Bayesian Network Models for Continuous-Time and Structured Data

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence (IDSIA)

January 19, 2023

→ Bayesian Networks: Definition and Assumptions

Continuous-Time Bayesian Networks

Bayesian Networks for Structured Data

Future Directions

A Graph and a Probability Distribution

A Bayesian network (BN) [20] is defined by:

- a network structure, a directed acyclic graph $\mathcal{G}$ in which each node corresponds to a random variable X_i;
- a global probability distribution $\mathbf{X}$ with parameters Θ, which can be factorised into smaller local probability distributions according to the arcs present in $\mathcal{G}$.

The main role of the network structure is to express the conditional independence relationships among the variables in the model through graphical separation, thus specifying the factorisation of the global distribution:

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid \Pi_{X_i}; \Theta_{X_i}) \quad \text{where} \quad \Pi_{X_i} = \{\text{parents of } X_i \text{ in } \mathcal{G}\}\,.$$

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

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

Structure learning consists in finding the DAG with the best

$$\mathrm{P}(\mathcal{G} \mid \mathcal{D}) \propto \underbrace{\mathrm{P}(\mathcal{G})}_{\text{graph prior}} \cdot \underbrace{\mathrm{P}(\mathcal{D} \mid \mathcal{G})}_{\text{marginal likelihood}} = \mathrm{P}(\mathcal{G}) \int \mathrm{P}(\mathcal{D} \mid \mathcal{G}, \Theta)\, \mathrm{P}(\Theta \mid \mathcal{G})\, d\Theta$$

which is known as score-based learning [10]. The alternative, constraint-based learning, uses tests following Pearl's work on causality [24]:

$$\underbrace{X_i \perp\!\!\!\perp_P X_j \mid \mathbf{S}_{X_i, X_j}}_{\text{conditional independence}} \Longrightarrow \underbrace{X_i \perp\!\!\!\perp_G X_j \mid \mathbf{S}_{X_i, X_j}}_{\text{graphical separation}}.$$

Parameter learning consists in estimating the parameter sets $\Theta_{X_i} \mid \Pi_{X_i}$.

The Classic Definition and Modern Extensions

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

- observations are independent and there are no missing values;
- all variables are observed, that is, there are no latent variables introducing confounding in the model;
- we measure probabilistic associations (or rather, independencies) and we cannot necessarily interpret them as causal.

What happens if we relax these assumptions? Many extensions suddenly become possible, see [19] for a recent review. In this talk we will discuss:

- Learning BNs from continuous-time dynamic data [5].
- Learning BNs from data in which data have structure, such as state-space data and collations of related data sets [2, 21].

We will not discuss learning BNs from incomplete data, but we are making progress on that front as well [4].

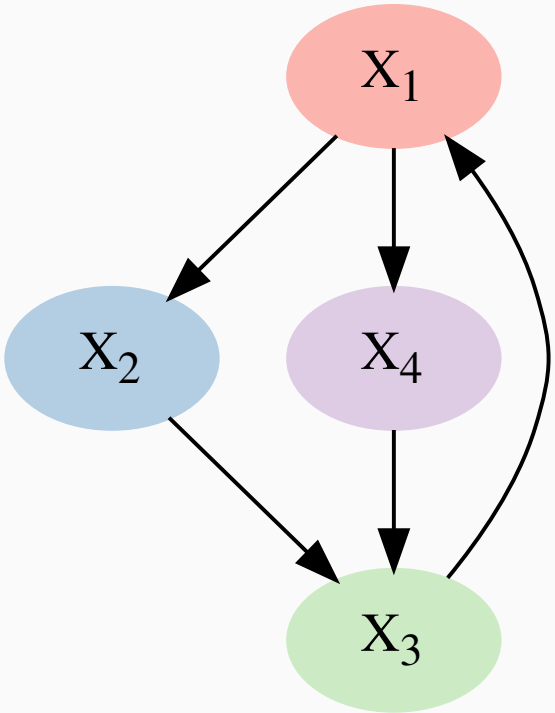
- ✓ Bayesian Networks: Definition and Assumptions
- → Continuous-Time Bayesian Networks
- Bayesian Networks for Structured Data
- Future Directions

Continuous-Time Bayesian Networks

Continuous-Time BNs (CTBNs) are a framework for modelling finite-state, continuous-time processes. Their graphical representation allows for natural, cyclic dependency graphs without having to specify a temporal granularity [17].

A CTBN consists of two components:

- A directed graph encoding conditional independencies.
- A conditional intensity matrix (CIM) $\mathbf{Q}_{X_i \mid \mathbf{u}}$ describing the evolution process of a variable with the parameters
  - $\mathbf{q}_{X_i}$: a set of intensities parameterising the exponential distributions over when the next transition occurs.
  - $\boldsymbol{\theta}_{X_i}$: a set of probabilities parameterising the distribution over where the state transitions.



Constraint-Based Structure Learning?

Score-based learning was covered by Nodelman [17] in his original work on CTBNs. For constraint-based structure learning, we need a new definition of conditional independence [5]:

> Let $\mathcal{N}$ be a CTBN with a graph $\mathcal{G}$ over $\mathbf{X}$. We say that $X_i \perp\!\!\!\perp X_j \mid \mathbf{S}_{X_i,X_j}$ if $\mathbf{Q}_{X_i \mid x,\mathbf{s}} = \mathbf{Q}_{X_i \mid \mathbf{s}}$ for all values x, s of X_j and $\mathbf{S}_{X_i,X_j}$.

