ways to apply EM to structure learning. Firstly, we can apply EM separately to each candidate DAG to be scored, as in the variational Bayes EM (Beal and Ghahramani, 2003). However, structure learning often involves many evaluations of the score function, thus making this approach computationally infeasible beyond small-scale problems. Secondly, we can embed structure learning in the M-step, estimating the expected sufficient statistics using the current best DAG. This approach is called Structural EM (Friedman, 1997, 1998), and is more efficient since it requires fewer applications of belief propagation.

Even so, Structural EM is computationally demanding. Balov (2013) proposed a more scalable approach for discrete BNs under MCAR called Node-Average Likelihood (NAL). While Balov (2013) defined NAL relying on the specific form of the multinomial log-likelihood, we will present it here using a more general definition that allows its extension to CGBNs. Starting from (1), he proposed to compute each term using the $\mathcal{D}_{(i)} \subseteq \mathcal{D}$ locally-complete data for which $X_i, \Pi_{X_i}$ are observed (that is, $\mathbf{Z}_{(i)} = \mathbf{Z}_{X_i,\Pi_{X_i}} = 1$):

$$\bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) = \frac{1}{|\mathcal{D}_{(i)}|} \sum_{\mathcal{D}_{(i)}} \log \mathrm{P}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}),$$

which is an empirical estimate of the average node log-likelihood $\mathrm{E}[\ell(X_i \,|\, \Pi_{X_i})]$. Replacing (4) with the above gives

$$\bar{\ell}(\mathcal{G}, \Theta \,|\, \mathcal{D}) = \sum\nolimits_{X_i \in \mathbf{X}} \bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i})$$

which Balov (2013) used to redefine the penalised log-likelihood score function as

$$S_{\mathrm{PL}}(\mathcal{G} \,|\, \mathcal{D}) = \bar{\ell}(\mathcal{G}, \Theta \,|\, \mathcal{D}) - \lambda_n h(\mathcal{G}) \tag{7}$$

and structure learning as $\widehat{\mathcal{G}} = \mathrm{argmax}_{\mathcal{G} \in \mathbb{G}} S_{\mathrm{PL}}(\mathcal{G} \,|\, \mathcal{D})$. NAL makes a more efficient use of incomplete data than discarding all incomplete samples, without incurring in the computational costs of EM approaches. However, comparing two DAGs means that NAL is evaluated on potentially different subsets of $\mathcal{D}$ for each $X_i$ in different DAGs; hence the usual results on MLEs and nested models do not apply. Balov (2013) proved both identifiability and consistency of score-based structure learning when using $S_{\mathrm{PL}}(\mathcal{G} \,|\, \mathcal{D})$ for discrete BNs. We will now prove both properties hold more generally, and in particular that they hold for CGBNs.

## 4. Properties of Node-Average Likelihood

We study two properties of NAL: whether the true DAG $\mathcal{G}_0$ is *identifiable*, and under which conditions $\widehat{\mathcal{G}}$ is a *consistent* estimator of $\mathcal{G}_0$. For each of our results, we reference the corresponding theorems that Balov (2013) derived for discrete BNs. Our contribution is the generalisation of these results under mild regularity conditions, which requires completely different proofs since Balov (2013) relied materially on the form of the multinomial log-likelihood and on not having both continuous and discrete parents in each $\Pi_{X_i}$.

### 4.1 Identifiability

In this section we show under what conditions NAL can be used to identify the true $\mathcal{G}_0$ of the BN from $\mathcal{D}$. Firstly, we prove that NAL is non-decreasing in the size of the parent sets and thus overfits like the log-likelihood does for complete data.

**Lemma 1 (L7.1)** *For any $X_i$ and disjoint $\mathbf{A}, \mathbf{B} \subset \mathbf{X}$ such that $\mathbf{Z_B} \perp\!\!\!\perp \{X_i, \mathbf{A}, \mathbf{B}\} \,|\, \mathbf{Z}_{X_i,\mathbf{A}} = 1$, we have $\bar{\ell}(X_i \,|\, \mathbf{A}) \leqslant \bar{\ell}(X_i \,|\, \mathbf{A}, \mathbf{B})$, with equality if and only if $X_i \perp\!\!\!\perp \mathbf{B} \,|(\mathbf{A}, \mathbf{Z}_{X_i,\mathbf{A}} = 1)$.*

Lemma 1 allows us to state that if $\mathcal{G}_0$ is identifiable, we can learn it by finding the simplest DAG that maximises NAL.

**Definition 2 (D3.1)** *$\mathcal{G}_0$ is identifiable if for any $\mathcal{G} \in \mathbb{G}$ we have $\bar{\ell}(\mathcal{G}, \Theta) \leq \bar{\ell}(\mathcal{G}_0, \Theta_0)$ when $\mathcal{G}_0 \subseteq \mathcal{G}$; and $\bar{\ell}(\mathcal{G}, \Theta) < \bar{\ell}(\mathcal{G}_0, \Theta_0)$ when $\mathcal{G}_0 \not\subseteq \mathcal{G}$.*

We next establish that under MCAR $\ell(\mathcal{G}, \Theta)$ attains its maximum at all $\mathcal{G} \supseteq \mathcal{G}_0$ and that all these DAGs induce the true distribution $\mathrm{P}_{\mathcal{G}_0}(\mathbf{X}) = \prod_{\mathbf{X}} \mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}_0)}, \Theta_{X_i})$.

**Proposition 3 (P3.1)** *Under MCAR, we have:*

*(i)* $\max_{\mathcal{G} \in \mathbb{G}} \bar{\ell}(\mathcal{G}, \Theta) = \bar{\ell}(\mathcal{G}_0, \Theta_0)$.

