Modelling Survey Data with Bayesian Networks

Marco Scutari

scutari@stats.ox.ac.uk
Department of Statistics
University of Oxford

May 18, 2015

Bayesian Networks

Bayesian networks (BNs) [6, 13] are defined by:

- a network structure, a directed acyclic graph $\mathcal{G} = (\mathbf{V}, A)$, in which each node $v_i \in \mathbf{V}$ corresponds to a random variable X_i;
- a global probability distribution, $\mathbf{X}$, which can be factorised into smaller local probability distributions according to the arcs $a_{ij} \in A$ 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:

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{p} \mathrm{P}(X_i \mid \Pi_{X_i}) \qquad \text{where} \qquad \Pi_{X_i} = \{\text{parents of } X_i\}$$

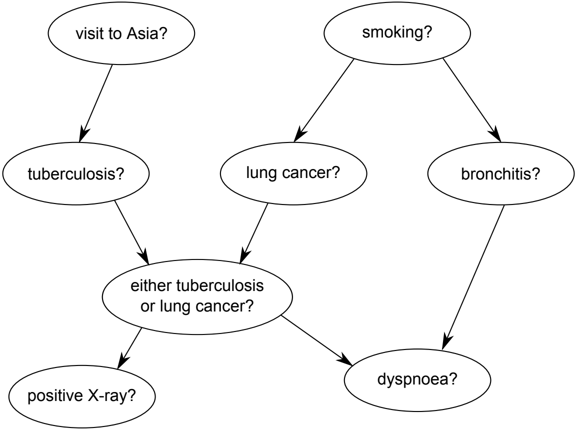
Discrete Bayesian Networks

In discrete BNs all X_i are defined to be either categorical or ordinal variables, and the parameters of interest are grouped in conditional probability tables (CPTs).

	$x_{i(1)}$	$\cdots$	$x_{i(p)}$	
$\Pi_{X_i(1)}$	π_{11}	$\cdots$	π_{1p}	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\Pi_{X_i(k)}$	π_{k1}	$\cdots$	π_{kp}	1

If the variables are ordinal, X_i and X_j are considered dependent if there is a trend, e.g. the levels of the first increase (decrease) as the levels of the second increase (decrease).

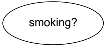
An Example: The ASIA Network (Global Distribution)



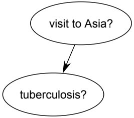
Lauritzen SL and Spiegelhalter DJ (1988). [7]

An Example: The ASIA Network (Local Distributions)

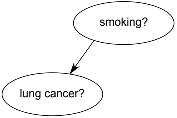
yes	no
0.01	0.99



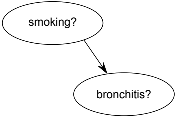
yes	no
0.50	0.50



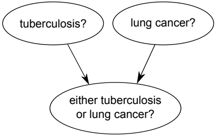
	yes	no
yes	0.05	0.95
no	0.01	0.99



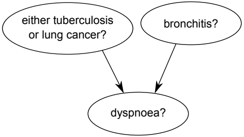
	yes	no
yes	0.10	0.90
no	0.01	0.99



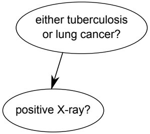
	yes	no
yes	0.60	0.40
no	0.30	0.70



	yes	no
yes:yes	1	0
yes:no	1	0
no:yes	1	0
no:no	0	1



	yes	no
yes:yes	0.90	0.10
yes:no	0.70	0.30
no:yes	0.80	0.20
no:no	0.10	0.90



	yes	no
yes	0.98	0.02
no	0.05	0.95

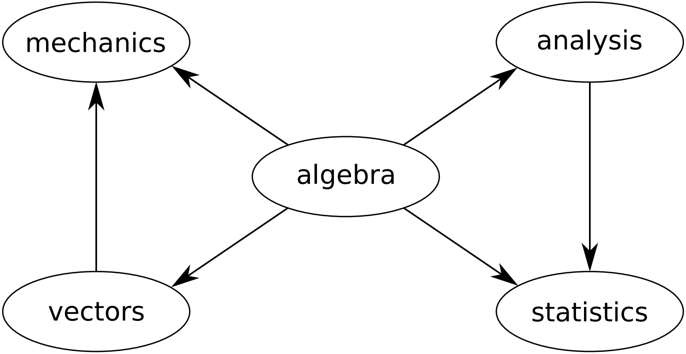
Continuous (Gaussian) Bayesian Networks

In continuous BNs the global distribution is assumed to be multivariate normal and the local distributions are univariate normals with independent variances. If we further assume that all dependencies are linear, the BN describes a hierarchical linear regression model with

$$X_i = \mu + X_{j_1}\beta_1 + \ldots + X_{j_k}\beta_k + \varepsilon_i \qquad \text{with} \qquad \varepsilon_i \sim N(0, \sigma_i^2).$$

As an extension of the above, hybrid BNs also include discrete variables which make the BN behave as a mixture or a random effects model.

