Challenges in Bayesian Network Modelling of Climate and Weather Data

Marco Scutari
scutari@idsia.ch

Dalle Molle Institute for
Artificial Intelligence (IDSIA)

November 6, 2019

Natural Systems are Complex Systems

Natural phenomena can only be modelled as complex systems in which

- there are many components that interact with each other;
- their interplay produces non-obvious behaviour;
- they develop over time and space in response to the surrounding environment.

Two scientific research fields in which this has increasingly become apparent are environmental sciences and biological sciences (genetics, systems biology, etc.).

Classic statistical models that focus on explaining or predicting a single component of such phenomena often fail to capture the big picture. Network models, on the other hand, focus on capturing the interplay between components from a systems perspective, without necessarily restricting their attention to a single one.

Bayesian networks (BNs) [9] implement this systems approach with:

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

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:

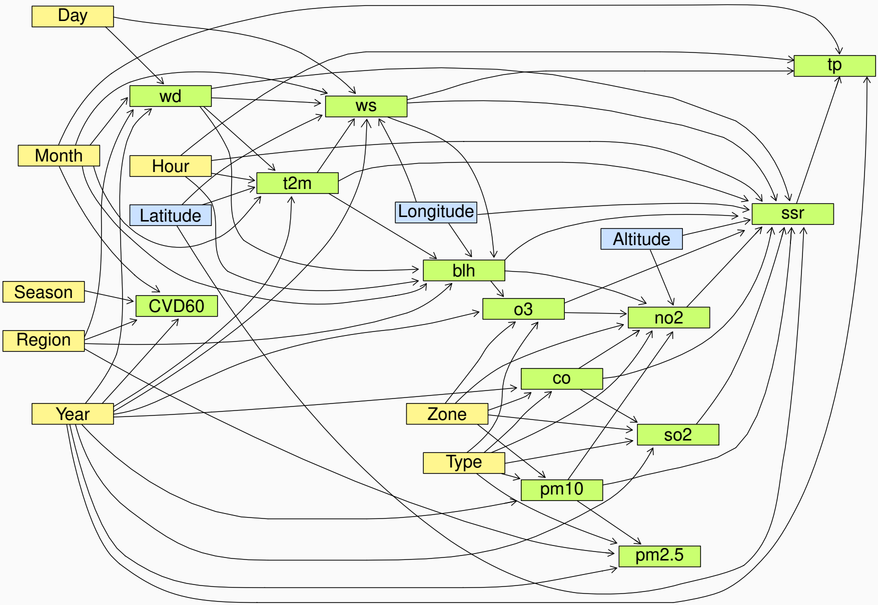
$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid \Pi_{X_i}; \Theta_{X_i}) \quad \text{where} \quad \Pi_{X_i} = \{\text{parents of } X_i\}.$$

Why Use Bayesian Networks?

Four main reasons:

- Both the network structure and the parameters can be learned efficiently from data [18]; and available prior information can be incorporated in the learning process as well [2, 13, 4].
- The network structure provides a high-level qualitative view of the phenomenon that can easily be used by non-statisticians.
- Automated reasoning can quantify the probability of any event of interest given available evidence using standard algorithms.
- With some additional assumptions BNs can be interpreted as causal models [14].

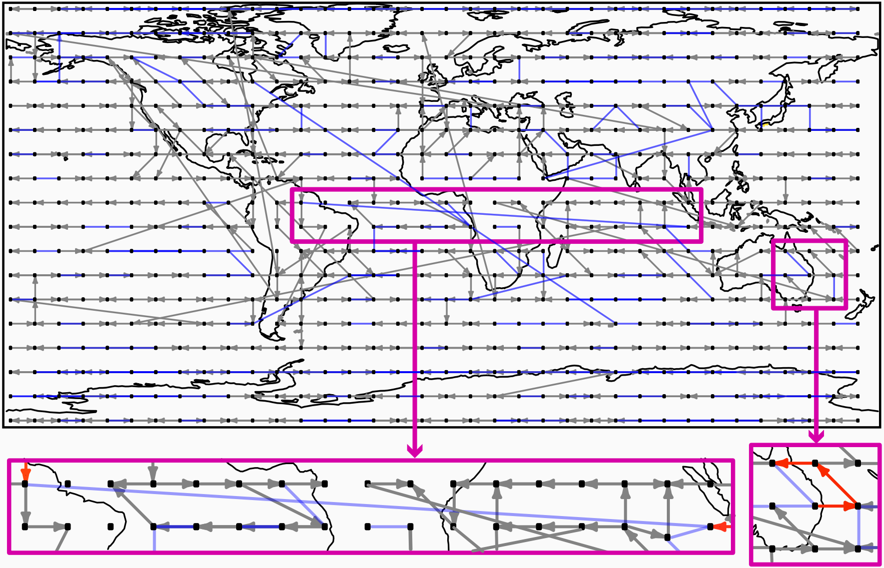
Several applications in environmental sciences: studying species dynamics [1, 19]; the impact of climate change on groundwater [12]; how to best manage water reservoirs under infrequent rainfalls [15]; the effects of El Niño [17]; and the impact of pollution [20].