*(ii)* *if* $\bar{\ell}(\mathcal{G}, \Theta) = \bar{\ell}(\mathcal{G}_0, \Theta_0)$, *then* $\mathrm{P}_{\mathcal{G}}(\mathbf{X}) = \mathrm{P}_{\mathcal{G}_0}(\mathbf{X})$.

*(iii)* *if* $\mathcal{G}_0 \subseteq \mathcal{G}$, *then* $\bar{\ell}(\mathcal{G}, \Theta) = \bar{\ell}(\mathcal{G}_0, \Theta_0)$.

The identifiability of $\mathcal{G}_0$ up to its equivalence class $[\mathcal{G}_0]$ follows from the above and is formally stated below.

**Corollary 4 (C3.2)** *$[\mathcal{G}_0]$ is identifiable under MCAR, that is*

$$\mathcal{G}_0 \cong \min \left\{ \mathcal{G}_* \in \mathbb{G} : \bar{\ell}(\mathcal{G}_*, \Theta_*) = \max_{\mathcal{G} \in \mathbb{G}} \bar{\ell}(\mathcal{G}, \Theta) \right\}.$$

### 4.2 Consistency

In this section we show that the candidate $\widehat{\mathcal{G}}$ chosen by maximising $S_{\mathrm{PL}}(\mathcal{G} \,|\, \mathcal{D})$ is a consistent estimator of the true $\mathcal{G}_0$ under MCAR, and the mild regularity conditions this requires.

#### 4.2.1 REGULARITY CONDITIONS

For all $\mathcal{G} \in \mathbb{G}$ and $X_i \in \mathbf{X}$:

(R1) $\widehat{\Theta}_{X_i}$ must exist, converge in probability to the population $\Theta_{X_i}$ with speed $O(n^{-1/2})$, and make $\nabla_{\Theta_{X_i}} \bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i})$ vanish.

(R2) The Hessian $\mathbf{H}_{\Theta_{X_i}}\big(\bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i})\big)$ exists and has finite expectation.

(R3) The variance $\nu^2 = \mathrm{VAR}\left(\bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i})\right) < \infty$.

Conditions (R1) and (R2) ensure that the NAL evaluated at the MLE is close to the NAL at the population $\Theta_{X_i}$ for large $n$; (R3) allows the use of various Central Limit Theorems (CLTs; Billingsley, 1995). (R1), (R2), (R3) hold for CGBNs.

**Proposition 5** *Consider a CGBN over $\mathbf{X}$. If $\Theta_{X_i} > 0$ for all $X_i \in \mathbf{\Delta}$, and $\sigma^2_{ij} > 0$ for all $X_i \in \mathbf{\Gamma}$ and all values $j$ of $\Delta_{X_i}$, then (R1), (R2), (R3) are satisfied.*

A more general approach would be to model the $X_i \in \mathbf{\Gamma}$ with a mixture of generalised linear models (McCullagh and Nelder, 1989). When using canonical link functions, (R1) and (R2) simplify due to the results in Fahrmeir and Kaufmann (1985). However, verifying (R1), (R2) and (R3) is non-trivial hence we leave this as a potentially tractable case for future work.

#### 4.2.2 CONSISTENCY RESULTS

Starting from the consistency of $\bar{\ell}(\mathcal{G}, \Theta \,|\, \mathcal{D})$, we will now establish for which sequences of $\lambda_n$ the $S_{\text{PL}}(\mathcal{G} \,|\, \mathcal{D})$ in (7) is consistent for complete data and under MCAR, that is, $\lim_{n\to\infty} P(\widehat{\mathcal{G}} = \mathcal{G}_0) = 1$. We note the sufficient conditions for the consistency of $\widehat{\mathcal{G}}$ from Balov (2013, P3.2, C4.1):

(C1) If $\mathcal{G}_0 \subseteq \mathcal{G}_1$ and $\mathcal{G}_0 \not\subseteq \mathcal{G}_2$, then $\lim_{n\to\infty} P(S_{\text{PL}}(\mathcal{G}_1 \,|\, \mathcal{D}) > S_{\text{PL}}(\mathcal{G}_2 \,|\, \mathcal{D})) = 1$.

(C2) If $\mathcal{G}_0 \subseteq \mathcal{G}_1$ and $\mathcal{G}_1 \subset \mathcal{G}_2$, then $\lim_{n\to\infty} P(S_{\text{PL}}(\mathcal{G}_1 \,|\, \mathcal{D}) > S_{\text{PL}}(\mathcal{G}_2 \,|\, \mathcal{D})) = 1$.

(C3) $\exists\, \mathcal{G}, X_i, X_j, i \neq j$ such that $\Pi_{X_i}^{(\mathcal{G}_0)} \subset \Pi_{X_i}^{(\mathcal{G})}$, $\Pi_{X_j}^{(\mathcal{G})} = \Pi_{X_j}^{(\mathcal{G}_0)}$ and $\mathrm{P}(\mathbf{Z}_{\Pi_{X_i}^{(\mathcal{G})} \setminus \Pi_{X_i}^{(\mathcal{G}_0)}} \,|\, \mathbf{Z}_{(i)}^{(\mathcal{G}_0)}) \in (0, 1)$.

These conditions can be trivially extended to equivalence classes as in Balov (2013, C3.3).

To prove consistency, we first establish some intermediate results. Unlike Balov (2013), we need a rigorous treatment of scoring DAGs that may represent misspecified models (White, 1982) that are not representable in terms of $\mathcal{G}_0$; an example would be a $X_i \in \mathbf{\Gamma}$ being a mixture of regressions in $\mathcal{G}$ and a single linear regressions in $\mathcal{G}_0$. In such cases minimising Kullback-Leibler distances to obtain MLEs does necessarily make them vanish as $n \to \infty$.