Note that conditional independence is not symmetric in CTBNs! To test it we need to test two separate hypotheses:

- Time To Transition: independence of the waiting times ($\mathbf{q}_{X_i}$), tested with an F test to compare their exponential distributions.
- State-to-State Transition: independence of the transitions ($\boldsymbol{\theta}_{X_i}$), tested with a two-sample χ^2 test or a Kolmogorov-Smirnov test.

We test time-to-transition hypothesis first and then, if the null is rejected, the state-to-state hypotheses. If both nulls are rejected, X_i and X_j are conditionally independent.

Time to Transition [3]: given the exponential waiting times $q_{x|\mathbf{s}}, q_{x|y,\mathbf{s}}$,

$$H_0 : \frac{q_{x|\mathbf{s}}}{q_{x|y,\mathbf{s}}} = 1 \qquad \text{with null } F_{r_a, r_b}$$

where $r_a = \sum_{x' \in X_i} M_{xx'|y,\mathbf{s}}$ and $r_b = \sum_{x' \in X_i} M_{xx'|\mathbf{s}}$.

State-to-State Transition [15]: given $\theta_{x|\mathbf{s}}, \theta_{x|y,\mathbf{s}}$,

$$H_0 : \theta_{x|\mathbf{s}} = \theta_{x|y,\mathbf{s}} \quad \text{with null} \quad \chi^2 = \sum_{x' \in X_i} \frac{(K \cdot M_{xx'|y,\mathbf{s}} - L \cdot M_{xx'|\mathbf{s}})^2}{M_{xx'|\mathbf{s}} + M_{xx'|y,\mathbf{s}}}$$

where $K = \sqrt{\frac{\sum_{i=1}^{k} M_{xx'|\mathbf{s}}}{\sum_{i=1}^{k} M_{xx'|y,\mathbf{s}}}}$ and $L = \frac{1}{K}$.

We reject the (conditional) independence between the two nodes if at least one null hypothesis is rejected.

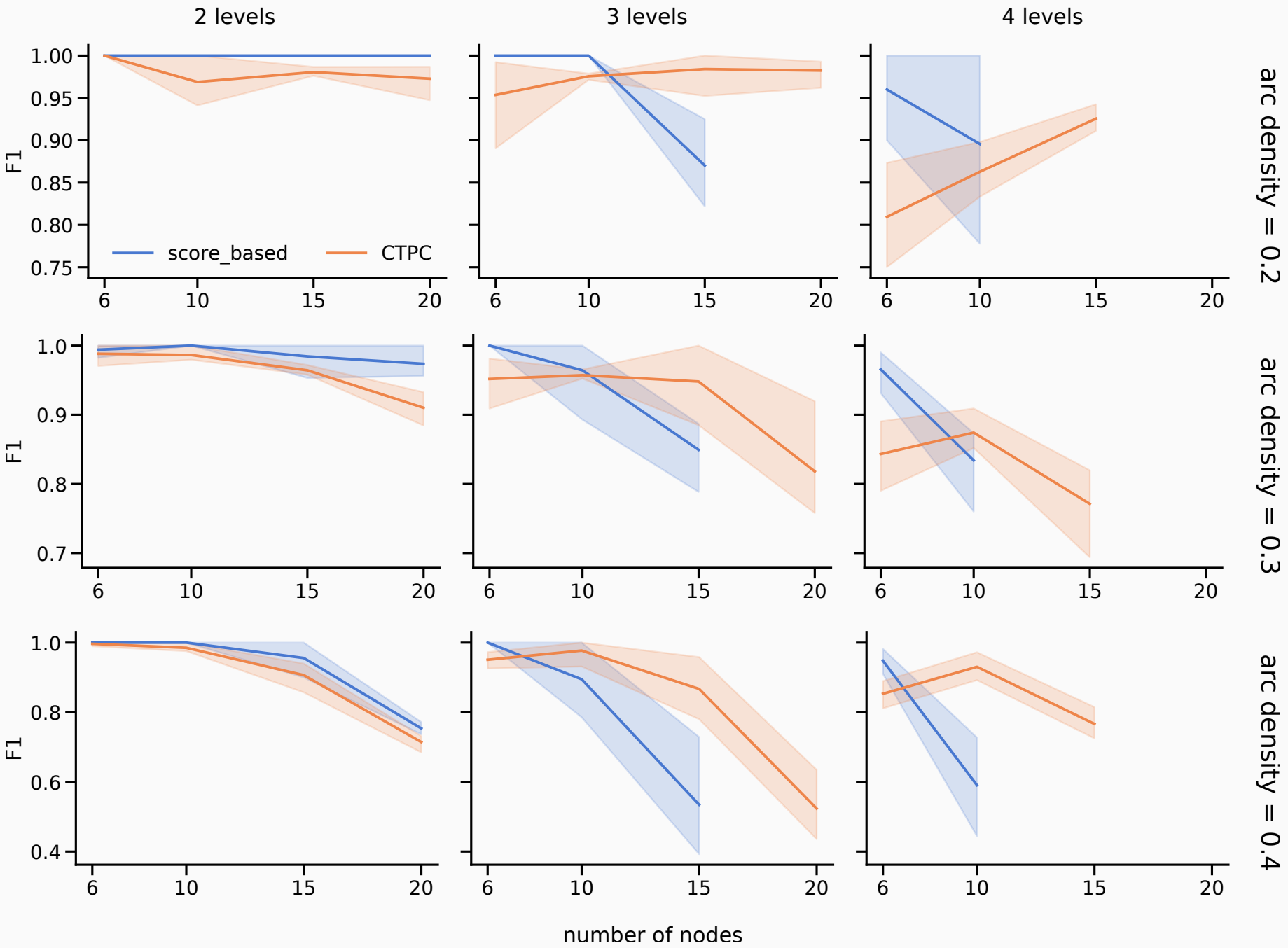
A PC Algorithm for Continuous-Time Bayesian Networks

