Using Mixed-Effect Models to Learn Bayesian Networks from Related Data Sets

Marco Scutari[1]
scutari@bnlearn.com

Christopher Marquis[2]
christopher.marquis@epfl.ch

Laura Azzimonti[1]
laura.azzimonti@idsia.ch

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

[2] Ecole Polytechnique Fédérale de Lausanne (EPFL)

October 5, 2022

→ Bayesian Networks

Heterogeneous Data

Using Mixed-Effects Models

Simulation Study

Conclusions and Extensions

Introduction

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

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

What are we assuming when trying to learn a BN? Typically that:

- observations are independent and identically distributed;
- there are no missing values;
- all variables are observed, that is, there are no latent variables introducing confounding in the model.

We revisit the implications of relaxing the first assumption, allowing for the data to be a collation of heterogeneous but related data sets. In our previous work [1], we attacked this problem for discrete data; here we consider continuous data and outline further extensions to hybrid and structured data.

✓ Bayesian Networks

→ Heterogeneous Data

Using Mixed-Effects Models

Simulation Study

Conclusions and Extensions

Homogeneous Data and Heterogeneous Data

If the data are homogeneous, all observations are independent and identically distributed:

$\mathbf{x}_k \sim \mathcal{B}$ with a common BN $\mathcal{B} = (\mathcal{G}, \Theta)$ for all observations $\mathbf{x}_k$.

Instead, we assume that observations belong to $F = 1, \ldots, f$ related data sets that share the same structure but have different parameters:

$\mathbf{x}_{k,j} \sim \mathcal{B}_j$ with $\mathcal{B}_j = (\mathcal{G}, \Theta_j)$ for some $j \in \{1, \ldots, f\}$.

Our aim is to model data that arise from the same generating model (hence the shared structure) but are collected under somehat different conditions, with somewhat different protocols or from somewhat different popuplations (hence the different parameters). This is what we call related data sets.

Complete Pooling: Gaussian BNs

Let's assume that all the variables $\mathbf{X} = \{X_i, i = 1, \ldots, N\}$ in the data are continuous, and that we know which observation belongs to which related data set (that is, F is known for all observations).

One option is to disregard the heterogeneous nature of the data and model them with a Gaussian BN (GBN) over $\mathbf{X}$. This is a complete pooling of the information in the related data sets into a single model, with local distributions:

$$X_i = \mu_i + \mathbf{\Pi}_{X_i}\boldsymbol{\beta}_i + \varepsilon_i, \qquad\qquad \varepsilon_i \sim N(\mathbf{0}, \sigma_i^2\mathbf{I}_n).$$

PROS: the parametrisation is simple, because $\mathbf{X}$ is a multivariate normal distribution and the X_i are linear regression models.

CONS: since we disregard F, the BN is a biased model. All observations are modelled with the same parameters as if they were homogeneous.

No Pooling: Conditional Gaussian BNs

At the oposite end of the spectrum, we could define a conditional Gaussian BN (CGBN) over $\{\mathbf{X}, F\}$ and make F a parent of each X_i so that

$$X_{ij} = \mu_{ij} + \boldsymbol{\Gamma}_{X_i}\boldsymbol{\beta}_{ij} + \varepsilon_{ij},$$
$$\varepsilon_{ij} \sim N(\mathbf{1}, \sigma^2_{ij}\mathbf{I}_{n_j}),\ j = 1, \ldots, f,\ \sum\nolimits_j n_j = n,\ \boldsymbol{\Gamma}_{X_i} = \boldsymbol{\Pi}_{X_i} \cap \mathbf{X}.$$

We have no pooling of information: each related data set is modelled by a separate linear regression model whose parameters are learned only from that data set's observations.

PROS: the BN is not biased since it can model the heterogeneity in the data. The parametrisation is still fairly simple: $\mathbf{X}$ is a mixture of multivariate normals and X_i are mistures of linear regression models.

CONS: the BN does not leverage the assumption that the data sets are related, which implies that the parameters should be similar across them.

