

Bayesian Networks

A Statistical Perspective

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence

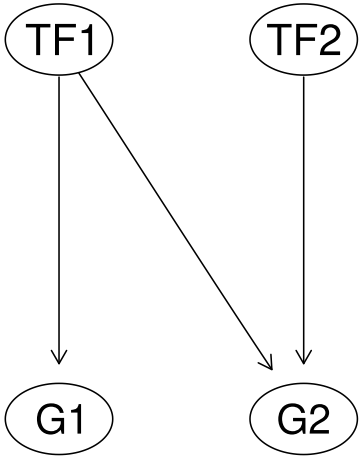
November 26, 2026

→ Graphs and Conditional Independence

Distributions and Parameter Learning

Structure Learning

Inference



The DAG implies the decomposition:

$$\mathrm{P}(\mathtt{G1}, \mathtt{G2}, \mathtt{TF1}, \mathtt{TF2}) = \mathrm{P}(\mathtt{G2} \mid \mathtt{TF1}, \mathtt{TF2})\, \mathrm{P}(\mathtt{G1} \mid \mathtt{TF1}) \times \mathrm{P}(\mathtt{TF1})\, \mathrm{P}(\mathtt{TF2})$$

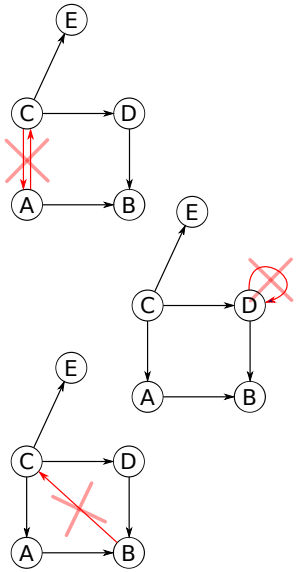
In general, $\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid \Pi_{X_i})$: each variable depends only on its parents Π_{X_i}, so the conditioning sets are smaller than those from the chain rule.

Bayesian networks [22, BNs] are defined by:

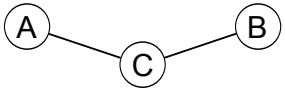
- a directed acyclic graph (DAG) $\mathcal{G}$, in which each node corresponds to a random variable X_i;
- a global probability distribution over $\mathbf{X} = \{X_1, \ldots, X_N\}$ which factorises into smaller local probability distributions according to the arcs present in the graph.

The DAG expresses the conditional independence relationships among the variables through graphical separation:

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

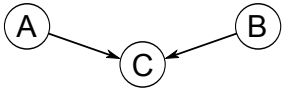


separation (undirected graphs)

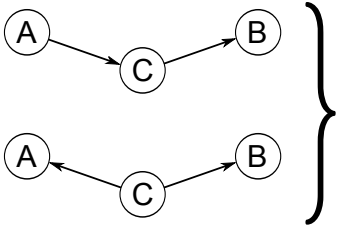


$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$
$$\mathrm{P}(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \mathrm{P}(\mathbf{A} \,|\, \mathbf{C})\, \mathrm{P}(\mathbf{B} \,|\, \mathbf{C})\, \mathrm{P}(\mathbf{C})$$

d-separation (directed acyclic graphs)



$$\mathbf{A} \not\perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$
$$\mathrm{P}(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \mathrm{P}(\mathbf{C} \,|\, \mathbf{A}, \mathbf{B})\, \mathrm{P}(\mathbf{A})\, \mathrm{P}(\mathbf{B})$$



$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$
$$\begin{aligned}\mathrm{P}(\mathbf{A}, \mathbf{B}, \mathbf{C}) &= \\ &= \mathrm{P}(\mathbf{B} \,|\, \mathbf{C})\, \mathrm{P}(\mathbf{C} \,|\, \mathbf{A})\, \mathrm{P}(\mathbf{A}) \\ &= \mathrm{P}(\mathbf{A} \,|\, \mathbf{C})\, \mathrm{P}(\mathbf{B} \,|\, \mathbf{C})\, \mathrm{P}(\mathbf{C})\end{aligned}$$

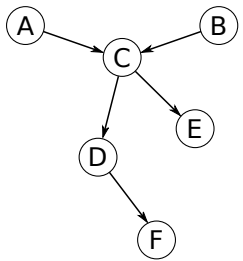
Extending the fundamental connections to every path between $\mathbf{A}$ and $\mathbf{B}$ gives d-separation.

$\mathbf{C}$ d-separates $\mathbf{A}$ from $\mathbf{B}$, denoted $\mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C}$, if along every path between a node in $\mathbf{A}$ and a node in $\mathbf{B}$ there is a node v such that either:

1. v has converging arcs (a collider), and neither v nor its descendants are in $\mathbf{C}$; or
2. v does not have converging arcs, and v is in $\mathbf{C}$.

Like independence, d-separation is symmetric. If $\mathbf{C}$ is empty, we are checking whether variables are marginally independent.

DAG



$\Longrightarrow$

Graphical separation

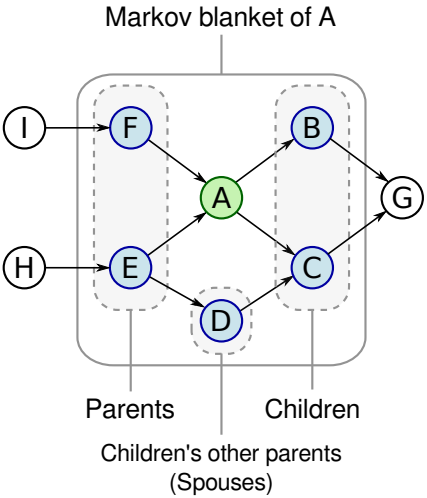
$$
\begin{aligned}
&A \perp\!\!\!\perp_G B \\
&A \perp\!\!\!\perp_G D \mid C \\
&B \perp\!\!\!\perp_G D \mid C \\
&A \perp\!\!\!\perp_G E \mid C \\
&B \perp\!\!\!\perp_G E \mid C \\
&D \perp\!\!\!\perp_G E \mid C \\
&C \perp\!\!\!\perp_G F \mid D \\
&\dots
\end{aligned}
$$