Given how different the definition of conditional independence is, we need to adapt the PC algorithm [6] to match.

1. Form a complete directed graph $\mathcal{G}$ over $\mathbf{X}$.
2. For each variable X_i:
   - 2.1 Set $\mathbf{U} = \{X_j \in \mathbf{X} : X_j \to X_i\}$, the current parent set.
   - 2.2 For increasing values $b = 0, \ldots, |\mathbf{U}|$:
     - 2.2.1 For each $X_j \in \mathbf{U}$, test $X_i \perp\!\!\!\perp X_j \mid \mathbf{S}_{X_i, X_j}$ for all possible subsets of size b of $\mathbf{U} \setminus X_j$.
     - 2.2.2 As soon as $X_i \perp\!\!\!\perp X_j \mid \mathbf{S}_{X_i, X_j}$ for some $\mathbf{S}_{X_i, X_j}$, remove $X_j \to X_i$ from $\mathcal{G}$ and X_j from $\mathbf{U}$.
3. Return $\mathcal{G}$.

We call this the Continuous-Time PC (CTPC) algorithm [5]. It has better structural accuracy than the score-based approach in [17], but both approaches are slow: they are only practical for less than 20 variables.

CTPC Versus Score-Based Learning

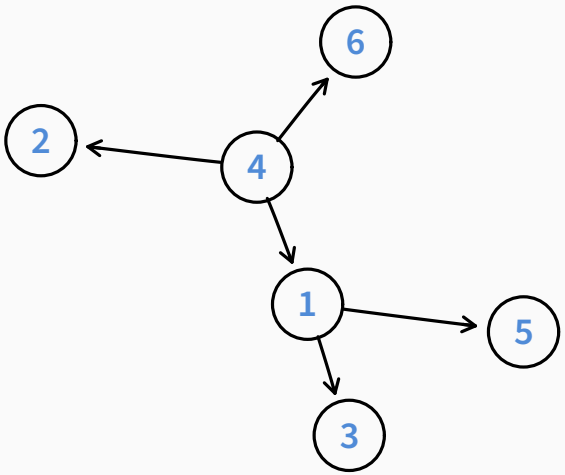


- ✓ Bayesian Networks: Definition and Assumptions
- ✓ Continuous-Time Bayesian Networks
- → **Bayesian Networks for Structured Data**
- Future Directions

Network models broadly fall into two groups:

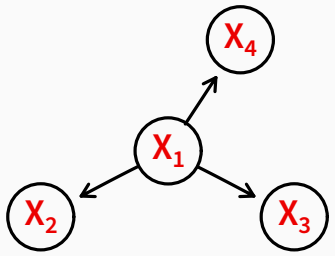
- Social networks in which nodes are associated with individuals and arcs represents their similarity, measured on the variables.
- Graphical models in which nodes are associated with the variables in the data and the arcs represent probabilistic associations measured on independent observations.

Social Network
(the structure of the data)



	X_1	X_2	X_3	X_4
1	5.25	4.14	1.45	5.25
2	4.10	0.70	2.49	6.26
3	7.11	7.96	3.15	8.54
4	0.38	8.45	7.62	4.71
5	7.03	6.46	2.45	4.07
6	1.09	3.98	1.09	2.50

Graphical Model
(the structure of the variables)



We often want to model both perspectives at the same time:

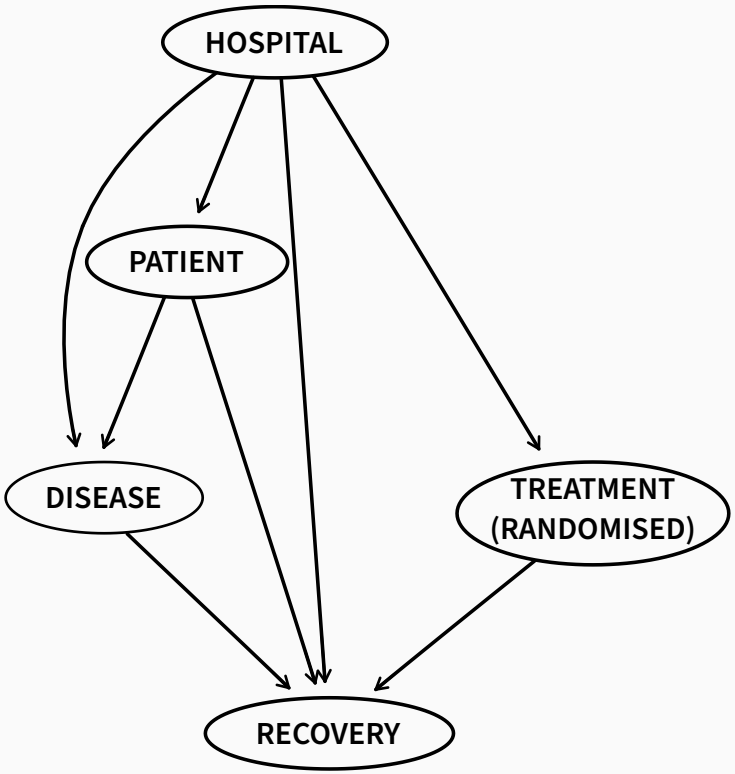
- In causal networks to measure the average treatment effect from multi-centre clinical trials, individuals treated in the same hospital and the treatment they receive are more homogeneous than those in different hospitals.
- In gene networks and multivariate genomic association models, similarity between the genotypes of individuals implies a similarity in their phenotypical traits.
- In state-space data, individuals close to each other in space are more likely to exhibit similar behaviour; and longitudinal measurements are more strongly associated within than between individuals.