Day
tp
wd
ws
Month
Hour
t2m
Latitude
Longitude
ssr
Altitude
blh
Season
CVD60
o3
no2
Region
co
Year
Zone
so2
Type
pm10
pm2.5

C. Vitolo, M. Scutari, M. Ghalaieny, A. Tucker and A. Russell (2018). "Modeling Air Pollution, Climate, and Health Data Using Bayesian Networks: A Case Study of the English Regions." Earth and Space Science, 5(4), 76–88. [20]

- Almost 50 million records spanning the period 1981–2014.
- 24 features: various air pollutants (O3, $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO) measured in 162 monitoring stations, their geographical characteristics (latitude, longitude, latitude, region and zone type), weather (wind speed and direction, temperature, rainfall, solar radiation, boundary layer height), demography and mortality rates.
- The model represents known processes in atmospheric chemistry with a good degree of accuracy.

Climate Data Analysis



M. Scutari, C. E. Graafland and J. M. Gutiérrez (2019). “Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms.” International Journal of Approximate Reasoning, 115:235–253. [17]

- Monthly surface temperature values on a global 10°-resolution regular grid from 1981 to 2010.
- Local dependencies are strong since they are the result of the short-term evolution of atmospheric thermodynamic processes. Distant teleconnected dependencies resulting from large-scale atmospheric oscillation patterns are in general weaker, but they are key for understanding regional climate variability.
- Altered probabilities of high temperatures in the Indian Ocean when El Niño-like evidence is introduced in the BN.

Assumptions and Limitations of Bayesian Networks

Two assumptions that are typically made in BN learning are particularly problematic:

- Complete Data: the data contain no missing values.
- Independent Observations: observations are jointly independent of each other.

Other common assumptions that may be problematic:

- Categorical variables are multinomial, continuous variables are Gaussian or mixtures of Gaussians.
- The network is sparse, with a number of arcs comparable to the number of nodes.

The computational complexity of learning can also be an issue: linear in the sample size but quadratic in the number of variables (and that is assuming the network is sparse).

We can learn the network structure from incomplete data using a variation of the EM algorithm called Structural EM [5, 6]:

- in the E-step, we complete the data by computing the expected sufficient statistics using the current network structure;
- in the M-step, we find the structure that maximises the expected likelihood or posterior probability for the completed data.

The parameters can be learned with the classic EM [10].

However:

- The Structural EM is extremely computationally intensive; the shortcuts used in practical implementations void its theoretical guarantees.
- There is no literature on this for continuous or hybrid data, only for categorical data.
- Data are assumed to be missing (completely) at random.

Take the Spatio-Temporal Structure of the Data into Account

For instance, the local distribution of a Gaussian variable with continuous parents is assumed to be

$$X_i = \mu_{X_i} + \Pi_{X_i}\beta_{X_i} + \varepsilon_{X_i}, \qquad \varepsilon_{X_i} \sim N\left(0, \Sigma_{X_i}\right), \Sigma_{X_i} = \sigma^2_{X_i}\mathrm{I}_n;$$

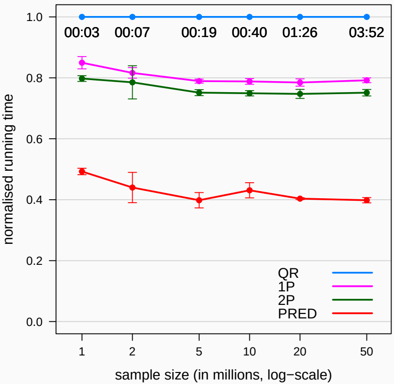
all the parameter estimators and goodness-of-fit scores are borrowed from classic linear regression.

The logical solution would be to use an appropriate covariance structure [3] such as an isotropic exponential structure

$$\Sigma_{X_i} = [\sigma_{jk}] \qquad\qquad \sigma_{jk} = \sigma^2 exp\left\{-d_{jk}/\theta\right\}$$

instead of $\sigma_{jk} = 0$ for all $j \neq k$. It comes at a cost in terms of speed, but it is feasible unlike the MCMC approaches for state space models such as [7].