Lemma 6 shows that if the difference in NAL between two DAGs is $O(n^{-\alpha})$, then the less complex DAG is chosen asymptotically if $\lambda_n \to 0$ slower than $n^{-\alpha}$.

**Lemma 6** *For $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{G}$, if (i) $\bar{\ell}(\mathcal{G}_1, \Theta_1 \,|\, \mathcal{D}) - \bar{\ell}(\mathcal{G}_2, \Theta_2 \,|\, \mathcal{D}) = O(n^{-\alpha})$ for $\alpha > 0$; (ii) $h(\mathcal{G}_1) < h(\mathcal{G}_2)$; (iii) $\lim_{n\to\infty} n^{\alpha}\lambda_n = \infty$; then $\lim_{n\to\infty} P(S_{\text{PL}}(\mathcal{G}_1 \,|\, \mathcal{D}) > S_{\text{PL}}(\mathcal{G}_2 \,|\, \mathcal{D})) = 1$.*

Lemma 7 establishes that the difference between the NAL at $\widehat{\Theta}_{X_i}$ and the NAL at the population $\Theta_{X_i}$ is $O(n^{-1})$, which is relevant due to Lemma 6. Working with the latter allows us to exploit conditional independencies and to apply CLTs by turning it into a sum of i.i.d. random variables.

**Lemma 7** *If (R1) and (R2) hold, then for all $\mathcal{G}$ and $X_i$:*

$$\bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) - \bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}) = O(n^{-1}).$$

Lemma 8 establishes that if $\mathcal{G} \supseteq \mathcal{G}_0$ the local distributions for $\mathcal{G}$ reduce to those of $\mathcal{G}_0$, and that population and sample distributions coincide. These results are crucial in linking $\bar{\ell}(\mathcal{G}, \Theta \,|\, \mathcal{D})$ and $\bar{\ell}(\mathcal{G}, \Theta)$ in Lemma 9.

**Lemma 8** *If $\mathcal{G} \supseteq \mathcal{G}_0$, then for each $X_i \in \mathbf{X}$:.*

*(i)* $\mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G})}, \Theta_{X_i}, \mathbf{Z}_{(i)}) = \mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}_0)}, \Theta_{X_i}, \mathbf{Z}_{(i)})$.

*(ii)* $\mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G})}, \Theta_{X_i}, \mathcal{D}) = \mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}_0)}, \Theta_{X_i}, \mathcal{D})$ *almost surely.*

*(iii)* $\mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G})}, \Theta_{X_i}, \mathcal{D}) = \mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G})}, \Theta_{X_i}, \mathbf{Z}_{(i)})$ *almost surely.*

Lemma 9 links the sample NAL and the population NAL and establishes the convergence rate of the former. It will be used to show (C1) in Theorem 10 by way of Proposition 3.

**Lemma 9 (L4.1)** *If (R1), (R2), (R3) are satisfied, then*

$$\lim_{n\to\infty} \mathrm{P}(\bar{\ell}(\mathcal{G}, \Theta \,|\, \mathcal{D}) \leq \bar{\ell}(\mathcal{G}, \Theta)) = 1 \qquad \text{for any } \mathcal{G} \in \mathbb{G}. \tag{8}$$

*Furthermore,* $\bar{\ell}(\mathcal{G}, \Theta \,|\, \mathcal{D}) - \bar{\ell}(\mathcal{G}, \Theta) = O(n^{-1/2})$ *if* $\mathcal{G}_0 \subseteq \mathcal{G}$.

Theorem 10 is the key result of this section, showing that BIC is consistent for complete data but it is not under MCAR. AIC is not consistent in either case.

**Theorem 10 (T4.1)** *Let* $\mathcal{G}_0$ *be identifiable,* $\lambda_n \to 0$ *as* $n \to \infty$*, and assume (R1), (R2) are satisfied. Then as* $n \to \infty$*:*

*(i) If* $\mathrm{P}(\mathbf{Z} = 1) = 1$ *and* $n\lambda_n \to \infty$*,* $\widehat{\mathcal{G}}$ *is consistent.*

*(ii) Under MCAR and (R3), if* $\sqrt{n}\lambda_n \to \infty$*,* $\widehat{\mathcal{G}}$ *is consistent.*

*(iii) Under MCAR, (C3) and (R3), if* $\liminf_{n\to\infty} \sqrt{n}\lambda_n < \infty$*, then* $\widehat{\mathcal{G}}$ *is not consistent.*

**Proof of Theorem 10.** For $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{G}$ such that $\mathcal{G}_0 \subseteq \mathcal{G}_1$, $\mathcal{G}_0 \not\subseteq \mathcal{G}_2$, we must show that $\lim_{n\to\infty} \mathrm{P}(S_{\mathrm{PL}}(\mathcal{G}_1 \,|\, \mathcal{D}) > S_{\mathrm{PL}}(\mathcal{G}_2 \,|\, \mathcal{D})) = 1$.

