Identifiability and Consistency of Bayesian Network Structure Learning from Incomplete Data

Tjebbe Bodewes [1]
tjebbe.bodewes@linacre.ox.ac.uk

Marco Scutari [2]
scutari@idsia.ch

[1] Zivver & Department of Statistics, University of Oxford

[2] Dalle Molle Institute for Artificial Intelligence (IDSIA)

September 24, 2020

Learning a Bayesian network $\mathbf{B} = (\mathcal{G}, \Theta)$ from a data set $\mathcal{D}$ involves:

$$\underbrace{\mathrm{P}(\mathbf{B} \mid \mathcal{D}) = \mathrm{P}(\mathcal{G}, \Theta \mid \mathcal{D})}_{\text{learning}} \quad = \quad \underbrace{\mathrm{P}(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \quad \cdot \quad \underbrace{\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}} .$$

Assuming complete data, we can decompose $\mathrm{P}(\mathcal{G} \mid \mathcal{D})$ into

$$\mathrm{P}(\mathcal{G} \mid \mathcal{D}) \propto \mathrm{P}(\mathcal{G})\, \mathrm{P}(\mathcal{D} \mid \mathcal{G}) = \mathrm{P}(\mathcal{G}) \int \mathrm{P}(\mathcal{D} \mid \mathcal{G}, \Theta)\, \mathrm{P}(\Theta \mid \mathcal{G}) d\Theta$$

where $\mathrm{P}(\mathcal{G})$ is the prior over the space of the DAGs and $\mathrm{P}(\mathcal{D} \mid \mathcal{G})$ is the marginal likelihood (ML) of the data; and then

$$\mathrm{P}(\mathcal{D} \mid \mathcal{G}) = \prod_{i=1}^{N} \left[\int \mathrm{P}(X_i \mid \Pi_{X_i}, \Theta_{X_i})\, \mathrm{P}(\Theta_{X_i} \mid \Pi_{X_i}) d\Theta_{X_i} \right] .$$

where Π_{X_i} are the parents of X_i in $\mathcal{G}$. BIC [9] is often used to approximate $\mathrm{P}(\mathcal{D} \mid \mathcal{G})$. Denote them with $S_{\mathrm{ML}}(\mathcal{G} \mid \mathcal{D})$ and $S_{\mathrm{BIC}}(\mathcal{G} \mid \mathcal{D})$ respectively.

Learning a Bayesian Network from Incomplete Data

When the data are incomplete, $S_{\mathrm{ML}}(\mathcal{G} \mid \mathcal{D})$ and $S_{\mathrm{BIC}}(\mathcal{G} \mid \mathcal{D})$ are no longer decomposable because we must integrate out missing values.

We can use Expectation-Maximisation (EM) [4]:

- in the E-step, we compute the expected sufficient statistics conditional on the observed data using belief propagation [7, 8, 10];
- in the M-step, we use complete-data learning methods with the expected sufficient statistics.

There are two ways of applying EM to structure learning:

- We can apply EM separately to each candidate DAG to be scored, as in the variational-Bayes EM [2].
- 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 [5, 6].

The latter is computationally feasible for medium and large problems, but still computationally demanding.

The Node-Averaged Likelihood

Balov [1] proposed a more scalable approach for discrete BNs called Node-Average Likelihood (NAL). NAL computes each term using the $\mathcal{D}_{(i)} \subseteq \mathcal{D}$ locally-complete data for which X_i, Π_{X_i} are observed:

$$\bar{\ell}(X_i \mid \Pi_{X_i}, \widehat{\Theta}_{X_i}) = \frac{1}{|\mathcal{D}_{(i)}|} \sum_{\mathcal{D}_{(i)}} \log \mathrm{P}(X_i \mid \Pi_{X_i}, \widehat{\Theta}_{X_i}) \to \mathrm{E}\left[\ell(X_i \mid \Pi_{X_i})\right],$$

which Balov used to define

$$S_{\mathrm{PL}}(\mathcal{G} \mid \mathcal{D}) = \bar{\ell}(\mathcal{G}, \Theta \mid \mathcal{D}) - \lambda_n h(\mathcal{G}), \qquad \lambda_n \in \mathbb{R}^+, h : \mathbb{G} \to \mathbb{R}^+$$

and structure learning as $\widehat{\mathcal{G}} = \mathrm{argmax}_{\mathcal{G} \in \mathbb{G}}\, S_{\mathrm{PL}}(\mathcal{G} \mid \mathcal{D})$.

Balov proved both identifiability and consistency of structure learning when using $S_{\mathrm{PL}}(\mathcal{G} \mid \mathcal{D})$ for discrete BNs. We will now prove both properties hold more generally, and in particular that they hold for conditional Gaussian BNs (CGBNs).

Identifiability (General)

Denote the true DAG as $\mathcal{G}_0$ and the equivalence class it belongs to as $[\mathcal{G}_0]$.