An Example: The Marks Network



Mardia KV, Kent JT and Bibby JM (1979) [10] and Whittaker J (1990). [16]

An Example: The Marks Network (Local Distributions)

$$\mathsf{ALG} = 50.60 + \varepsilon_{\mathsf{ALG}} \sim N(0, 10.62^2)$$

$$\mathsf{ANL} = -3.57 + 0.99\mathsf{ALG} + \varepsilon_{\mathsf{ANL}} \sim N(0, 10.50^2)$$

$$\mathsf{MECH} = -12.36 + 0.54\mathsf{ALG} + 0.46\mathsf{VECT} + \varepsilon_{\mathsf{MECH}} \sim N(0, 13.97^2)$$

$$\mathsf{STAT} = -11.19 + 0.76\mathsf{ALG} + 0.31\mathsf{ANL} + \varepsilon_{\mathsf{STAT}} \sim N(0, 12.60^2)$$

$$\mathsf{VECT} = 12.41 + 0.75\mathsf{ALG} + \varepsilon_{\mathsf{VECT}} \sim N(0, 10.48^2)$$

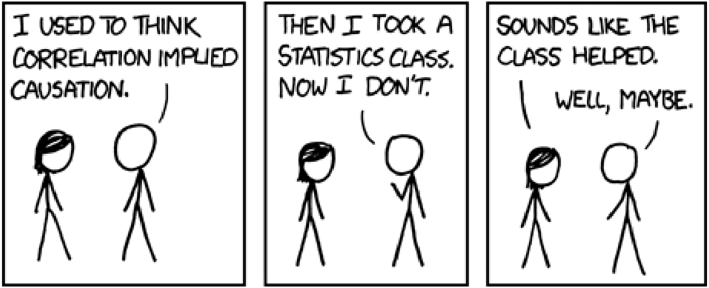
Causal Interpretation of Bayesian Networks

> *It seems that if conditional independence judgments are byproducts of stored causal relationships, then tapping and representing those relationships directly would be a more natural and more reliable way of expressing what we know or believe about the world. This is indeed the philosophy behind causal BNs.*
>
> *Judea Pearl [14]*

This is the reason why building a BN from expert knowledge in practice codifies known and expected causal relationships for a given phenomenon. Three additional assumptions are needed:

- each variable $X_i \in \mathbf{X}$ is conditionally independent of its non-effects, both direct and indirect, given its direct causes;
- there must exist a DAG faithful to the probability distribution $\mathbf{P}$ of $\mathbf{X}$;
- there must be no latent variables (unobserved variables influencing the variables in the network) acting as confounding factors.

Obligatory XKCD



http://xkcd.com/552/

Bayesian Networks and Experimental Design

The link between BNs and survey data analysis is that, as the latter, they can be applied to

1. observational data, letting model estimation learn all the dependencies between the variables. For this to make sense we implicitly assume our sample is representative of the population;
2. experimental data, whose dependence structure is set (at least in part) by the design;

In addition, BNs make it easy to combine either type of data with interventional data (e.g. data with variables whose values are actively set by the experimenter) to disambiguate the directions of causality.

Variables that are under the control of the experimenter, because of either interventions or randomisation, cannot have incoming arcs in the BN because they are not (supposed to be) subject to external influences.

Addressing Confounding

A confounder is defined as an extraneous variable that is associated with both the variable of interest and the variables used to explain it. If such a variable is included in the BN:

- we can condition or marginalise it to remove its influence from the inference on the rest of the model;
- we can treat it an intervention and perform a counterfactual query [14], the causal equivalent of the conditional probability query above.

If such a variable is not in the BN:

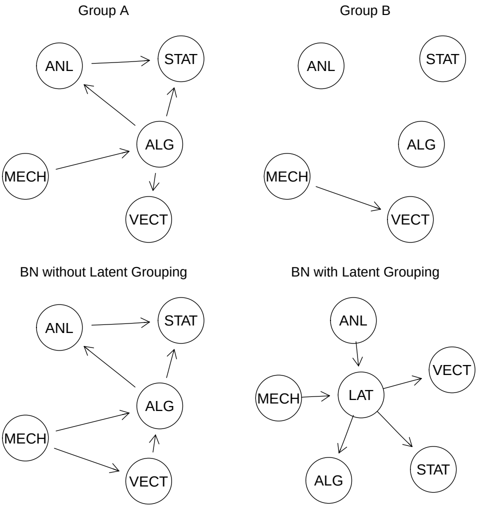
- if the structure is considered fixed, at least in the neighbourhood of the confounder, a standard application of the EM algorithm [9] can be used to impute the parameters;
- if the structure is also unknown, the structural EM [2] can be used to learn iteratively the parameter given the structure (E step) and the structure given the parameters (M step).