Failure to properly account for the similarity between individuals artificially inflates the strength of the apparent relationships between the variables, resulting in dense biased networks.

Multi-Centre Clinical Trials: Related Data Sets

Hospitals produce separate data sets which are then collated together for the analysis. Inevitable differences in their implementations of the clinical trial make those data sets related but not identical.
We want to learn the BN as a hierarchical model that separates the shared average effect encoded by each arc from hospital-specific effects, pooling information across hospitals.

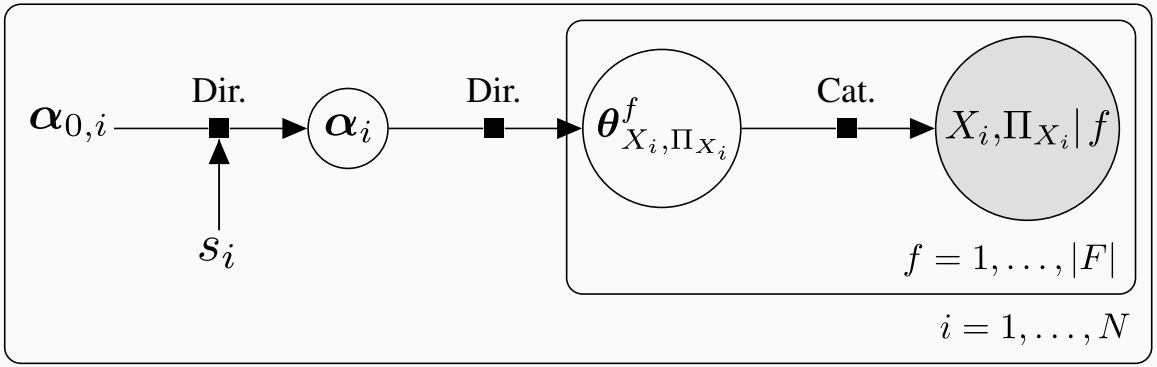
Assumptions: the structure of the BN is the same for all hospitals, but the parameters differ between hospitals. The assignment of each individual to hospital is known.



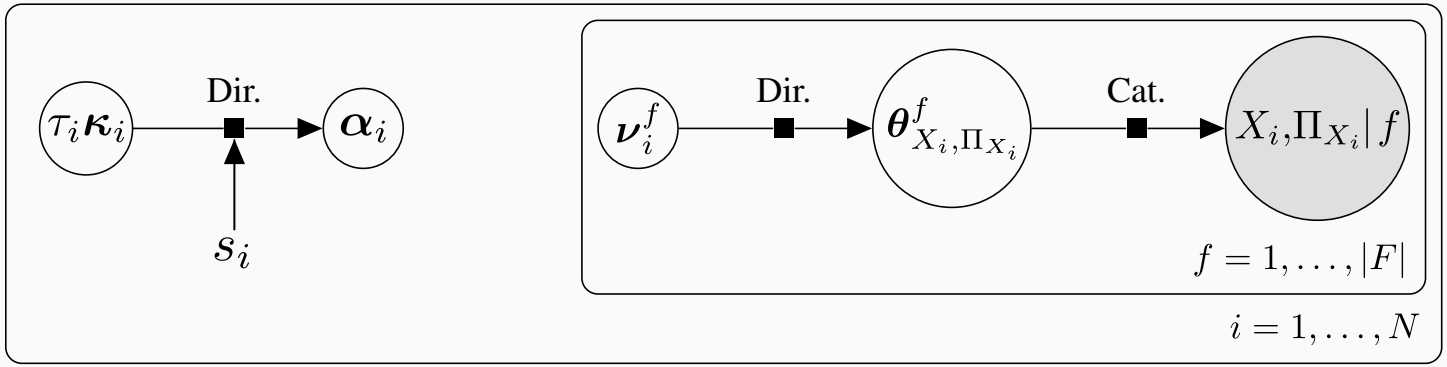
The mathematical formulation:

- a variational Bayesian score with a hierarchical prior [2];
- using mixed-effects models [21].

