Challenges in Bayesian Network Modelling of Climate and Weather Data

Marco Scutari
scutari@idsia.ch

Dalle Molle Institute for Artificial Intelligence

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 $P(X)$ with parameters $\Theta$, 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:

$$P(X) = \prod_{i=1}^{N} P(X_i \mid \Pi_{X_i}; \Theta_{X_i})$$

where $\Pi_{X_i} = \{\text{parents of } X_i\}$. 
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].
Almost 50 million records spanning the period 1981–2014.

24 features: various air pollutants (O₃, PM₂.₅, PM₁₀, SO₂, NO₂, 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

- **Monthly surface temperature values** on a global $10^\circ$-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.
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).
Learning from Incomplete Data

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

$$X_i = \mu X_i + \prod X_i \beta X_i + \varepsilon X_i, \quad \varepsilon X_i \sim N(0, \Sigma X_i), \Sigma X_i = \sigma^2 X_i I_n;$$

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

$$\Sigma X_i = [\sigma_{jk}], \quad \sigma_{jk} = \sigma^2 \exp \left\{-\frac{d_{jk}}{\theta}\right\}$$

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].
Improve Computational Efficiency

- 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]!)

![Graph showing normalised running time vs. sample size for different methods.]
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!


N. Friedman. 
The Bayesian Structural EM Algorithm. 


S. Karan, M. Eichhorn, B. Hurlburt, G. Iraci, and J. Zola. 
Fast Counting in Machine Learning Applications. 

D. Koller and N. Friedman. 
*Probabilistic Graphical Models: Principles and Techniques*. 

S. L. Lauritzen. 
The EM Algorithm for Graphical Association Models with Missing Data. 

C. Lippert, J. Listgarten, Y. Liu, C. M. Cadie, R. I. Davidson, and D. Heckerman. 
FaST Linear Mixed Models for Genome-Wide Association Studies. 


M. Scutari, C. Vitolo, and A. Tucker. 
Learning Bayesian Networks from Big Data with Greedy Search: Computational Complexity and Efficient Implementation. 

Spatio-Temporal Bayesian Network Models with Latent Variables for Revealing Trophic Dynamics and Functional Networks in Fisheries Ecology. 

Modelling Air Pollution, Climate and Health Data Using Bayesian Networks: a Case Study of the English Regions. 