Confounding and Latent Variables: An Example

Edwards [1] noted that the students whose marks were recorded apparently belonged to two groups (which we will call A and B) with substantially different academic profiles. He then assigned each student to one of those two groups using the EM algorithm to impute group membership as a latent variable (LAT). The EM algorithm assigned the first 52 students (with the exception of number 45) to belong to group A, and the remainder to group B.

The BNs learned from group A and group B are completely different. And they are both different from the BN learned from the whole data set, with and without LAT.

The Marks Network, Once More



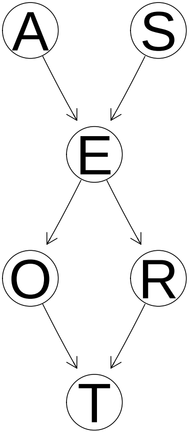
An Example: Train Use Survey

Consider a simple, hypothetical survey whose aim is to investigate the usage patterns of different means of transport, with a focus on cars and trains (disclaimer: liberally inspired by [5]).

- Age (A): *young* for individuals below 30 years old, *adult* for individuals between 30 and 60 years old, and *old* for people older than 60.
- Sex (S): *male* or *female*.
- Education (E): *up to high school* or *university degree*.
- Occupation (O): *employee* or *self-employed*.
- Residence (R): the size of the city the individual lives in, recorded as either *small* or *big*.
- Travel (T): the means of transport favoured by the individual, recorded either as *car*, *train* or *other*.

The nature of the variables recorded in the survey suggests how they may be related with each other.

The Train Use Survey as a Bayesian Network (v1)

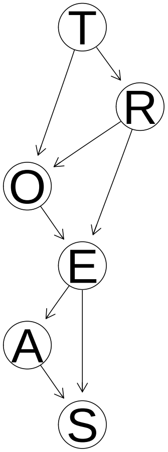


That is a prognostic view of the survey as a BN:

1. the blocks in the experimental design on top (e.g. stuff from the registry office);
2. the variables of interest in the middle (e.g. socio-economic indicators);
3. the object of the survey at the bottom (e.g. means of transport).

Variables that can be thought as "causes" are on above variables that can be considered their "effect", and confounders are on above everything else.

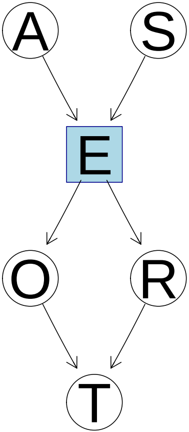
The Train Use Survey as a Bayesian Network (v2)



That is a diagnostic view of the survey as a BN: it encodes the same dependence relationships as the prognostic view but is laid out to have "effects" on top and "causes" at the bottom.

Depending on the phenomenon and the goals of the survey, one may have a graph that makes more sense than the other; but they are equivalent for any subsequent inference. For discrete BNs, one representation may have fewer parameters than the other.

Conditional Probability Queries

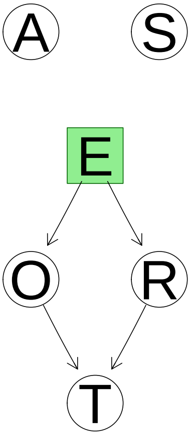


In a conditional probability query:

1. we condition on the distribution of one or more variables, but
2. the probabilistic dependencies are left intact.

This is because we are investigating the phenomenon as it was observed from the data, and therefore we let the conditioning propagate to all other variables. So the distribution of i.e. `A` is updated to `A` | `E` in the same way as `O` is updated to `O` | `E`.

Counterfactual Queries

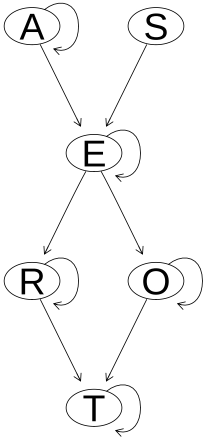


In a counterfactual query:

1. we take complete control of the distribution of one or more variables, and
2. the probabilistic dependencies of those nodes (e.g. incoming arcs) are removed from the BN.

This is because we are considering an alternate scenario than that it was observed from the data, and we let the conditioning propagate only to variables downstream (the "effects", not the "causes"). So the distribution of i.e. `A` remains unaffected but `O` is updated to `O` | `E`.