A Variational Bayesian Score for Discrete Variables



Hierarchical Model



Variational Approximation

Thus we get the Bayesian Hierarchical Dirichlet (BHD) score:

$$\mathrm{P}(\mathcal{D} \mid F, \mathcal{G}) \approx \prod_{i=1}^{N} \prod_{f=1}^{|F|} \prod_{j=1}^{|\Pi_{X_i}|} \left[\frac{\Gamma(s_i \hat{\kappa}_{ij})}{\Gamma(s_i \hat{\kappa}_{ij} + n^f_{ij})} \prod_{k=1}^{|X_i|} \frac{\Gamma(s_i \hat{\kappa}_{ijk} + n^f_{ijk})}{\Gamma(s_i \hat{\kappa}_{ijk})} \right]$$

where $s_i \hat{\kappa}_{ijk}$ = the posterior mean of α_{ijk} under the variational model.

The BHD score:

- has better structural accuracy when we are modelling related data sets;
- it gets increasingly better as the number of related grows;
- it gets increasingly better as the size of (at least some of) the individual related data sets grows.

However, this approach is not flexible because we need a separate set of mathematical derivations for each structured-data scenario and for each type of random variable. In this respect, a better alternative is to use mixed-effects models (LMEs) [7, 18] as the local distributions for the X_i:

- generalised linear mixed-effects models (GLMMs) can model all types of variables in the exponential family (binomial, multinomial, Poisson, Gaussian, etc.) but
- in practice they make us reintroduce some linearity assumptions.

What Does That Look Like?

In a Gaussian BN, each node X_i has distribution

$$X_i = \mu_{X_i} + \Pi_{X_i}\beta_{X_i} + \varepsilon_{X_i} \qquad \text{with} \qquad \varepsilon_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I}_n). \quad (1)$$

Adding a grouping node F like clinical trial centres would make it a conditional Gaussian BN in which we fit a separate linear regression for each data set j identified by F:

$$X_i = \mu_{ij} + \Pi_{X_i}\beta_{ij} + \varepsilon_{X_i} \qquad \text{with} \qquad \varepsilon_{X_i} \sim N(0, \sigma^2_{ij}\mathbf{I}_{n_j}). \quad (2)$$

A mixed-effects model that takes (1) and adds random effects for all Π_{X_i}

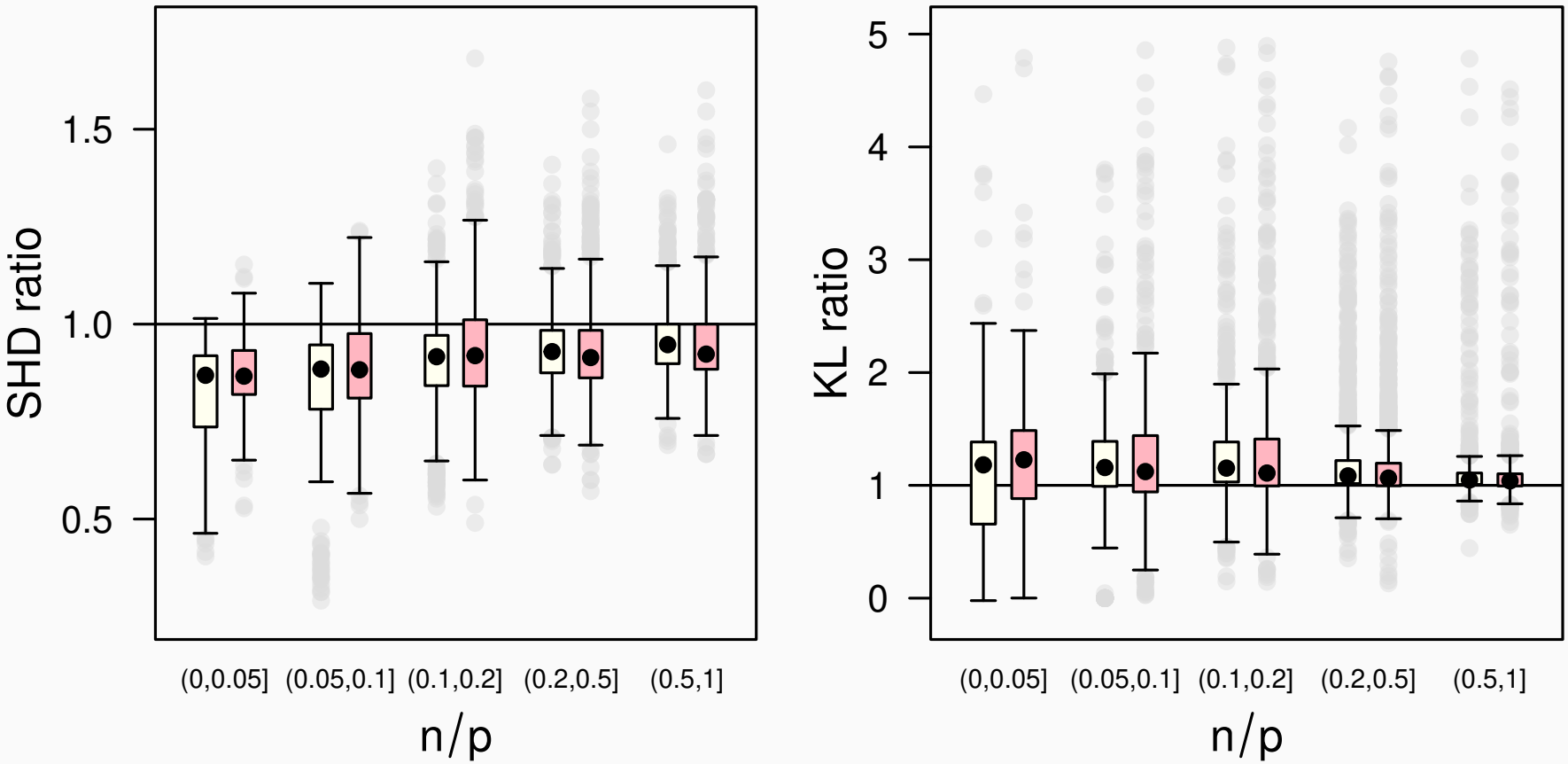
$$X_i = \mu_{X_i} + \Pi_{X_i}\beta_{X_i} + \mathbf{Z}\mathbf{b}_{X_i} + \varepsilon_{X_i},\ \mathbf{b}_{X_i} \sim N(\mathbf{0}, \Sigma), \varepsilon_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I}_n)$$