Lemma 1 implies that $\bar{\ell}(\mathcal{G}_1, \Theta_1) \geqslant \bar{\ell}(\mathcal{G}_0, \Theta_0)$, while identifiability gives $\bar{\ell}(\mathcal{G}_2, \Theta_2) < \bar{\ell}(\mathcal{G}_0, \Theta_0)$, resulting in $\bar{\ell}(\mathcal{G}_1, \Theta_1) > \bar{\ell}(\mathcal{G}_2, \Theta_2)$. Lemma 9 implies that $\bar{\ell}(\mathcal{G}_1, \Theta_1 \,|\, \mathcal{D}) \to \bar{\ell}(\mathcal{G}_1, \Theta_1)$ and $\mathrm{P}(\bar{\ell}(\mathcal{G}_2, \Theta_2 \,|\, \mathcal{D}) \leqslant \bar{\ell}(\mathcal{G}_2, \Theta_2)) = 1$ as $n \to \infty$. We deduce that $\lim_{n\to\infty} \mathrm{P}(\bar{\ell}(\mathcal{G}_1, \Theta_1 \,|\, \mathcal{D}) > \bar{\ell}(\mathcal{G}_2, \Theta_2 \,|\, \mathcal{D})) = 1$. As $\lambda_n \to 0$ for $n \to \infty$, we have $\lim_{n\to\infty} \mathrm{P}(S_{\mathrm{PL}}(\mathcal{G}_1 \,|\, \mathcal{D}) > S_{\mathrm{PL}}(\mathcal{G}_2 \,|\, \mathcal{D})) = 1$ and condition (C1) is satisfied for parts (i)-(iii).

We show that under the conditions in parts (i) and (ii), $\lim_{n\to\infty} \mathrm{P}(S_{\mathrm{PL}}(\mathcal{G}_1 \,|\, \mathcal{D})) > S_{\mathrm{PL}}(\mathcal{G}_2 \,|\, \mathcal{D})) = 1$ for $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{G}$ such that $\mathcal{G}_0 \subseteq \mathcal{G}_1 \subset \mathcal{G}_2$. This means that (C2) is satisfied and $\widehat{\mathcal{G}}$ is consistent. We then show that under the assumptions in part (iii), there exists a $\mathcal{G} \supset \mathcal{G}_0$ such that $\lim_{n\to\infty} \mathrm{P}(S_{\mathrm{PL}}(\mathcal{G} \,|\, \mathcal{D}) > S_{\mathrm{PL}}(\mathcal{G}_0 \,|\, \mathcal{D})) > 0$, which implies inconsistency of $\widehat{\mathcal{G}}$. Note that if $\mathrm{P}(\mathbf{Z} = 1) = 1$ as in part (i), then $\mathbf{X} \perp\!\!\!\perp \mathbf{Z}$. Therefore, results derived under MCAR hold for all three parts.

Let $\mathcal{G} \supseteq \mathcal{G}_0$. In the sample NAL, we first apply Lemma 7 to replace the MLEs with their population values and use Lemma 8(ii) to eliminate the redundant parents:

$$\bar{\ell}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G})}, \widehat{\Theta}_{X_i}) = \bar{\ell}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G})}, \Theta_{X_i}) + O(n^{-1}) = \frac{1}{|\mathcal{D}_{(i)}^{(\mathcal{G})}|} \sum_{\mathcal{D}_{(i)}^{(\mathcal{G})}} \log \mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}_0)}, \Theta_{X_i}) + O(n^{-1}).$$

The difference between the last expression and $\bar{\ell}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}_0)}, \Theta_{X_i})$, in which $\mathcal{D}_{(i)}^{(\mathcal{G}_0)}$ would appear instead of $\mathcal{D}_{(i)}^{(\mathcal{G})}$, represents the difference in NAL due to $\mathbf{Z}_{(i)}^{(\mathcal{G})} \neq \mathbf{Z}_{(i)}^{(\mathcal{G}_0)}$. We denote it as $d(\mathcal{D}_{(i)}) + O(n^{-1})$ in the following; the contrast between part (i) and parts (ii), (iii) follows from its behaviour for complete and MCAR data.

*Part (i):* For all $\mathcal{G} \supseteq \mathcal{G}_0$, $\bar{\ell}(\mathcal{G}, \Theta \,|\, \mathcal{D}) - \bar{\ell}(\mathcal{G}_0, \Theta_0 \,|\, \mathcal{D}) = O(n^{-1})$. As $\mathcal{G}_0 \subseteq \mathcal{G}_1 \subset \mathcal{G}_2$, it follows that $\bar{\ell}(\mathcal{G}_1, \Theta_1 \,|\, \mathcal{D}) - \bar{\ell}(\mathcal{G}_2, \Theta_2 \,|\, \mathcal{D}) = O(n^{-1})$ and $h(\mathcal{G}_1) < h(\mathcal{G}_2)$. From Lemma 6,

$$\lim_{n\to\infty} \mathrm{P}(S_{\mathrm{PL}}(\mathcal{G}_1 \,|\, \mathcal{D}) > S_{\mathrm{PL}}(\mathcal{G}_2 \,|\, \mathcal{D})) = 1.$$

Hence (C2) is satisfied and part (i) follows.

*Part (ii):* Consider $\mathcal{G} \supset \mathcal{G}_0$: $\Pi_{X_i}^{(\mathcal{G})} \supset \Pi_{X_i}^{(\mathcal{G}_0)}$ for at least one $X_i$ and $\mathbf{Z}_{(i)}^{(\mathcal{G})}$, $\mathbf{Z}_{(i)}^{(\mathcal{G}_0)}$ are potentially different, hence $\mathrm{P}(d(\mathcal{D}_{(i)}) \neq 0) > 0$. Then by (R3) and Balov (2013, L7.4):

$$\sqrt{n} d(\mathcal{D}_{(i)}) \to N(0, \gamma\nu^2),$$

where $\gamma = \left(1 - \mathrm{P}(\mathbf{Z}_{(i)}^{(\mathcal{G})} \mid \mathbf{Z}_{(i)}^{(\mathcal{G}_0)})\right) / \mathrm{P}(\mathbf{Z}_{(i)}^{(\mathcal{G})} \mid \mathbf{Z}_{(i)}^{(\mathcal{G}_0)})$. Reintroducing the $O(n^{-1})$ term and noting that $\sqrt{n}O(n^{-1}) \to 0$,

$$\sqrt{n}\left(\bar{\ell}(X_i \mid \Pi_{X_i}^{(\mathcal{G})}, \widehat{\Theta}_{X_i}) - \bar{\ell}(X_i \mid \Pi_{X_i}^{(\mathcal{G}_0)}, \widehat{\Theta}_{X_i})\right) \to N(0, \gamma\nu^2). \tag{9}$$