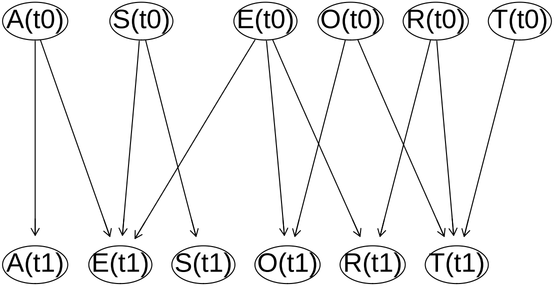
Dynamic Bayesian Networks



Dynamic BNs [11] are the temporal extension of classic BNs, which are sometimes referred to as static BNs.

- They are implicitly assumed to represent a Markov chain of order 1 — not because it is impossible to model higher-order dependencies but because we usually do not have data good/large enough to do that.
- All dependencies are assumed to flow along the arrow of time, and dependencies between variables at the same time point are generally not allowed.
- We can model feedback loops!

Unrolling and Static Bayesian Networks



All dynamic BNs can be unrolled into static BNs by duplicating nodes as required by the Markov order. Thus, there is not practical difference as far as subsequent inference is concerned.

Bayesian Networks and Panel Data

Dynamic BNs thus allow to model panel data along the same lines as normal surveys. The main differences are:

- Model estimation is much easier, because all arc directions follow the arrow of time as per the Granger causality principle [3]. No equivalence classes of BNs that are probabilistically indistinguishable.
- Model estimation is not as straightforward, because dynamic BNs have more parameters and thus require large sample sizes [4], regularisations based on strong sparsity-inducing priors [12], or other simplifying assumptions [8].
- Non-stationarity is also an issue [15], especially for discrete BNs.

Vector Auto-Regressive (VAR) processes are trivially rewritten as continuous dynamic BNs, and the same is true of discrete time Markov processes (discrete BNs), longitudinal and mixed effects models (hybrid BNs). So most models used for panel data can be expressed as BNs, which allows for standardised inference and causal inference.

Conclusions

- BNs allow an intuitive representation of dependencies for use in exploratory analysis, qualitative reasoning on the data, and to guide further modelling and inference.
- BNs provide a standardised formal treatment of causality for both static and dynamic data.
- Model estimation is largely abstracted from the nature of the data, both in the types of variables and in the sampling scheme.
- Models for both survey and panel data can be rewritten as (static or dynamic) BNS; that is, BNs subsume and generalise a number of classic models.

References I

D. I. Edwards.
Introduction to Graphical Modelling.
Springer, 2nd edition, 2000.

N. Friedman.
The Bayesian Structural EM Algorithm.
In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI1998)*, pages 129–138, 1998.

C. W. J. Granger.
Some Recent Development in a Concept of Causality.
Journal of Econometrics, 39(1–2):199–211, 1988.

D. Husmeier.
Sensitivity and Specificity of Inferring Genetic Regulatory Interactions from Microarray Experiments with Dynamic Bayesian Networks.
Bioinformatics, 19(17):2271–2282, 2003.

R .S. Kenett, G. Perruca, and S. Salini.
Modern Analysis of Customer Surveys: With Applications Using R, chapter 11.
Wiley, 2012.

References II

D. Koller and N. Friedman.
Probabilistic Graphical Models: Principles and Techniques.
MIT Press, 2009.

S. L. Lauritzen and D. J. Spiegelhalter.
Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion).
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 50(2):157–224, 1988.

S. Lèbre.
Inferring Dynamic Genetic Networks with Low Order Independencies.
Statistical Applications in Genetics and Molecular Biology, page 9, 2009.

G. J. MacLachlan and T. Krishnan.
The EM Algorithm and Extensions.
Wiley, 2nd edition, 2008.

K. V. Mardia, J. T. Kent, and J. M. Bibby.
Multivariate Analysis.
Academic Press, 1979.

References III

K. P. Murphy.
Dynamic Bayesian Networks: Representation, Inference and Learning.
PhD thesis.

R. Opgen-Rhein and K. Strimmer.
Learning Causal Networks from Systems Biology Time Course Data: an Effective Model Selection Procedure for the Vector Autoregressive Process.
BMC Bioinformatics, 8(Suppl. 2):S3, 2007.

J. Pearl.
Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
Morgan Kaufmann, 1988.

J. Pearl.
Causality: Models, Reasoning and Inference.
Cambridge University Press, 2nd edition, 2009.

J. W. Robinson and A. J. Hartemink.
Learning Non-Stationary Dynamic Bayesian Networks.
Journal of Machine Learning Research, 11:3647–3680.

References IV

J. Whittaker.
Graphical Models in Applied Multivariate Statistics.
Wiley, 1990.