$\Longrightarrow$

Probabilistic independence

$$
\begin{aligned}
&A \perp\!\!\!\perp_P B \\
&A \perp\!\!\!\perp_P D \mid C \\
&B \perp\!\!\!\perp_P D \mid C \\
&A \perp\!\!\!\perp_P E \mid C \\
&B \perp\!\!\!\perp_P E \mid C \\
&D \perp\!\!\!\perp_P E \mid C \\
&C \perp\!\!\!\perp_P F \mid D \\
&\dots
\end{aligned}
$$

Formally, the DAG is an independence map of the probability distribution of $\mathbf{X}$, with graphical separation ($\perp\!\!\!\perp_G$) implying probabilistic independence ($\perp\!\!\!\perp_P$).



The set of nodes that graphically isolates a target node from the rest of the DAG is called its Markov blanket: its parents, its children and the other nodes sharing a child.

Since $\perp\!\!\!\perp_G$ implies $\perp\!\!\!\perp_P$, we can restrict ourselves to the Markov blanket to perform any kind of inference on the target node: it solves the variable selection problem for predicting the target.

Symmetry: if A is in the Markov blanket of B, then B is in the Markov blanket of A.

Different DAGs, Same Distribution: Equivalence Classes

$A \to C \to B$ and $A \leftarrow C \to B$ are probabilistically equivalent: we can reverse arcs as long as we do not create or remove a v-structure ($A \to C \leftarrow B$, with no arc between A and B).

DAGs fall into equivalence classes identified by the underlying undirected graph and the v-structures. The other arcs are either:

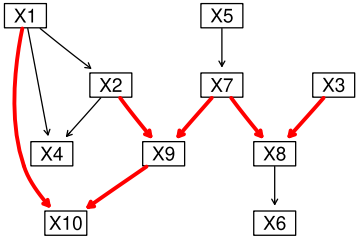
- compelled, because the other direction would introduce cycles or new v-structures;
- undetermined.

The result is a completed partially directed acyclic graph (CPDAG).

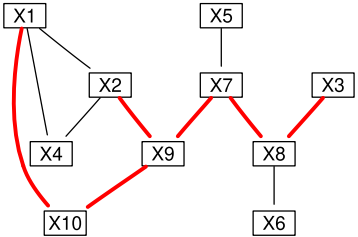
Data alone cannot orient all arcs: arcs are not causal effects yet.

DAG, CPDAG and Equivalent DAGs

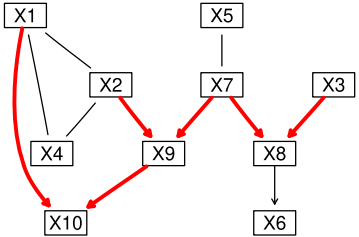
DAG



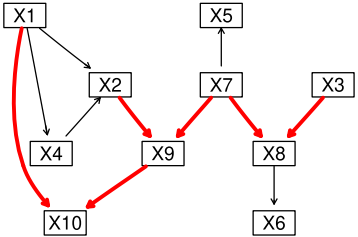
Skeleton



CPDAG



An Equivalent DAG



V-structures are in red: they are the arcs that keep their direction in the CPDAG.

Reversing $X_5 \to X_7$ or $X_2 \to X_4$ gives another DAG in the same equivalence class.

- ✓ Graphs and Conditional Independence
- → **Distributions and Parameter Learning**
- Structure Learning
- Inference

Model selection and estimation are collectively known as learning, and are usually performed as a two-step process:

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

Structure learning is done in practice as

$$\mathrm{P}(\mathcal{G} \mid \mathcal{D}) \propto \mathrm{P}(\mathcal{G})\,\mathrm{P}(\mathcal{D} \mid \mathcal{G}), \quad \mathrm{P}(\mathcal{D} \mid \mathcal{G}) = \prod_{i=1}^{N} \left[\int \mathrm{P}(X_i \mid \Pi_{X_i}, \Theta_{X_i})\,\mathrm{P}(\Theta_{X_i} \mid \Pi_{X_i}) d\Theta_{X_i} \right]$$

and parameter learning as

$$\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D}) = \prod_{i=1}^{N} \mathrm{P}(\Theta_{X_i} \mid \Pi_{X_i}, \mathcal{D}).$$

Thanks to the decomposition, we only manipulate one local distribution at a time.

What About the Probability Distributions?

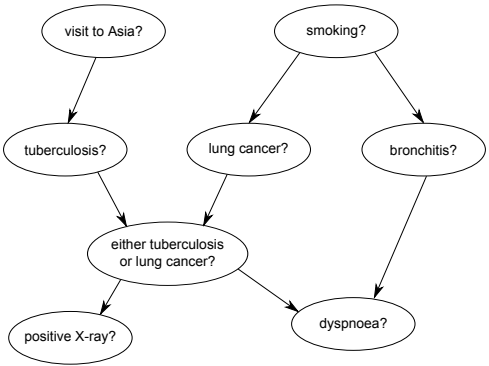
The second component of a BN is the probability distribution $\mathrm{P}(\mathbf{X})$. The choice should be such that the BN:

- can be learned efficiently from data;
- is flexible (distributional assumptions should not be too limiting);
- is easy to query to perform inference.

The three most common choices are:

- discrete BNs: $\mathbf{X}$ and the $X_i \mid \Pi_{X_i}$ are multinomial;
- Gaussian BNs (GBNs): $\mathbf{X}$ is multivariate normal, the $X_i \mid \Pi_{X_i}$ are univariate normal;
- conditional linear Gaussian BNs (CLGBNs): $\mathbf{X}$ is a mixture of multivariate normals.

Exact inference is possible in these three cases, hence their popularity.



ASIA [15]: total parameters of $\mathbf{X}$: $2^8 - 1 = 255$.