As $N(0, \gamma\nu^2)$ is $O(1)$, we obtain that $\bar{\ell}(\mathcal{G}, \Theta \mid \mathcal{D}) - \bar{\ell}(\mathcal{G}_0, \Theta_0 \mid \mathcal{D}) = O(n^{-1/2})$ for all $\mathcal{G} \supset \mathcal{G}_0$. It follows that $\bar{\ell}(\mathcal{G}_1, \Theta_1 \mid \mathcal{D}) - \bar{\ell}(\mathcal{G}_2, \Theta_2 \mid \mathcal{D}) = O(n^{-1/2})$ for all $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{G}$ such that $\mathcal{G}_0 \subseteq \mathcal{G}_1 \subset \mathcal{G}_2$. Applying Lemma 6 as in part (i), we obtain that if $\lim_{n\to\infty} \sqrt{n}\lambda_n = \infty$, then

$$\limsup_{n\to\infty} \mathrm{P}(S_{\mathrm{PL}}(\mathcal{G}_1 \mid \mathcal{D}) > S_{\mathrm{PL}}(\mathcal{G}_2 \mid \mathcal{D})) = 1.$$

Hence (C2) is satisfied and part (ii) follows.

*Part (iii):* By Condition (C3), there exist a $\mathcal{G} \in \mathbb{G}$ and an $X_i$ such that $\Pi_{X_j}^{(\mathcal{G})} = \Pi_{X_j}^{(\mathcal{G}_0)}$ for all $j \neq i$ and $\Pi_{X_i}^{(\mathcal{G})} \supset \Pi_{X_i}^{(\mathcal{G}_0)}$. Let $h(\mathcal{G}) - h(\mathcal{G}_0) = c > 0$. Then $S_{\mathrm{PL}}(\mathcal{G} \mid \mathcal{D}) > S_{\mathrm{PL}}(\mathcal{G}_0 \mid \mathcal{D})$ implies:

$$\begin{aligned} &\sqrt{n}\left(\bar{\ell}(\mathcal{G}, \Theta \mid \mathcal{D}) - \bar{\ell}(\mathcal{G}_0, \Theta_0 \mid \mathcal{D})\right) > \sqrt{n}\lambda_n \left(h(\mathcal{G}) - h(\mathcal{G}_0)\right), \\ &\sqrt{n}\left(\bar{\ell}(X_i \mid \Pi_{X_i}^{(\mathcal{G})}, \widehat{\Theta}_{X_i}) - \bar{\ell}(X_i \mid \Pi_{X_i}^{(\mathcal{G}_0)}, \widehat{\Theta}_{X_i})\right) > \sqrt{n}\lambda_n c. \end{aligned} \tag{10}$$

As in (9), the left-hand side converges to $N(0, \gamma\nu^2)$.

Furthermore, the sequence $\sqrt{n}\lambda_n$ is bounded and we can apply the Bolzano-Weierstrass theorem (Bartle and Sherbert, 2000). There must be a subsequence $\{\sqrt{n}\lambda_n\}_{n'}$ such that $\lim_{n'\to\infty} \sqrt{n'}\lambda_{n'} = \lambda_0 < \infty$. Combining this with (10) and the asymptotic normality above, we have

$$\lim_{n\to\infty} \mathrm{P}(S_{\mathrm{PL}}(\mathcal{G} \mid \mathcal{D}) > S_{\mathrm{PL}}(\mathcal{G}_0 \mid \mathcal{D})) = 1 - \Phi\left(\frac{c\lambda_0}{\sqrt{\gamma\nu^2}}\right) > 0,$$

where $\Phi$ is standard normal CDF. This proves part (iii). ■

For BIC, $n\lambda_n = \log(n)/2 \to \infty$ and $\sqrt{n}\lambda_n = \log(n)/(2\sqrt{n}) \to 0$; for AIC, $n\lambda_n = 1$ and $\sqrt{n}\lambda_n = 1/\sqrt{n} \to 0$. Hence BIC satisfies (i) but not (ii); and AIC does not satisfy either (i) or (ii), confirming and extending (80) in Bozdogan (1987).

Finally, the following corollary justifies the use of NAL in practical BN structure learning (including CGBNs).

**Corollary 11 (C4.2)** *Assume (R1), (R2), (R3) are satisfied. For almost all MCAR distributions* $\mathbf{Z}$, $[\widehat{\mathcal{G}}]$ *is a consistent estimator of* $[\mathcal{G}_0]$ *if and only if* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$.

**Proof of Corollary 11.** From Corollary 4, we have that if $\mathbf{X}$ is MCAR, $[\mathcal{G}_0]$ is identifiable in $\mathbb{G}$. Balov (2013) argues that (C3) holds for almost all distributions of $\mathbf{Z}$ if the set of independence relationships implied by $\mathcal{G}_0$ is non-empty. The result then follows from the proof of Theorem 10, with $\mathcal{G}_0$ replaced by any $\mathcal{G}_* \in [\mathcal{G}_0]$. ■

## 5. Conclusions

Common approaches to BN structure learning from incomplete data, such as the Structural EM, embed EM within score-based algorithms thus incurring in a significant computational cost. NAL is a competitive but faster alternative; in this paper we proved its consistency and model identifiability for CGBNs, showing NAL's wide applicability beyond its original formulation. We also established general sufficient conditions for consistency that can be readily checked for other classes of BNs.