has the same form as (2),

$$X_i = (\mu_{ij} + b_{0j}) + \Pi_{X_i}(\beta_{X_i} + \mathbf{b}_{ij}) + \varepsilon_{X_i},$$

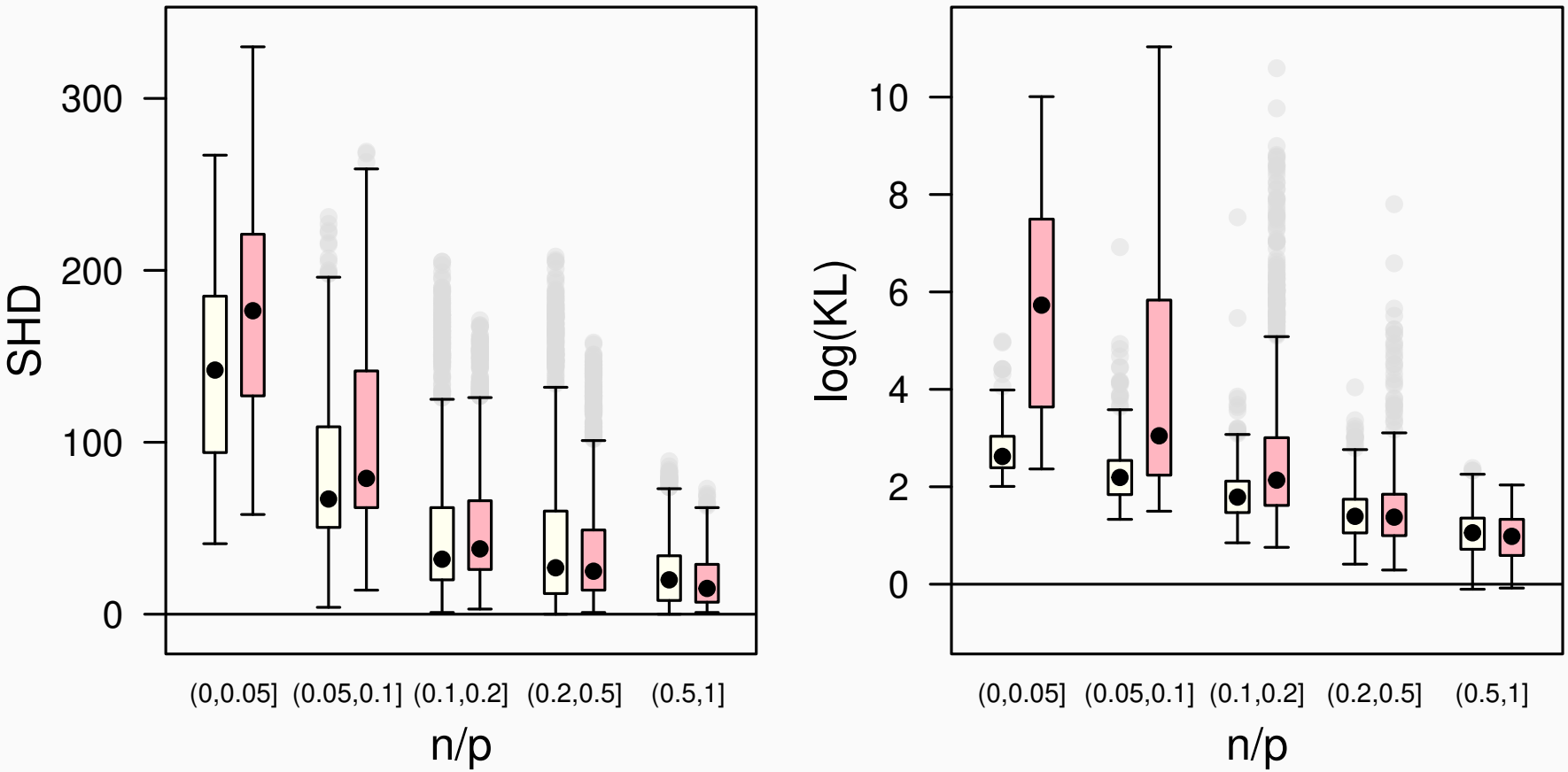
but pools information across data sets much like BHD does [21].

Pooling Versus No Pooling: Homogeneous Data



If the data are just a single homogeneous data set, introducing mixed effects does not degrade performance.