The global distribution is $\mathbf{X} \sim Mul(\pi)$, with one probability for each combination of values of $\mathbf{X}$. They are too many ($l^{|\mathbf{X}|} - 1$ if each variable takes l values), too small and have no structure.

We then assume:

1. positivity: all probabilities are strictly positive;
2. parameter independence across parent configurations;
3. parameter modularity across local distributions.

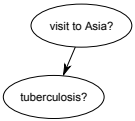
Conditional Probability Tables (CPTs)



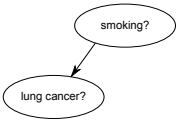
yes	no
0.01	0.99



yes	no
0.50	0.50



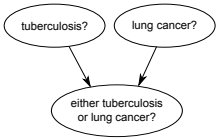
	yes	no
yes	0.05	0.95
no	0.01	0.99



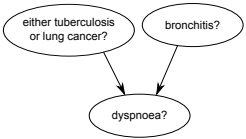
	yes	no
yes	0.10	0.90
no	0.01	0.99



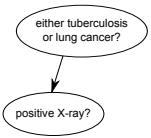
	yes	no
yes	0.60	0.40
no	0.30	0.70



	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

Then $X_i \mid \Pi_{X_i} \sim \mathit{Mul}(\pi_{ik|j})$ with $\pi_{ik|j} = \mathrm{P}(X_i = k \mid \Pi_{X_i} = j)$: the local distributions take the form of conditional probability tables (CPTs), with $(l-1)l^{|\Pi_{X_i}|}$ parameters each.

Overall parameters of the $X_i \mid \Pi_{X_i}$: 18 instead of 255.

The $\pi_{ik|j}$ still do not separate the effects of the Π_{X_i}.

A CPT is equivalent to a saturated multinomial logistic regression for X_i, with the Π_{X_i} and their interactions of all orders as regressors.

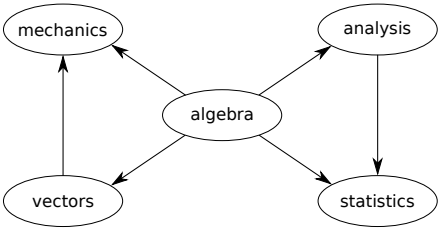
For each value k of X_i, via the multinomial link function:

$$\log(\pi_{ik|\bullet}/\pi_{i1|\bullet}) = \eta_{ik|\bullet} \qquad \eta_{ik|\bullet} = \mathbf{Z}_i\boldsymbol{\gamma}_{ik}$$

where $\mathbf{Z}_i$ is the saturated design matrix and the $\boldsymbol{\gamma}_{ik}$ are the regression coefficients.

Now we can:

- separate main and interaction effects; and
- simplify the structure of Θ_{X_i} by removing higher-order interactions as needed.



MARKS [16]: total parameters of $\mathbf{X}$: $5 + 15 = 20$.

The global distribution is $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ is the vector of the means ($|\mathbf{X}| \times 1$) and Σ is the covariance matrix ($|\mathbf{X}| \times |\mathbf{X}|$).

Classical multivariate statistics tells us that

1. if we assume that Σ is full rank, we can compute the precision matrix $\Omega = \Sigma^{-1}$;
2. if $\Omega_{ij} = 0$ then $X_i \perp\!\!\!\perp_P X_j \mid \mathbf{X} \setminus \{X_i, X_j\}$.

In a GBN, $\Omega_{ij} = 0$ if X_i and X_j are neither adjacent nor share a child in the DAG. As a result, GBNs can only model linear dependencies.

The local distributions $\mathrm{P}(X_i \mid \Pi_{X_i})$ are linear regressions:

$$X_i = \mu_{X_i} + \Pi_{X_i}\boldsymbol{\beta}_{X_i} + \varepsilon_{X_i}, \qquad \varepsilon_{X_i} \sim N(0, \sigma^2_{X_i}).$$

Equivalently, they can be parametrised with partial correlations. If $\Omega^{(i)}$ is the inverse of the covariance matrix of $\{X_i, \Pi_{X_i}\}$, with X_i first:

$$\beta_{X_i,j} = -\frac{\Omega^{(i)}_{1,j}}{\Omega^{(i)}_{1,1}}, \qquad \sigma^2_{X_i} = \frac{1}{\Omega^{(i)}_{1,1}}, \qquad \rho_{ij} = -\frac{\Omega^{(i)}_{1,j}}{\sqrt{\Omega^{(i)}_{1,1}\Omega^{(i)}_{j,j}}},$$

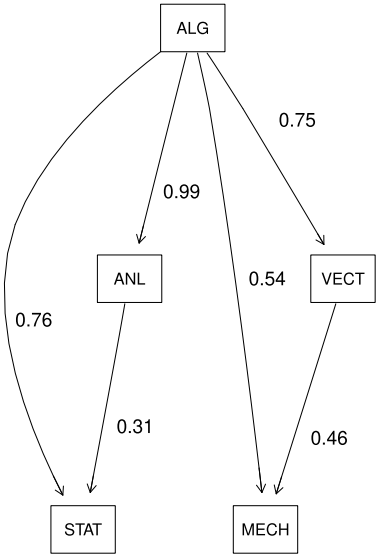
where ρ_{ij} is the partial correlation between X_i and its jth parent given the other parents.

For the MARKS data:

$$\begin{aligned}
\mathsf{ALG} &= 50.60 + \varepsilon_{\mathsf{ALG}} \sim N(0, 112.8) \\
\mathsf{ANL} &= -3.57 + 0.99\mathsf{ALG} + \varepsilon_{\mathsf{ANL}} \sim N(0, 110.25) \\
\mathsf{MECH} &= -12.36 + 0.54\mathsf{ALG} + 0.46\mathsf{VECT} + \varepsilon_{\mathsf{MECH}} \sim N(0, 195.2) \\
\mathsf{STAT} &= -11.19 + 0.76\mathsf{ALG} + 0.31\mathsf{ANL} + \varepsilon_{\mathsf{STAT}} \sim N(0, 158.8) \\
\mathsf{VECT} &= 12.41 + 0.75\mathsf{ALG} + \varepsilon_{\mathsf{VECT}} \sim N(0, 109.8)
\end{aligned}$$

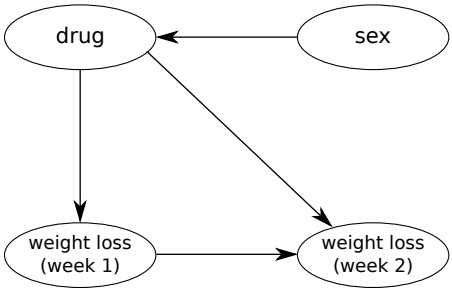
Parents only contribute main effects, no interactions, in their natural form, no transforms. Overall parameters of the $X_i \mid \Pi_{X_i}$: $11 + 5 = 16$.