## Appendix A. Proofs

**Proof of Lemma 1.** Let $\mathbf{Z}_{(\mathbf{A})} = \mathbf{Z}_{X_i,\mathbf{A}}$ and $\mathbf{Z}_{(\mathbf{AB})} = \mathbf{Z}_{X_i,\mathbf{A},\mathbf{B}}$. By the law of total probability and $\mathbf{Z_B} \perp\!\!\!\perp X_i, \mathbf{B} \,|\, \mathbf{Z}_{(\mathbf{A})}$: $\mathrm{P}(X_i \,|\, \mathbf{A}, \mathbf{Z}_{(\mathbf{A})}) = \mathrm{E}\left[\mathrm{P}(X_i \,|\, \mathbf{A}, \mathbf{B}, \mathbf{Z}_{(\mathbf{AB})}) \,|\, \mathbf{A}, \mathbf{Z}_{(\mathbf{AB})}\right]$. Since $\mathbf{Z_B} \perp\!\!\!\perp \mathbf{A} \,|\, \mathbf{Z}_{(\mathbf{A})}$, by Jensen's inequality

$$\bar{\ell}(X_i \,|\, \mathbf{A}) = \mathrm{E}\left[\int \mathrm{P}(X_i \,|\, \mathbf{A}, \mathbf{Z}_{(\mathbf{A})}) \log \mathrm{P}(X_i \,|\, \mathbf{A}, \mathbf{Z}_{(\mathbf{A})})\, dX_i \,\middle|\, \mathbf{Z}_{(\mathbf{A})}\right] \leqslant$$
$$\mathrm{E}\left[\int \mathrm{P}(X_i \,|\, \mathbf{A}, \mathbf{B}, \mathbf{Z}_{(\mathbf{AB})}) \log \mathrm{P}(X_i \,|\, \mathbf{A}, \mathbf{B}, \mathbf{Z}_{(\mathbf{AB})})\, dX_i \,\middle|\, \mathbf{Z}_{(\mathbf{AB})}\right] = \bar{\ell}(X_i \,|\, \mathbf{A}, \mathbf{B}),$$

with equality if and only if $X_i \perp\!\!\!\perp \mathbf{B} \,|\, (\mathbf{A}, \mathbf{Z}_{(\mathbf{A})})$. ■

**Proof of Proposition 3.** Under MCAR, $\mathrm{P}_{\mathcal{G}_0}(\mathbf{X} \,|\, \mathbf{Z}) = \mathrm{P}_{\mathcal{G}_0}(\mathbf{X})$ and

$$\bar{\ell}(\mathcal{G}, \Theta) \approx \mathrm{E}\left[\ell(\mathcal{G}, \Theta)\right] = \mathrm{E}\left[\ell(\mathcal{G}_0, \Theta_0)\right] - \mathrm{KL}(\mathrm{P}_{\mathcal{G}_0}(\mathbf{X}) \,||\, \mathrm{P}_{\mathcal{G}}(\mathbf{X})).$$

Hence $\max_{\mathcal{G}\in\mathbb{G}} \ell(\mathcal{G}, \Theta) = \ell(\mathcal{G}_0, \Theta_0)$ with the maximum attained only when the KL vanishes because $\mathrm{P}_{\mathcal{G}_0}(\mathbf{X}) = \mathrm{P}_{\mathcal{G}}(\mathbf{X})$. The maximum can always be attained because $\mathcal{G}_0 \in \mathbb{G}$. This proves (i) and (ii). For (iii), $\Pi_{X_i}^{(\mathcal{G}_0)} \subseteq \Pi_{X_i}^{(\mathcal{G})}$ for all $X_i$ and due to Lemma 1 $\bar{\ell}(\mathcal{G}, \Theta) \geqslant \bar{\ell}(\mathcal{G}_0, \Theta_0)$, while (i) implies $\bar{\ell}(\mathcal{G}, \Theta) \leqslant \bar{\ell}(\mathcal{G}_0, \Theta_0)$. $\bar{\ell}(\mathcal{G}, \Theta) = \bar{\ell}(\mathcal{G}_0, \Theta_0)$ follows. ■

**Proof of Corollary 4.** Let $\mathcal{G}_*$ satisfy $\bar{\ell}(\mathcal{G}_*, \Theta_*) = \max_{\mathcal{G}\in\mathbb{G}} \bar{\ell}(\mathcal{G}, \Theta)$. By (i) of Proposition 3, $\bar{\ell}(\mathcal{G}_*, \Theta_*) = \bar{\ell}(\mathcal{G}_0, \Theta_0)$. Then, by (ii) we have $\mathrm{P}_{\mathcal{G}}(\mathbf{X}) = \mathrm{P}_{\mathcal{G}_0}(\mathbf{X})$. The minimum of all such $\mathcal{G}_*$ is a valid DAG. The result follows from Chickering (1995). ■

**Proof of Proposition 5.** *Condition (R1)*: Let $\mathcal{I}_j$ be the subset of observations for which $\Delta_{X_i} = j$. By assumption, $P(\Delta_{X_i} = j) > 0$, thus this set is non-empty for $n \to \infty$. If $X_i \in \mathbf{\Delta}$, $\widehat{\Theta}_{X_i}$ satisfies $\nabla_{\Theta_{X_i}} \bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) = 0$ and is given by $\widehat{\pi}_{ik\,|\,j} = |\mathcal{I}_j|^{-1} \sum_{\mathcal{I}_j} \mathbb{1}[X_i = k]$. The Lindeberg-Lévy CLT gives $\widehat{\pi}_{ik\,|\,j} = \pi_{ik\,|\,j} + O(n^{-1/2})$ as desired.