- Many algorithms display embarrassing or coarse-grained parallelism [16].
- There are many approaches in statistical genetics that optimise sequential linear model evaluation [11], including for correlated observations.
- For discrete data, there are efficient data structures that can be leveraged [8].



(Classic closed-form results can help too [18]!)

Conclusions and Remarks

- BNs are naturally suited to modelling complex systems as networks.
- BNs have several key advantages: they can incorporate prior information while learning them from data; they are easy to interpret for non-statisticians; and they allow automated and causal reasoning.
- Their fundamental assumptions must be weakened to improve their usability in environmental sciences, to handle incomplete and spatio-temporal data effectively.
- Computational complexity is also an issue, but there is literature to draw from for inspiration.

Acknowledgements

Catharina Elisabeth Graafland
José Manuel Gutiérrez
Institute of Physics of Cantabria (CSIC-UC)

Allan Tucker
Andrew Russell
Mohamed Ghalaieny
Brunel University London

Claudia Vitolo
European Centre for Medium-Range Weather Forecasts

Thanks!

Any questions?

- A. Aderhold, D. Husmeier, J. J. Lennon, C. M. Beale, and V. A. Smith.
  Hierarchical Bayesian Models in Ecology: Reconstructing Species Interaction Networks from Non-Homogeneous Species Abundance Data.
  Ecological Informatics, 11:55–64, 2012.
- R. Castelo and A. Siebes.
  Priors on Network Structures. Biasing the Search for Bayesian Networks.
  International Journal of Approximate Reasoning, 24(1):39–57, 2000.
- P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger.
  Analysis of Longitudinal Data.
  Oxford University Press, 2nd edition, 2013.
- M. J. Druzdzel and L. C. van der Gaag.
  Elicitation of Probabilities for Belief Networks: Combining Qualitative and Quantitative Information.
  In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 141–148, 1995.
- N. Friedman.
  Learning Belief Networks in the Presence of Missing Values and Hidden Variables.
  In *Proceedings of the 14th International Conference on Machine Learning*, pages 125–133, 1997.

- N. Friedman.
  The Bayesian Structural EM Algorithm.
  In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 129–138, 1998.
- I. D. Jonsen, R. A. Myers, and J. M. Flemming.
  Meta-Analysis of Animal Movement Using State-Space Models.
  Ecology, 84(11):3055–3063, 2003.
- S. Karan, M. Eichhorn, B. Hurlburt, G. Iraci, and J. Zola.
  Fast Counting in Machine Learning Applications.
  In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pages 540–549, 2018.
- D. Koller and N. Friedman.
  Probabilistic Graphical Models: Principles and Techniques.
  MIT Press, 2009.
- S. L. Lauritzen.
  The EM Algorithm for Graphical Association Models with Missing Data.
  Computational Statistics and Data Analysis, 19(2):191–201, 1995.
- 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.

- J.-L. Molina, D. Pulido-Velázquez, J. L. García-Aróstegui, and M. Pulido-Velázquez.
  Dynamic Bayesian Networks as a Decision Support Tool for Assessing Climate Change Impacts on Highly Stressed Groundwater Systems.
  Journal of Hydrology, 479:113–129, 2013.
- S. Mukherjee and T. P. Speed.
  Network Inference Using Informative Priors.
  Proceedings of the National Academy of Sciences, 105(38):14313–14318, 2008.
- J. Pearl and D. Mackenzie.
  The Book of Why: the New Science of Cause and Effect.
  Basic Books, 2018.
- R. F. Ropero, M. J. Flores, R. Rumí, and P. A. Aguilera.
  Applications of Hybrid Dynamic Bayesian Networks to Water Reservoir Management.
  Environmetrics, 28:e2432, 2017.
- M. Scutari.
  Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimised Implementations in the bnlearn R Package.
  Journal of Statistical Software, 77(2):1–20, 2017.
- 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.

- M. Scutari, C. Vitolo, and A. Tucker.
  Learning Bayesian Networks from Big Data with Greedy Search: Computational Complexity and Efficient Implementation.
  Statistics and Computing, 25(9):1095–1108, 2019.
- N. Trifonova, A. Kenny, D. Maxwell, D. Duplisea, J. Fernandes, and A. Tucker.
  Spatio-Temporal Bayesian Network Models with Latent Variables for Revealing Trophic Dynamics and Functional Networks in Fisheries Ecology.
  Ecological Informatics, 30:142–158, 2015.
- 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.