The regression coefficients are the path coefficients of Wright's path analysis [26].



CLGBNs combine discrete BNs and GBNs to obtain a mixture-of-Gaussians distribution:

- continuous nodes cannot be parents of discrete nodes;
- discrete nodes have CPTs;
- continuous nodes have one linear regression for each configuration of the discrete parents, with the continuous parents as regressors.



RATS' WEIGHTS: weight loss in a drug trial [4].

The discrete nodes define the components of a mixture, each component is the joint distribution of the continuous nodes. Here $3 \times 2 = 6$ components, each a bivariate normal.

The local distribution for the second weight loss is:

$$\begin{aligned}
\texttt{WL2}_{\texttt{D1}} &= 1.02 + 0.89\texttt{WL1} + \varepsilon_{\texttt{WL2|D1}} \sim N(0, 3.2) \\
\texttt{WL2}_{\texttt{D2}} &= -1.68 + 1.35\texttt{WL1} + \varepsilon_{\texttt{WL2|D2}} \sim N(0, 4) \\
\texttt{WL2}_{\texttt{D3}} &= -1.83 + 0.82\texttt{WL1} + \varepsilon_{\texttt{WL2|D3}} \sim N(0, 1.9)
\end{aligned}$$

We can write each coefficient as the deviation from the respective means across the drugs:

$$\begin{aligned}
\texttt{WL2}_{\texttt{D1}} &= (-0.83 + 1.85) + (1.02 - 0.13)\texttt{WL1} + \varepsilon_{\texttt{WL2|D1}} \\
\texttt{WL2}_{\texttt{D2}} &= (-0.83 - 0.85) + (1.02 + 0.33)\texttt{WL1} + \varepsilon_{\texttt{WL2|D2}} \\
\texttt{WL2}_{\texttt{D3}} &= (-0.83 - 1.00) + (1.02 - 0.20)\texttt{WL1} + \varepsilon_{\texttt{WL2|D3}}
\end{aligned}$$

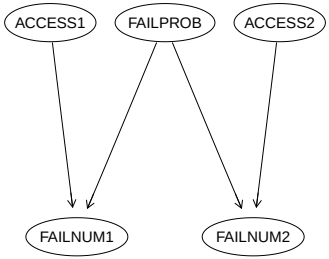
This is the form of a mixed-effects model: fixed intercept (-0.83) and slope (1.02), random intercept and slope for each drug.

In additive BNs [19, 12], each local distribution is a generalised linear model in which the parents enter additively on the scale of the link function $g(\cdot)$:

$$g\left(\mathrm{E}(X_i \mid \Pi_{X_i})\right) = \mu_{X_i} + \Pi_{X_i}\boldsymbol{\beta}_{X_i}.$$

- Binomial (logit), Gaussian (identity), Poisson (log) and multinomial nodes in the same BN, with no constraints on which arcs may be present.
- No interactions between the parents: far fewer parameters than a CPT.
- Scores are marginal likelihoods or BIC; grouped data are handled with random effects.
- A parametric bootstrap from the learned BN identifies and removes the arcs that are due to overfitting.

We are free to choose any $\mathrm{P}(X_i \mid \Pi_{X_i}; \Theta_{X_i})$, framing any hierarchical model as a BN. Nodes should be the variables in the model, not their parameters, to interpret it like a classical BN.



Reliability testing, from [5].

$$\texttt{FAILPROB} \sim \textit{Beta}(2, 5)$$
$$\texttt{ACCESS1} \sim \textit{Pois}(20)$$
$$\texttt{ACCESS2} \sim \textit{Pois}(400)$$
$$\texttt{FAILNUM1} \sim \textit{Bi}(\texttt{ACCESS1}, \texttt{FAILPROB})$$
$$\texttt{FAILNUM2} \sim \textit{Bi}(\texttt{ACCESS2}, \texttt{FAILPROB})$$

We gain flexibility and software such as Stan.

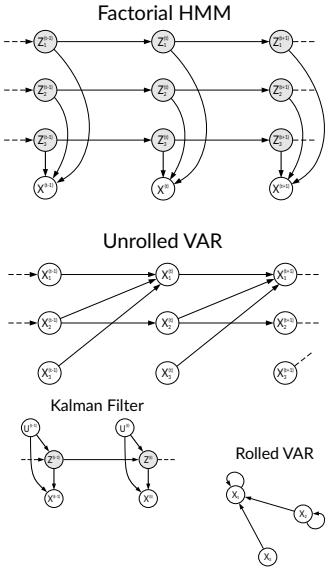
We lose closed-form results for densities, parameter estimates, scores and tests.

Any stochastic process can be treated as a dynamic BN:

- all variables $\mathbf{X}$ become separate nodes in each time point;
- arcs only go forward in time, so the graph is still a DAG and can model feedback loops.

We typically assume order $p = 1$, no instantaneous arcs, and a homogeneous, stationary time series.

A VAR process is a Gaussian dynamic BN, with $X_j^{(t-l)} \in \Pi_{X_i^{(t)}}$ if and only if $A_l[i, j] \neq 0$.



Given the DAG, we estimate the parameters of each local distribution, one at a time.