Now consider $X_i \in \mathbf{\Gamma}$. From (2), if we let $\mathbf{W} = [\mathbf{1}\ \Gamma_{X_i}]$ conditional on $\mathcal{I}_j$ and we absorb $\mu_{ij}$ into $\boldsymbol{\beta}_{ij}$, the MLEs for $\boldsymbol{\beta}_{ij}$ and $\sigma^2_{ij}$ are the classic

$$\widehat{\boldsymbol{\beta}}_{ij} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T X_i = \beta_k + (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\varepsilon,$$
$$\widehat{\sigma}^2_{ij} = \frac{\widehat{e}^T\widehat{e}}{n} = \frac{\varepsilon^T\varepsilon}{n} - \frac{1}{n}\left(\frac{\mathbf{W}^T\varepsilon}{\sqrt{n}}\right)\left(\frac{\mathbf{W}^T\mathbf{W}}{n}\right)^{-1}\left(\frac{\mathbf{W}^T\varepsilon}{\sqrt{n}}\right).$$

with $\widehat{e} = X_i - \mathbf{W}^T\widehat{\boldsymbol{\beta}}_{ij}$. $(\mathbf{W}^T\mathbf{W}/n)^{-1} \to Q^{-1}$ and $Q$ is invertible under our assumptions. Hence $\widehat{\boldsymbol{\beta}}_{ij}$ and $\widehat{\sigma}^2_{ij}$ satisfy $\nabla_{\Theta_{X_i}} \bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) = 0$, and as $\mathbf{W}$ and $\varepsilon$ are uncorrelated

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{ij} - \boldsymbol{\beta}_{ij}) = \left(\mathbf{W}^T\mathbf{W}/n\right)^{-1}\left(\mathbf{W}^T\varepsilon/\sqrt{n}\right) = O(1) \qquad \text{and} \qquad \sqrt{n}\left(\widehat{\sigma}^2_{ij} - \sigma^2_{ij}\right) = O(1)$$

as shown, for instance, in Heij et al. (2004). (R1) follows.

*Condition (R2)*: For $X_i \in \mathbf{\Delta}$, the diagonal elements of the Hessian are $-1/\pi_{jk}^2$, while the off-diagonal elements are zero. Hence the Hessian is finite for $\pi_{jk} > 0$.

For $X_i \in \mathbf{\Gamma}$, the Hessian of $\boldsymbol{\beta}_{ij}, \sigma_{ij}^2$ for each $\Delta_{X_i} = j$ is

$$-\left(\sigma_{ij}^2\right)^{-2} \begin{bmatrix} \sigma_{ij}^2 \mathbf{W}\mathbf{W}^T & e\mathbf{W}^T \\ e^T\mathbf{W} & (2e^2 - \sigma_{ij}^2)/(2\sigma_{ij}^2) \end{bmatrix},$$

where $e = X_i - \mathbf{W}^T\boldsymbol{\beta}_{ij}$. All elements of this matrix are constants, Gaussian random variables or squares and cross-products thereof. As $\sigma_{ij}^2 > 0$ by assumption, the expectation of all elements of the Hessian is finite and (R2) is satisfied.

*Condition (R3)*: By Jensen's inequality, it suffices to show that $\mathrm{E}[\bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i})^2] < \infty$. If $X_i \in \mathbf{\Delta}$,

$$\mathrm{E}\left[\bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i})^2\right] = \sum\nolimits_{j,k} \pi_{ik\,|\,j} (\log \pi_{ik\,|\,j})^2 < \infty$$

as we are summing a finite number of finite terms under the assumption that all $\pi_{ik\,|\,j} > 0$. For $X_i \in \mathbf{\Gamma}$,

$$\mathrm{E}\left[\bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i})^2\right] = \sum\nolimits_{j} \mathrm{P}(\Delta_{X_i} = j)\, \mathrm{E}\left[(\log \mathrm{P}(X_i \,|\, \Gamma_{X_i}, \Delta_{X_i} = j, \boldsymbol{\beta}_{ij}, \sigma_{ij}^2))^2\right]$$

where the expectations on the right-hand side, disregarding constants, take the form

$$\mathrm{E}\left[\left((X_i - \Gamma_{X_i}\boldsymbol{\beta}_{ij})^2/(2\sigma_{ij}^2)\right)^2\right] < \infty$$

since Gaussian moments are finite. Thus (R3) is satisfied. ■

**Proof of Lemma 6.** The difference in scores, scaled by $n^\alpha$, is

$$n^\alpha(S_{\mathrm{PL}}(\mathcal{G}_1 \,|\, \mathcal{D}) - S_{\mathrm{PL}}(\mathcal{G}_2 \,|\, \mathcal{D})) = n^\alpha(\bar{\ell}(\mathcal{G}_1, \Theta_1 \,|\, \mathcal{D}) - \bar{\ell}(\mathcal{G}_2, \Theta_2 \,|\, \mathcal{D})) + n^\alpha \lambda_n (h(\mathcal{G}_2) - h(\mathcal{G}_1)).$$

The first term is $O(1)$, while the second diverges as $n \to \infty$. The result follows. ■

**Proof of Lemma 7.** If $n \to \infty$, under (R1) and (R2)

$$\bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}) = \bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) + \frac{1}{2}(\Theta_{X_i} - \widehat{\Theta}_{X_i})^T \mathbf{H}_{\Theta_{X_i}}\left(\bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i})\right)(\Theta_{X_i} - \widehat{\Theta}_{X_i}).$$