✓ Bayesian Networks

✓ Heterogeneous Data

→ Using Mixed-Effects Models

Simulation Study

Conclusions and Extensions

Partial Pooling: Mixed-Effects Models inside CGBNs

In between these two extremes, we choose to use mixed-effects models (LMEs) [2, 3] as the local distributions for the X_i:

$$X_i = \mu_i + \mathbf{\Pi}_{X_i}\boldsymbol{\beta}_i + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i, \quad \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{\Sigma}_i),\ \varepsilon_i \sim N(\mathbf{0}, \sigma_i^2\mathbf{I}_n),$$

where

- $\mathbf{\Pi}_{X_i}$ is the design matrix associated to the $\mathbf{\Gamma}_{X_i}$;
- $\boldsymbol{\beta}_i$ is the vector of fixed effects;
- $\mathbf{Z}_i$ is the design matrix of the random effects, which encodes which observation belongs to which related data set;
- $\mathbf{b}_i$ is the vector of random effects.

LMEs perform a partial pooling of information: they implictly shrink of the parameters associated to the related data sets towards their common average (the fixed effects) based on sample and effect sizes.

Partial Pooling: Mixed-Effects Models inside CGBNs

When used as a local distribution, an LME with random effects for all the parents can be written in the same form as the local distribution of a continuous variable in a CGBN:

$$X_{ij} = (\mu_{ij} + b_{ij0}) + \mathbf{\Pi}_{X_i}(\boldsymbol{\beta}_i + \mathbf{b}_{ij}) + \varepsilon_{ij},$$

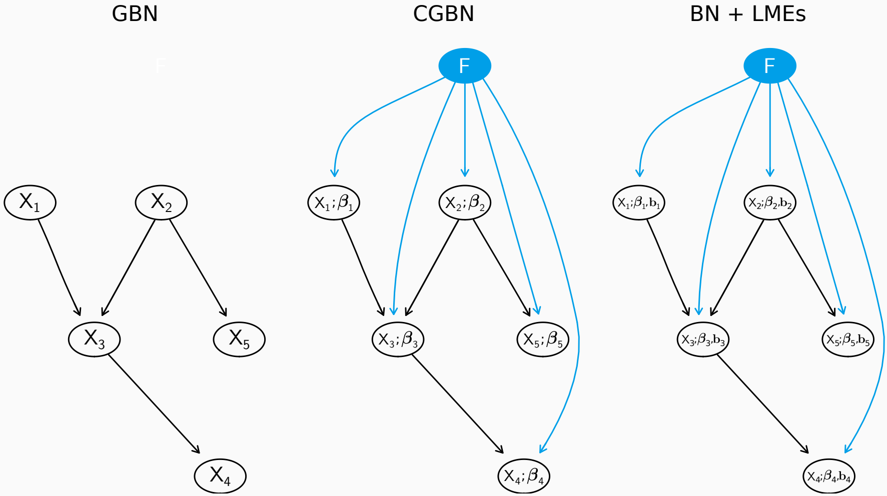
$$\begin{pmatrix} b_{ij0} \\ \mathbf{b}_{ij} \end{pmatrix} \sim N(\mathbf{0}, \tilde{\mathbf{\Sigma}}_i),\ \varepsilon_{ij} \sim N(\mathbf{0}, \sigma_i^2 \mathbf{I}_{n_j}),$$

The random effects represent the deviations of the regression coefficients for the individual related data sets from the fixed effects.

PROS: the BN is unbiased and uses the information in the related data sets to the best effect, pooling information and regularising the parameters through shrinkage.

CONS: learning is somewhat slower since the local distributions are more complex.

GBNs, CGBNs, BNs with LMEs



GBN

CGBN

BN + LMEs

X_1
X_2
X_3
X_5
X_4
F
$X_1;\beta_1$
$X_2;\beta_2$
$X_3;\beta_3$
$X_5;\beta_5$
$X_4;\beta_4$
F
$X_1;\beta_1,\mathbf{b}_1$
$X_2;\beta_2,\mathbf{b}_2$
$X_3;\beta_3,\mathbf{b}_3$
$X_5;\beta_5,\mathbf{b}_5$
$X_4;\beta_4,\mathbf{b}_4$

- ✓ Bayesian Networks
- ✓ Heterogeneous Data
- ✓ Using Mixed-Effects Models
- → **Simulation Study**
- Conclusions and Extensions

Design of the Simulation Study

We studied the properties of this choice of local distributions with a simulation study spanning:

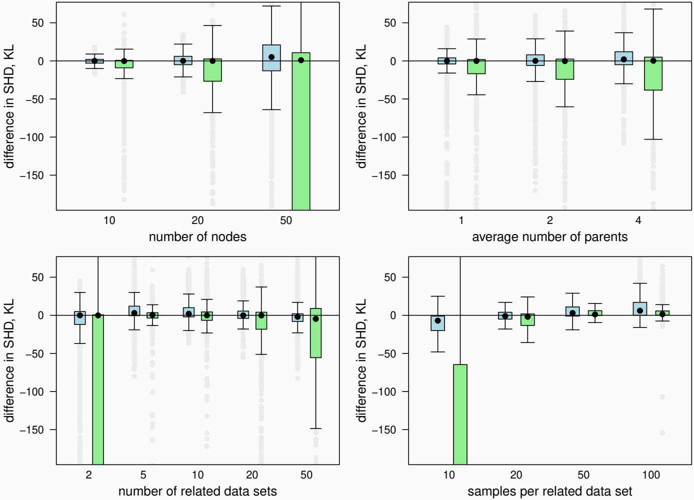
1. 5 DAGs for each combination of $N = 10, 20, 50$, $\overline{|\Pi_{X_i}|} = 1, 2, 4$, $|F| = 2, 5, 10, 20, 50$, and arcs pointing from F to each node.
2. For each DAG, for each X_i and each related data set, we sample

$$\boldsymbol{\beta}_{ij} \sim N(\boldsymbol{\beta}_i + \mathbf{b}_{ij}, \sigma^2_{\boldsymbol{\beta}_{ij}} \mathbf{I}_{|\Pi_{X_i}|+1}),$$
$$\boldsymbol{\beta}_i = \mathbf{2},\ \mathbf{b}_{ij} \sim N(\mathbf{0}, \mathbf{I}_{|\Pi_{X_i}|+1}), \sigma^2_{\boldsymbol{\beta}_{ij}} \sim \chi^2_1$$

   and we set the standard error of the residuals σ^2_{ij} so that the Π_{X_i} explain 85% of the variance of X_i.

We generate 5 data sets for each of $n_j = 10, 20, 50, 100$ from each generating model $\mathcal{B}_{\mathrm{TRUE}}$. For each data set, we learn a GBN on $\mathbf{X}$ ($\mathcal{B}_{\mathrm{GBN}}$, complete pooling), a CGBN on $\{\mathbf{X}, F\}$ ($\mathcal{B}_{\mathrm{CGBN}}$, no pooling) and our LME solution on $\{\mathbf{X}, F\}$ ($\mathcal{B}_{\mathrm{LME}}$, partial pooling). Structure learning is implemented using hill-climbing and BIC.

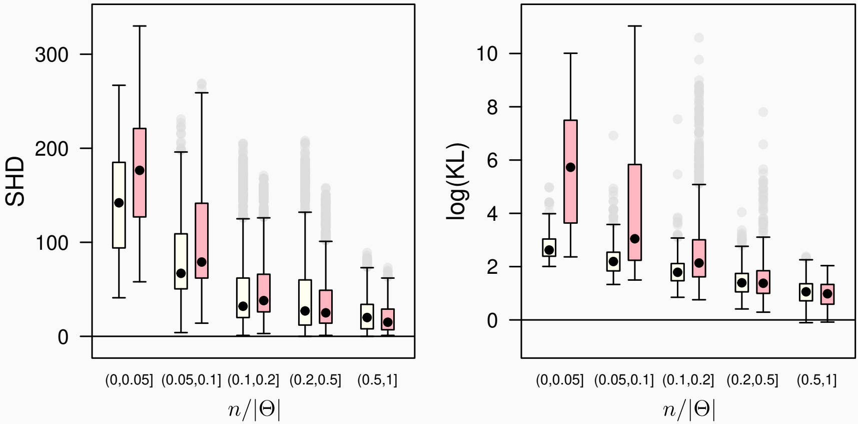
Balanced Data Sets



$\mathrm{SHD}(\mathcal{B}_{\mathrm{LME}}) - \mathrm{SHD}(\mathcal{B}_{\mathrm{CGBN}})$ (blue) and $\mathrm{KL}(\mathcal{B}_{\mathrm{TRUE}}, \mathcal{B}_{\mathrm{LME}}) - \mathrm{KL}(\mathcal{B}_{\mathrm{TRUE}}, \mathcal{B}_{\mathrm{CGBN}})$ (green). Negative values favour $\mathcal{B}_{\mathrm{LME}}$.