- Maximum likelihood estimators: the usual empirical estimators. They are unstable in most multivariate problems and often non-smooth.
- Bayesian posterior estimators: with conjugate priors, to stay in closed form [11]. The Dirichlet for discrete BNs,

$$Dir(\alpha_{ik|j}) \xrightarrow{data} Dir(\alpha_{ik|j} + n_{ik|j}), \qquad \hat{\pi}_{ik|j} = \frac{\alpha_{ik|j} + n_{ik|j}}{\sum_k (\alpha_{ik|j} + n_{ik|j})},$$

  and the Inverse Wishart for GBNs. The imaginary sample size $\sum_k \alpha_{ik|j}$ is the weight of the prior, in observations.
- Regularised estimators: the local distributions are regressions, so we can estimate them as penalised or random-effects regressions.

Penalised regressions for GBNs:

$$\widehat{\boldsymbol{\beta}}_{X_i} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| X_i - \mu_{X_i} - \Pi_{X_i}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_q^q ,$$

with $q = 2$ (ridge) or $q = 1$ (lasso, which also drops parents). The same penalties on the logistic form of a CPT, or in additive BNs. Both are posterior modes under Gaussian or Laplace priors [8].

Random-effects regressions for CLGBNs: the coefficients for each configuration j of the discrete parents are $\boldsymbol{\beta}_{X_i|j} \sim N(\boldsymbol{\beta}_{X_i}, \Psi)$, so their estimates are pulled towards the fixed effects, more strongly when n_j is small. For a random intercept with variance τ^2 [18, 24]:

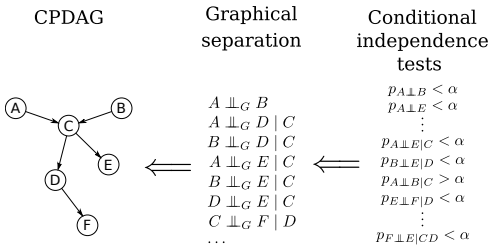
$$\hat{b}_j = \frac{\tau^2}{\tau^2 + \sigma^2/n_j} \left(\bar{x}_{i|j} - \hat{\mu}_{X_i} \right).$$

- ✓ Graphs and Conditional Independence
- ✓ Distributions and Parameter Learning
- → Structure Learning
- Inference

Assumptions for Structure Learning from Data

- There is a one-to-one correspondence between the nodes in the DAG and the random variables in $\mathbf{X}$, with no variable a deterministic function of others.
- All the relationships between the variables are conditional independencies, the only kind a BN can express.
- The global distribution is strictly positive: every combination of values is possible, even if really unlikely. This makes the model uniquely identifiable.
- Observations are independent. Temporal or spatial dependence must be built into the network, as in dynamic BNs.

Constraint-based algorithms use conditional independence tests to learn "constraints" from the data, and assume that the DAG is a perfect map.



Score-based algorithms find the DAG that maximises a goodness-of-fit score, typically decomposable as

$$\mathrm{Score}(\mathcal{G}) = \sum_i \mathrm{Score}(X_i \mid \Pi_{X_i}),$$

with heuristics such as hill-climbing.

Differentiable algorithms maximise the score by continuous optimisation over a weighted adjacency matrix W, subject to a smooth acyclicity constraint such as DAGMA's [1]:

$$h(W) = -\log\det(sI - W \circ W) + N \log s = 0.$$

BNs are defined from graphical separation:

$$\mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C} \implies \mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C}.$$

Constraint-based algorithms such as the PC algorithm [3] also assume the reverse:

$$\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \iff \mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C}.$$

This is a much stronger assumption (faithfulness), which has pros and cons:

- It is impossible to verify.
- It is a sufficient assumption to uniquely identify Markov blankets, and thus we no longer need to assume that $\mathrm{P}(\mathbf{X})$ is strictly positive everywhere.
- Not all $\mathrm{P}(\mathbf{X})$ have a faithful DAG.

For discrete BNs, the tests for $X \perp\!\!\!\perp_P Y \mid \mathbf{Z}$ are Pearson's X^2 and the log-likelihood ratio G^2 (the mutual information) on the observed frequencies n_{ijk}:

$$\mathrm{G}^2(X, Y \mid \mathbf{Z}) = 2 \sum_{i=1}^{R} \sum_{j=1}^{C} \sum_{k=1}^{L} n_{ijk} \log \frac{n_{ijk} n_{++k}}{n_{i+k} n_{+jk}} \sim \chi^2_{(R-1)(C-1)L}.$$

For GBNs, the exact t-test and Fisher's Z for the partial correlations $\rho_{XY|\mathbf{Z}}$:

$$\mathrm{Z}(X, Y \mid \mathbf{Z}) = \log \left(\frac{1 + \rho_{XY|\mathbf{Z}}}{1 - \rho_{XY|\mathbf{Z}}} \right) \frac{\sqrt{n - |\mathbf{Z}| - 3}}{2} \sim N(0, 1).$$

For CLGBNs, log-likelihood ratios between nested regressions.

The threshold α is a tuning parameter, not a type I error rate: the tests are many, sequential and dependent.

An exhaustive search is unfeasible in practice, regardless of the network score used in the process. However, we can use heuristics in combination with decomposable scores

$$\text{Score}(\mathcal{G}) = \sum_{i=1}^{N} \text{Score}(X_i \mid \Pi_{X_i})$$

such as

$$\text{BIC}(\mathcal{G}) = \sum_{i=1}^{N} \log \text{P}(X_i \mid \Pi_{X_i}; \widehat{\Theta}_{X_i}) - \frac{|\Theta_{X_i}|}{2} \log n,$$

$$\text{BDe}(\mathcal{G}), \text{BGe}(\mathcal{G}) = \sum_{i=1}^{N} \log \left[\int \text{P}(X_i \mid \Pi_{X_i}, \Theta_{X_i}) \, \text{P}(\Theta_{X_i} \mid \Pi_{X_i}) d\Theta_{X_i} \right],$$

because we only compare BNs that differ in one local distribution at a time.

AIC overfits a lot. BIC may under-fit a bit but it is a good default to use.