All elements of the Hessian are bounded by (R2), thus

$$\bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) - \bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}) = O(\|\Theta_{X_i} - \widehat{\Theta}_{X_i}\|^2),$$

which is $O(n^{-1})$. The result follows. ■

**Proof of Lemma 8.** Under MCAR, we can drop $\mathbf{Z}_{(i)}$. If $\mathcal{G} = \mathcal{G}_0$, (i) and (ii) are trivial, while (iii) holds by definition. If $\mathcal{G} \supset \mathcal{G}_0$, there exists an $X_i$ such that $\Pi_{X_i}^{(\mathcal{G}_0)} \subset \Pi_{X_i}^{(\mathcal{G})}$ and for which $X_i \perp\!\!\!\perp \Pi_{X_i}^{(\mathcal{G})} \setminus \Pi_{X_i}^{(\mathcal{G}_0)} \,|\, \Pi_{X_i}^{(\mathcal{G}_0)}$, then (i) follows directly. For (ii), note that $\mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G})}) = \mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}^0)})$. By definition, estimating $\Theta_{X_i}$ minimises the KL divergence between $\mathcal{G}, \Theta$ and $\mathcal{G}_0, \Theta_0$ for the given $\mathcal{D}$, which happens when the corresponding distributions are equal almost surely leading to (ii). We obtain (iii) from (ii) as $\mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G})}, \Theta_{X_i}, \mathcal{D}) = \mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}^0)}, \Theta_{X_i}, \mathcal{D})$

$$\mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}^0)}, \Theta_{X_i}, \mathcal{D}) = \mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}^0)}, \Theta_{X_i}) = \mathrm{P}(X_i \,|\, \Pi_{X_i}^{(\mathcal{G}^0)}, \Theta_{X_i}, \mathbf{Z}_{(i)}).$$

■

**Proof of Lemma 9.** Under MCAR, we can drop the $\mathbf{Z}_{(i)}$. Given (4) and the regularity conditions (R1), (R2), Lemma 7 gives

$$\bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) = \bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}) + O(n^{-1}). \tag{11}$$

As $\bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i})$ is a sum of i.i.d. random variables whose mean exists and is finite, we can apply the Weak Law of Large Numbers to the right hand side of (11). Then

$$\bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) \approx \bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}) - \mathrm{E}[\mathrm{KL}(\mathrm{P}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}) \,||\, \mathrm{P}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}))].$$

$\mathrm{KL}(\cdot) \geqslant 0$ implies that $\lim_{n\to\infty} \mathrm{P}(\bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) \leq \bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i})) = 1$ and (8) follows.

To prove the second assertion, assume that $\mathcal{G}_0 \subseteq \mathcal{G}$. By Lemma 8(iii), we have $\mathrm{P}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}, \mathbf{Z}_{(i)}) = \mathrm{P}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i}, \mathcal{D})$ almost surely. Hence the KL divergence above is zero and $\bar{\ell}(X_i \,|\, \Pi_{X_i}, \widehat{\Theta}_{X_i}) \to_p \bar{\ell}(X_i \,|\, \Pi_{X_i}, \Theta_{X_i})$ with a rate of convergence of $O(n^{-1/2})$ by the Lindeberg-Lévy CLT, dominating the $O(n^{-1})$ in (11). ■

## References


H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716 – 723, 1974.

N. Balov. Consistent Model Selection of Discrete Bayesian Networks from Incomplete Data. *Electronic Journal of Statistics*, 7:1047–1077, 2013.

R. G. Bartle and D. R. Sherbert. *Introduction to Real Analysis*. Wiley, 2000.

M. Beal and Z. Ghahramani. The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. *Bayesian Statistics*, 7:453–464, 2003.

P. Billingsley. *Probability and Measure*. Wiley, 1995.

H. Bozdogan. Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions. *Psychometrika*, 52(3):345–370, 1987.

D. M. Chickering. A Transformational Characterization of Equivalent Bayesian Network Structures. In *UAI*, pages 87–98, 1995.

D. M. Chickering and D. Heckerman. Learning Bayesian Networks is NP-hard. Technical Report MSR-TR-94-17, Microsoft Corporation, 1994.

J. Cussens. Bayesian Network Learning with Cutting Planes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 153–160, 2012.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, pages 1–38, 1977.

L. Fahrmeir and H. Kaufmann. Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *The Annals of Statistics*, pages 342–368, 1985.

N. Friedman. Learning Belief Networks in the Presence of Missing Values and Hidden Variables. In *ICML*, pages 125–133, 1997.

N. Friedman. The Bayesian Structural EM Algorithm. In *UAI*, pages 129–138, 1998.

D. Geiger and D. Heckerman. Learning Gaussian Networks. In *UAI*, pages 235–243, 1994.

F. Glover and M. Laguna. *Tabu search*. Springer, 1998.

D. Heckerman and D. Geiger. Learning Bayesian Networks: a Unification for Discrete and Gaussian Domains. In *UAI*, pages 274–284, 1995.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3):197–243, 1995.

C. Heij, , P. de Boer, P. H. Franses, T. Kloek, and H. K. van Dijk. *Econometric Methods with Applications in Business and Economics*. OUP Oxford, 2004.

S. L. Lauritzen. The EM algorithm for Graphical Association Models with Missing Data. *Computational Statistics & Data Analysis*, 19(2):191–201, 1995.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 1st edition, 1987.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. CRC press, 2nd edition, 1989.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

J. Rissanen. *Information and Complexity in Statistical Models*. Springer, 2007.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

M. Scutari. Dirichlet Bayesian Network Scores and the Maximum Entropy Principle. *Proceedings of Machine Learning Research (AMBN 2017)*, 73:9–20, 2017.

M. Scutari. Bayesian network models for incomplete and dynamic data. *Statistica Neerlandica*, 2020. In print.

M. Scutari, C. E. Graafland, and J. M. Gutiérrez. Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms. *International Journal of Approximate Reasoning*, 115:235–253, 2019.

G. Shafer and P. P. Shenoy. Probability propagation. *Annals of Mathematics and Artificial Intelligence*, 2(1-4):327–351, 1990.

H. White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25, 1982.