Pooling Versus No Pooling: Heterogeneous Data



If the data really are a collation of related data sets, introducing mixed effects improves both structural (SHD) and parametric accuracy (KL). The difference becomes more marked if the related data sets are unbalanced.

Statistical genetics has long used the random effects in mixed-effects models to encode population structure (that is, related individuals), using pedigrees or genomic data [1]. In the context of BNs, the foundational Genomic BLUP model becomes

$$X_i = \mu_{X_i} + \Pi_{X_i}\boldsymbol{\beta}_{X_i} + \mathbf{g} + \varepsilon_{X_i},\ \mathbf{g} \sim N(0, \mathbf{K}), \varepsilon_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I}_n)$$

where the kinship matrix $\mathbf{K}$ encodes the relatedness. If $\mathbf{K} \propto \mathbf{G}\mathbf{G}^T$, with $\mathbf{G}$ the matrix of the genotypes, this model is equivalent to a ridge regression, which is in turn equivalent to a random-effects model [9, 23]:

$$X_i = \mu_{X_i} + \Pi_{X_i}\boldsymbol{\beta}_{X_i} + \mathbf{G}\mathbf{b}_{X_i} + \varepsilon_{X_i},\ \varepsilon_{X_i} \sim N(0, \sigma^2_{X_i}\mathbf{I}_n).$$

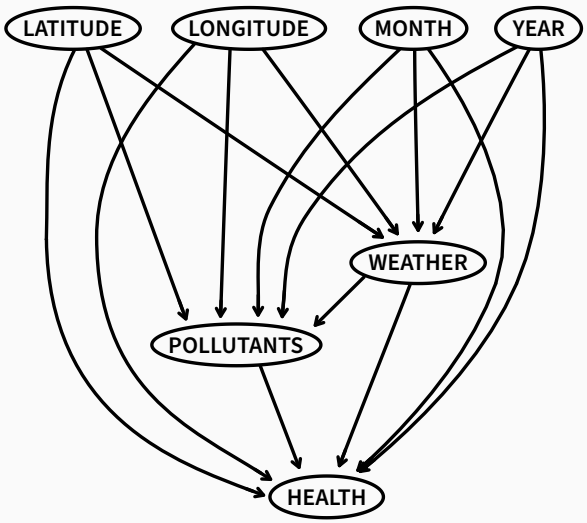
Hence we can add the random effects $\mathbf{G}\mathbf{b}_{X_i}$ to a local regression as soon as one of the genotypes is a parent of X_i to implicitly incorporate kinship.

(Or we can give up modelling individual genotypes and do it like [13].)

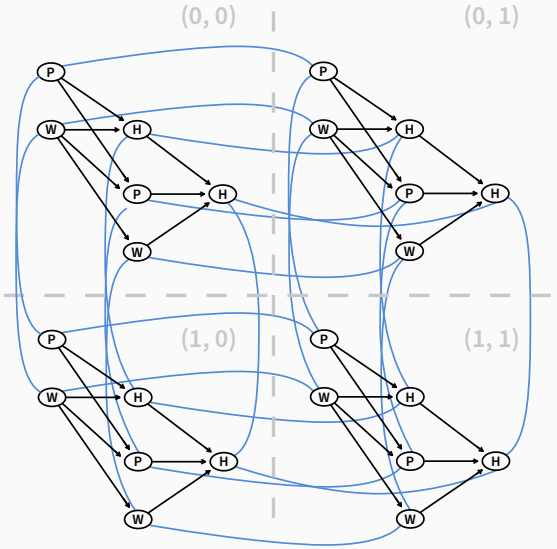
Structured Data over Time and Space

Dynamic BNs can model temporal data as vector auto-regressive processes by duplicating nodes across time points [16], spatial networks over a grid [12]. However, this is impractical when observations are irregularly spaced and spread over time. A mixed-effects model can incorporate spatial and temporal autocorrelation into local distributions using random effects similarly to kinship in statistical genetics models.

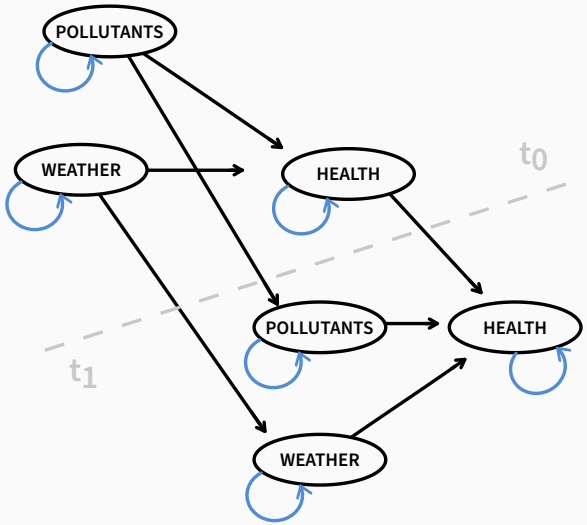
A Naive Static BNs
(data are independent across time and space)



Nodes Replicated in Time and Space
(too many nodes and arcs)



Dynamic BN with Spatial Correlation
(an interpretable causal network)



- ✓ Bayesian Networks: Definition and Assumptions
- ✓ Continuous-Time Bayesian Networks
- ✓ Bayesian Networks for Structured Data
- → **Future Directions**

Conclusions and Future Directions

Bayesian networks are a fundamental tool in machine learning: they subsume many models [19] and handle incomplete data [4], continuous-time time series [5] and collections of related data sets [2].