For discrete BNs, with complete data, a Dirichlet conjugate prior $\Theta_{X_i} \mid \Pi_{X_i} \sim \mathit{Dir}(\alpha_{ijk})$, positivity, parameter independence and modularity [9]:

$$\mathrm{BD}(\mathcal{G}, \mathcal{D}; \boldsymbol{\alpha}) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \left[\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \right]$$

where r_i is the number of states of X_i; q_i the number of configurations of Π_{X_i}; $n_{ij} = \sum_k n_{ijk}$; $\alpha_{ij} = \sum_k \alpha_{ijk}$. The usual choice is BDeu, $\alpha_{ijk} = \alpha/(r_i q_i)$, with α the imaginary sample size.

For GBNs, with a normal-Wishart conjugate prior [14]:

$$\mathrm{BGe}(\mathcal{G}, \mathcal{D}) = \prod_{i=1}^{N} \frac{\mathrm{P}(\mathcal{D}_{\{X_i, \Pi_{X_i}\}})}{\mathrm{P}(\mathcal{D}_{\Pi_{X_i}})}, \qquad \mathrm{P}(\mathcal{D}_{\mathbf{Y}}) \propto \frac{|T_{\mathbf{YY}}|^{a/2}}{|R_{\mathbf{YY}}|^{(a+n)/2}},$$

where T is the prior scale matrix with a degrees of freedom, and R is T updated with the sample covariance matrix.

The most common choice for $\mathrm{P}(\mathcal{G})$ is the uniform distribution because it feels like it is uninformative. However [20]:

- It favours dense graphs: each pair of nodes is connected with probability $2/3$.
- Arcs incident on a common node are correlated in the prior, so false positives and false negatives can propagate.

The marginal uniform prior [21] includes an arc with probability $\leqslant 1/4$, its reverse with the same probability, and no arc with probability $\geqslant 1/2$, independently for each pair of nodes.

Setting the probability of inclusion to $c\frac{2}{N-1}, c \in [1, 3]$ gives $O(cN)$ expected arcs, which often works even better.

Experts find it difficult to suggest values for the parameters, so their knowledge is best used on the structure, with an informative prior:

- limiting the number of parents of each node;
- setting different probabilities of inclusion for different arcs [2];
- forcing (whitelisting) or forbidding (blacklisting) arcs;
- setting the topological ordering of the nodes.

These priors can rule out most DAGs, shrinking the search space by orders of magnitude.

Energy-based priors [10, 17] are softer:

$$\mathrm{P}(\mathcal{G}) \propto \exp\left(-\beta E(\mathcal{G})\right),$$

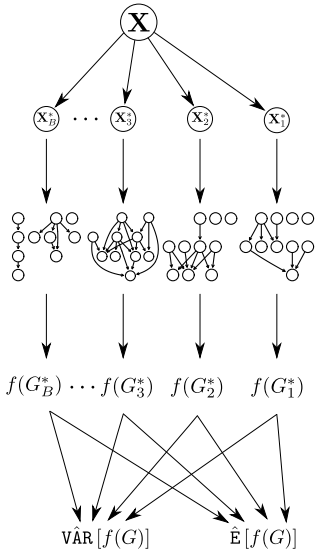
where the energy $E(\mathcal{G})$ measures how much $\mathcal{G}$ disagrees with prior knowledge and $\beta \geqslant 0$ sets its weight ($\beta = 0$ gives the uniform prior).

Model Averaging: We Should Always Do That!

Structure learning from limited amounts of data is inherently noisy. Perturbing the data, learning multiple DAGs and then averaging them is most effective in removing that noise [6]:

- searching from different starting points;
- learning from bootstrap samples (bagging).

The frequency of an arc quantifies our confidence that it is real, not the size or the sign of its effect. A consensus DAG keeps the arcs above a threshold estimated from the data [25].



Instead, we can work with the whole posterior distribution over the DAGs, for any feature f of the DAG such as the presence of an arc:

$$\mathrm{P}(f \mid \mathcal{D}) = \sum_{\mathcal{G}} f(\mathcal{G}) \, \mathrm{P}(\mathcal{G} \mid \mathcal{D}), \qquad \mathrm{P}(\mathcal{G} \mid \mathcal{D}) \propto \mathrm{P}(\mathcal{G}) \, \mathrm{P}(\mathcal{D} \mid \mathcal{G}).$$

The number of DAGs grows super-exponentially with N, so we sample from $\mathrm{P}(\mathcal{G} \mid \mathcal{D})$ with MCMC, using a BD or BGe marginal likelihood and a graphical prior:

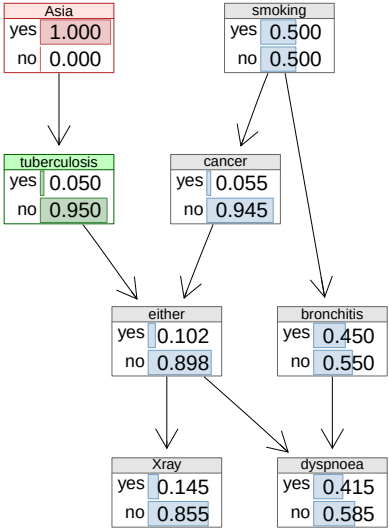
- structure MCMC adds, removes and reverses single arcs, and mixes slowly;
- order MCMC samples topological orderings and mixes much better, but it biases the posterior over DAGs [7];
- partition MCMC samples ordered partitions of the nodes, and removes that bias [13].

- ✓ Graphs and Conditional Independence
- ✓ Distributions and Parameter Learning
- ✓ Structure Learning
- → Inference

Questions that can be asked are called queries and are typically about an event of interest (in green) given some evidence we have observed (in red).

The two most common queries are:

- conditional probability queries ("What is the probability of tuberculosis for someone who visited Asia?"); and
- most probable explanation queries ("What is the most likely scenario for tuberculosis infection?")



Exact and Approximate Inference

Inference automates quantitative assessments of probabilistic statements in BNs: you ask queries on some event given some evidence you have observed, and you get the answer.