Under MCAR, we have:

1. $\max_{\mathcal{G}\in\mathbb{G}} \bar{\ell}(\mathcal{G}, \Theta) = \bar{\ell}(\mathcal{G}_0, \Theta_0)$.
2. If $\bar{\ell}(\mathcal{G}, \Theta) = \bar{\ell}(\mathcal{G}_0, \Theta_0)$, then $\mathrm{P}_{\mathcal{G}}(\mathbf{X}) = \mathrm{P}_{\mathcal{G}_0}(\mathbf{X})$.
3. If $\mathcal{G}_0 \subseteq \mathcal{G}$, then $\bar{\ell}(\mathcal{G}, \Theta) = \bar{\ell}(\mathcal{G}_0, \Theta_0)$.

Identifiability follows from the above.

$[\mathcal{G}_0]$ is identifiable under MCAR, that is

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

Consistency (for CGBNs)

From [1], the sufficient conditions for consistency are:

1. If $\mathcal{G}_0 \subseteq \mathcal{G}_1, \mathcal{G}_0 \nsubseteq \mathcal{G}_2, \lim_{n\to\infty} \mathrm{P}\left(S_{\mathrm{PL}}(\mathcal{G}_1 \mid \mathcal{D}) > S_{\mathrm{PL}}(\mathcal{G}_2 \mid \mathcal{D})\right) = 1.$
2. If $\mathcal{G}_0 \subseteq \mathcal{G}_1, \mathcal{G}_1 \subset \mathcal{G}_2, \lim_{n\to\infty} \mathrm{P}\left(S_{\mathrm{PL}}(\mathcal{G}_1 \mid \mathcal{D}) > S_{\mathrm{PL}}(\mathcal{G}_2 \mid \mathcal{D})\right) = 1.$
3. $\exists\, \mathcal{G} : \Pi_{X_i}^{(\mathcal{G}_0)} \subset \Pi_{X_i}^{(\mathcal{G})}, \Pi_{X_j}^{(\mathcal{G})} = \Pi_{X_j}^{(\mathcal{G}_0)}, \Pi_{X_i}^{(\mathcal{G})} \setminus \Pi_{X_i}^{(\mathcal{G}_0)}$ are neither always observed nor never observed (thus $\mathcal{G}_0$ must not be a maximal DAG).

Under some regularity conditions, we show when they hold for CGBNs:

Let $\mathcal{G}_0$ be identifiable, $\lambda_n \to 0$ as $n \to \infty$, and assume MLEs and NAL's Hessian exist finite. Then as $n \to \infty$:

1. If $n\lambda_n \to \infty$, $\widehat{\mathcal{G}}$ is consistent.
2. Under MCAR and $\mathrm{VAR}(\mathrm{NAL}) < \infty$, if $\sqrt{n}\lambda_n \to \infty$, $\widehat{\mathcal{G}}$ is consistent.
3. Under the above and condition 3, if $\liminf_{n\to\infty} \sqrt{n}\lambda_n < \infty$, then $\widehat{\mathcal{G}}$ is not consistent.

Conclusions

- In $S_{\mathrm{BIC}}(\mathcal{G} \mid \mathcal{D})$, $n\lambda_n = \log(n)/2 \to \infty$ and $\sqrt{n}\lambda_n = \log(n)/(2\sqrt{n}) \to 0$, so BIC satisfies the first condition but not the second in the main result. Hence BIC is consistent for complete data but not for incomplete data.
- The equivalent $S_{\mathrm{AIC}}(\mathcal{G} \mid \mathcal{D})$ does not satisfy either condition which confirms and extends the results in [3]. Hence AIC is not consistent for either complete or incomplete data.
- How to choose λ_n is an open problem.
- Proving results is complicated because
  - $S_{\mathrm{PL}}(\mathcal{G} \mid \mathcal{D})$ is fitted on different subsets of $\mathcal{D}$ for different $\mathcal{G}$, so models are not nested;
  - variables have heterogeneous distributions;
  - DAGs that may represent misspecified models [11] are not representable in terms of $\mathcal{G}_0$ so minimising Kullback-Leibler distances to obtain MLEs does necessarily make them vanish as $n \to \infty$.

Thanks!

Any questions?

References I

- N. Balov.
  Consistent Model Selection of Discrete Bayesian Networks from Incomplete Data.
  Electronic Journal of Statistics, 7:1047–1077, 2013.
- 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.
- H. Bozdogan.
  Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions.
  Psychometrika, 52(3):345–370, 1987.
- 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.
- 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.

References II

- S. L. Lauritzen.
  The EM algorithm for Graphical Association Models with Missing Data.
  Computational Statistics & Data Analysis, 19(2):191–201, 1995.
- J. Pearl.
  Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
  Morgan Kaufmann Publishers Inc., 1988.
- G. Schwarz.
  Estimating the Dimension of a Model.
  The Annals of Statistics, 6(2):461–464, 1978.
- 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.