$\mathcal{B}_{\mathrm{LME}}$ have lower SHD and KL than $\mathcal{B}_{\mathrm{GBN}}$ for 95% of the data sets.

Balanced and Unbalanced Data Sets

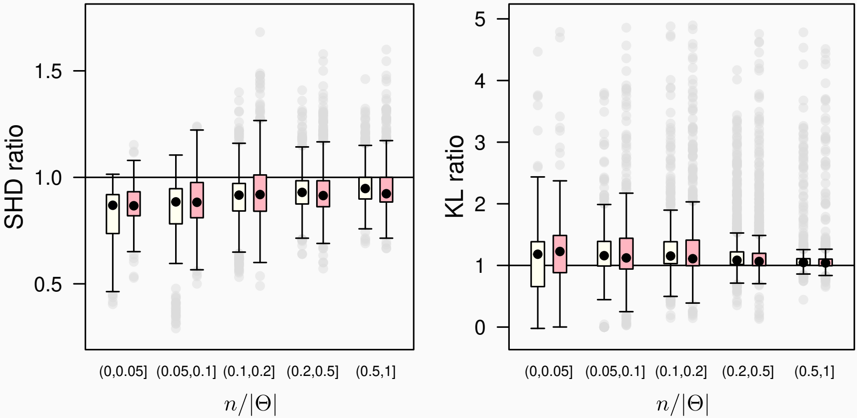


$\mathcal{B}_{\mathrm{LME}}$ (ivory) and the $\mathcal{B}_{\mathrm{CGBN}}$ (pink) compared to $\mathcal{B}_{\mathrm{TRUE}}$.

The more the related data sets have unbalanced sample sizes, the more $\mathcal{B}_{\mathrm{LME}}$ dominates $\mathcal{B}_{\mathrm{CGBN}}$. This is the result of pooling information between related data sets: those that have fewer observations borrow information from the larger ones, thus allowing us to learn more accurate models. $\mathcal{B}_{\mathrm{CGBN}}$ do not perform any pooling and therefore perform increasingly poorly.

Homogeneous Data

What if there is no heterogeneity in the data, but we use $\mathcal{B}_{\mathrm{LME}}$ anyway? In this scenario, $\mathcal{B}_{\mathrm{GBN}}$ is the correctly specified model while both $\mathcal{B}_{\mathrm{LME}}$ and $\mathcal{B}_{\mathrm{CGBN}}$ are over-parametrised.



$\mathcal{B}_{\mathrm{LME}}$ (ivory) and the $\mathcal{B}_{\mathrm{CGBN}}$ (pink) compared to $\mathcal{B}_{\mathrm{GBN}}$.

✓ Bayesian Networks

✓ Heterogeneous Data

✓ Using Mixed-Effects Models

✓ Simulation Study

→ Conclusions and Extensions

Conclusions and Extensions

- The automatic pooling of information between the related data sets results in BNs with better structural and parametric accuracy for small sample sizes and unbalanced data sets.
- Using LMEs as local distributions outperforms GBNs. It is at least as good as using standard CGBNs in terms of SHD, and outperforms CGBNs in terms of KL.
- LMEs perform well even when the data are homogeneous.
- We can do more!
    - We can drop the assumption that F is a parent of all other nodes.
    - Generalised LMEs as local distributions can handle a diverse set of distributions.
    - LMEs can model more complex structures in the data: cryptic relatedness (in genetics), spatial dependencies, temporal dependencies.

Thanks!

Any questions?

References I

- L. Azzimonti, G. Corani, and M. Scutari.
  A Bayesian Hierarchical Score for Structure Learning from Related Data Sets.
  International Journal of Approximate Reasoning, 142:248–265, 2021.
- E. Demidenko.
  Mixed Models: Theory and Applications with R.
  Wiley, 2nd edition, 2009.
- J. C. Pinheiro and D. M. Bates.
  Mixed-effects models in S and S-PLUS.
  Springer, 2000.