Exact Inference	Approximate Inference
Variable elimination, junction trees.	Logic sampling and likelihood weighting.
Heavier computational cost.	More scalable.
Akin to symbolic computations. They provide exact answers.	Like posterior inference, answers have simulation noise.

Exact inference is possible for discrete BNs and GBNs (conditioning a multivariate normal, in closed form); for CLGBNs it is much more expensive; for general BNs it is impossible.

INPUT: a BN, evidence E and query event Q.

1. Order the variables in $\mathbf{X}$ according to the topological ordering in the DAG, so that parents come before children.

2. Set $n_E = 0$ and $n_{E,Q} = 0$.

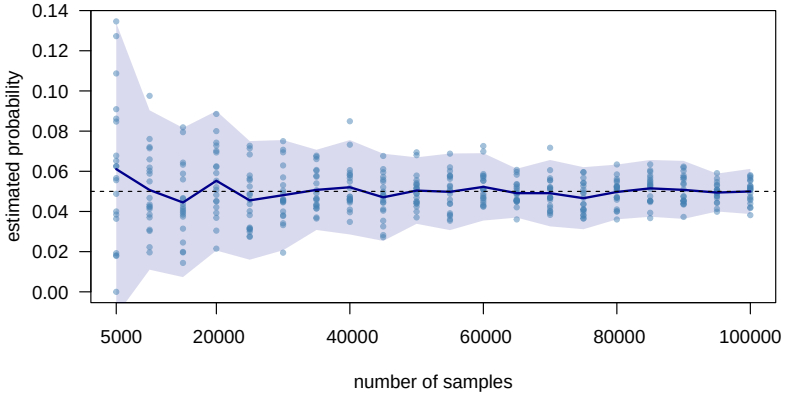
3. For a suitably large number of samples $\mathbf{x}$:

   3.1 generate a random value from each $X_i \mid \Pi_{X_i}$, in order, using the Π_{X_i} we already generated;

   3.2 if $\mathbf{x}$ includes E, set $n_E = n_E + 1$;

   3.3 if $\mathbf{x}$ includes both Q and E, set $n_{E,Q} = n_{E,Q} + 1$.

4. The answer to the query is the estimated probability $n_{E,Q}/n_E$.



$\mathrm{P}(\text{tuberculosis} \mid \text{Asia}) = 0.05$, estimated 20 times for each number of samples. Sampling variability decreases, but it never goes to zero.

Summary

- BNs use DAGs to encode conditional independence through d-separation and to factorise the global distribution into local distributions.
- Discrete BNs, GBNs and CLGBNs correspond to logistic regressions, linear regressions and mixed-effects models; general BNs are hierarchical Bayesian models.
- Structure learning algorithms are constraint-based, score-based or differentiable. There is no "best" algorithm [23]: priors and model averaging matter as much.
- Inference is either exact or approximate.
- DAGs in the same equivalence class cannot be told apart from data: Part 2 covers when arcs are causal, Part 3 dynamic BNs and model averaging in practice.

That's all!

Happy to discuss in more detail.

scutari@bnlearn.com

References I

[1] K. Bello, B. Aragam, and P. Ravikumar.
DAGMA: Learning Dags via M-Matrices and a Log-Determinant Acyclicity Characterization.
Advances in Neural Information Processing Systems, 35:8226–8239, 2022.

[2] 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.

[3] D. Colombo and M. H. Maathuis.
Order-Independent Constraint-Based Causal Structure Learning.
Journal of Machine Learning Research, 15:3921–3962, 2014.

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

[5] N. Fenton and M. Neil.
Risk Assessment and Decision Analysis with Bayesian Networks.
CRC Press, 2nd edition, 2018.

References II

[6] N. Friedman, M. Goldszmidt, and A. Wyner.
Data Analysis with Bayesian Networks: A Bootstrap Approach.
In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 196–205, 1999.

[7] N. Friedman and D. Koller.
Being Bayesian about Bayesian Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks.
Machine Learning, 50(1–2):95–126, 2003.

[8] T. Hastie, R. Tibshirani, and M. Wainwright.
Statistical Learning with Sparsity: The Lasso and Generalizations.
CRC Press, 2015.

[9] D. Heckerman, D. Geiger, and D. M. Chickering.
Learning Bayesian Networks: The Combination of Knowledge and Statistical Data.
Machine Learning, 20(3):197–243, 1995.

[10] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano.
Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks.
Journal of Bioinformatics and Computational Biology, 2(1):77–98, 2004.

References III

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

[12] G. Kratzer, F. Lewis, A. Comin, M. Pittavino, and R. Furrer. Additive Bayesian Network Modeling with the R Package Abn. *Journal of Statistical Software*, 105(8), 2023.

[13] J. Kuipers and G. Moffa. Partition MCMC for Inference on Acyclic Digraphs. *Journal of the American Statistical Association*, 112(517):282–299, 2017.

[14] J. Kuipers, G. Moffa, and D. Heckerman. Addendum on the Scoring of Gaussian Directed Acyclic Graphical Models. *The Annals of Statistics*, 42(4):1689–1691, 2014.

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

References IV

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

[17] S. Mukherjee and T. P. Speed.
Network Inference Using Informative Priors.
Proceedings of the National Academy of Sciences, 105(38):14313–14318, 2008.

[18] J. C. Pinheiro and D. M. Bates.
Mixed-Effects Models in S and S-PLUS.
Springer, 2000.

[19] M. Pittavino, A. Dreyfus, C. Heuer, J. Benschop, P. Wilson, J. Collins-Emerson, P. R. Torgerson, and R. Furrer.
Comparison between Generalized Linear Modelling and Additive Bayesian Network; Identification of Factors Associated with the Incidence of Antibodies against Leptospira interrogans sv Pomona in Meat Workers in New Zealand.
Acta Tropica, 173:191–199, 2017.

[20] M. Scutari.
On the Prior and Posterior Distributions Used in Graphical Modelling.
Bayesian Analysis, 8(3):505–532, 2013.