What next?

- Making CTBNs into Markov decision processes [11, 22] to model as streaming health data where we administer medical treatments in real time.
- Incorporating all the computational tricks used in the statistical genetics literature [8, 14, 26] to speed up learning.
- A reanalysis of a complex environmental data set such as [25] to explore BNs with a spatio-temporal structure.

Acknowledgements

Christopher Marquis
École Polytechnique Fédérale de Lausanne (EPFL)

Alessandro Bregoli
Fabio Stella
Università degli Studi di Milano-Bicocca

Søren Wengel Mogensen
Lunds Universitet

Laura Azzimonti
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)

Thanks!

Any questions?

References I

- W. Astle and D. J. Balding.
  Population Structure and Cryptic Relatedness in Genetic Association Studies.
  Statistical Science, 24(4):451–471, 2009.
- 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.
- L. Bain and M. Englehardt.
  Statistical Analysis of Reliability and Life-Testing Models: Theory and Methods.
  CRC Press, 1991.
- T. Bodewes and M. Scutari.
  Learning Bayesian Networks from Incomplete Data with the Node-Averaged Likelihood.
  International Journal of Approximate Reasoning, 138:145–160, 2021.
- A. Bregoli, M. Scutari, and F. Stella.
  A Constraint-Based Algorithm for the Structural Learning of Continuous-Time Bayesian Networks.
  International Journal of Approximate Reasoning, 138:105–122, 2021.
- D. Colombo and M. H. Maathuis.
  Order-Independent Constraint-Based Causal Structure Learning.
  Journal of Machine Learning Research, 15:3921–3962, 2014.

References II

- E. Demidenko.
  Mixed Models: Theory and Applications with R.
  Wiley, 2nd edition, 2009.
- J. H. Sul H. M. Kang, S. K. Service, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin.
  Variance Component Model to Account for Sample Structure in Genome-Wide Association Studies.
  Nature Genetics, 42:348–254, 2010.
- D. Habier, R. L. Fernando, and J. C. M. Dekkers.
  The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values.
  Genetics, 177:2389–2397, 2007.
- D. Heckerman and D. Geiger.
  Learning Bayesian Networks: a Unification for Discrete and Gaussian Domains.
  In *UAI*, pages 274–284, 1995.
- K. F. Kan and C. R. Shelton.
  Solving Structured Continuous-Time Markov Decision Processes.
  In *ISAIM*, 2008.
- C. Krapu, R. Stuart, and A. Rose.
  A Review of Bayesian Networks for Spatial Data.
  ACM Transactions on Spatial Algorithms and Systems, pages 1–21, 2022.

References III

- W. Kruijer, P. Behrouzi, D. Bustos-Korts, M. X. Rodríguez-Álvarez, S. M. Mahmoudi, B.Yandell, E. Wit, and F. A. van Eeuwijk.
  Reconstruction of Networks with Direct and Indirect Genetic Effects.
  Genetics, 214(4):781–807, 2020.
- C. Lippert, J. Listgarten, Y. Liu, C. M. Cadie, R. I. Davidson, and D. Heckerman.
  FaST Linear Mixed Models for Genome-Wide Association Studies.
  Nature Methods, 8(10):833–837, 2011.
- B. Mitchell.
  A Comparison of Chi-Square and Kolmogorov-Smirnov Tests.
  The Royal Geographical Society, 3(4):237–241, 1971.
- K. Murphy.
  Dynamic Bayesian Networks: Representation, Inference and Learning.
  PhD thesis, UC Berkeley, Computer Science Division, 2002.
- U. D. Nodelman.
  Continuous Time Bayesian Networks.
  PhD thesis, Stanford University, 2007.
- J. C. Pinheiro and D. M. Bates.
  Mixed-effects models in S and S-PLUS.
  Springer, 2000.

References IV

- M. Scutari.
  Bayesian Network Models for Incomplete and Dynamic Data.
  Statistica Neerlandica, 74(3):397–419, 2020.
- M. Scutari and J.-B. Denis.
  Bayesian Networks with Examples in R.
  Chapman & Hall, 2nd edition, 2021.
- M. Scutari, C. Marquis, and L. Azzimonti.
  Using Mixed-Effect Models to Learn Bayesian Networks from Related Data Sets.
  Proceedings of Machine Learning Research (PGM 2022), 2022.
- L. Sturlaugson, L.Perreault, and J. W. Sheppard.
  Factored Performance Functions and Decision Making in Continuous Time Bayesian Networks.
  Journal of Applied Logic, 22:28–45, 2017.
- P.M. VanRaden.
  Efficient Methods to Compute Genomic Predictions.
  J. Dairy Sci., 91(11):4414–4423, 2008.
- T. S. Verma and J. Pearl.
  Equivalence and Synthesis of Causal Models.
  In *UAI*, pages 255–268, 1990.

References V

- C. Vitolo, M. Scutari, A. Tucker, and A. Russell.
  Modelling Air Pollution, Climate and Health Data Using Bayesian Networks: a Case Study of the English Regions.
  Earth and Space Science, 5(4):76–88, 2018.
- X. Zhou and M. Stephens.
  Genome-Wide Efficient Mixed-Model Analysis for Association Studies.
  Nature Genetics, 44:821–824, 2012.