References V

M. Scutari.
[21] An Empirical-Bayes Score for Discrete Bayesian Networks.
In *Proceedings of the 8th International Conference on Probabilistic Graphical Models*, volume 52 of *Proceedings of Machine Learning Research*, pages 438–448, 2016.

M. Scutari and J.-B. Denis.
[22] *Bayesian Networks with Examples in R*.
Chapman & Hall, 2nd edition, 2021.

M. Scutari, C. E. Graafland, and J. M. Gutiérrez.
[23] 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. Marquis, and L. Azzimonti.
[24] Using Mixed-Effect Models to Learn Bayesian Networks from Related Data Sets.
Proceedings of Machine Learning Research (PGM 2022), 186:73–84, 2022.

M. Scutari and R. Nagarajan.
[25] On Identifying Significant Edges in Graphical Models of Molecular Networks.
Artificial Intelligence in Medicine, 57(3):207–217, 2013.

References VI

S. Wright.
[26] The Method of Path Coefficients.
The Annals of Mathematical Statistics, 5(3):161–215, 1934.



Causality

An Introduction

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence

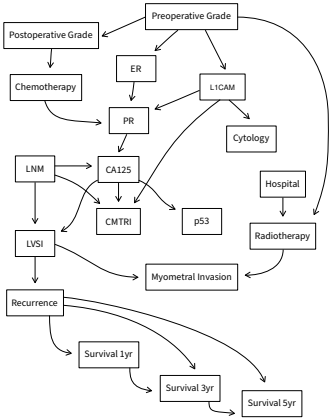
November 26, 2026

Bayesian networks [17, BNs] are defined by:

- a directed acyclic graph (DAG) $\mathcal{G}$, in which each node corresponds to a random variable X_i;
- a global probability distribution over $\mathbf{X} = \{X_1, \ldots, X_N\}$ which factorises into smaller local probability distributions according to the arcs present in the graph.

The DAG expresses the conditional independence relationships among the variables through graphical separation:

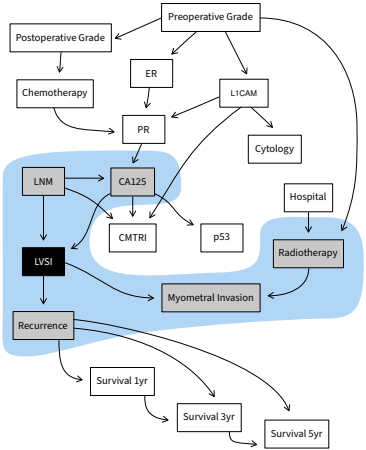
$$\mathrm{P}(\mathbf{X}; \Theta) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid \Pi_{X_i}; \Theta_{X_i}), \quad \Pi_{X_i} = \{\text{parents of } X_i\}.$$



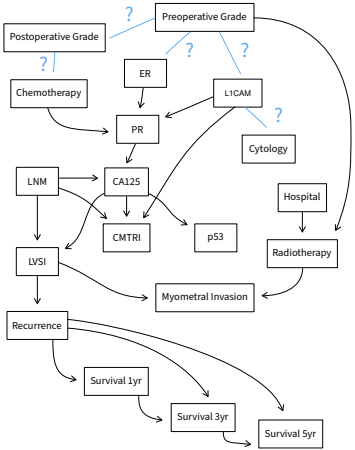
Endometrial cancer, AIME 2023 [21].

Key Properties of Bayesian Networks

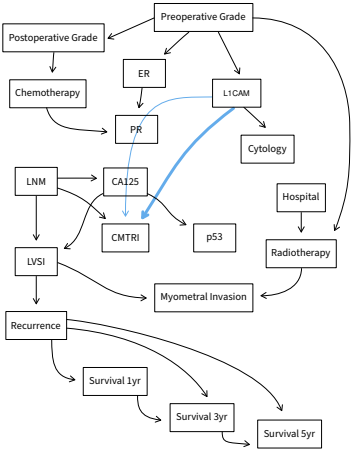
Markov Blankets: Feature Selection



Equivalence Classes: Identifiability



Path Analysis: Mediation



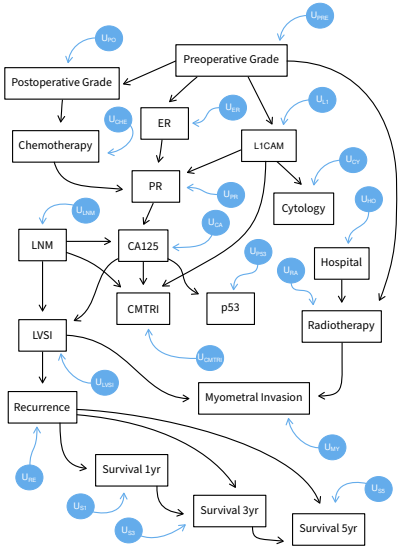
A structural causal model [1, SCM] is defined by:

- some endogenous (deterministic) variables $\mathbf{X}$;
- some exogenous (stochastic) variables $\mathbf{U}$;
- a DAG $\mathcal{G}$ with nodes mapping to $\{X_i, U_i\} \subseteq \mathbf{X} \cup \mathbf{U}$;
- one structural equation $f_i(\cdot)$ for each X_i.

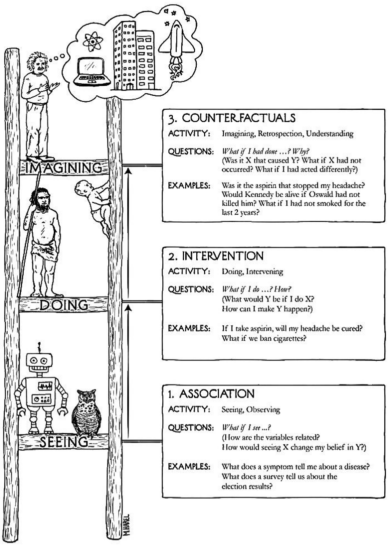
Then the X_i are defined by the structural equations:

$$X_i := f_i(\Pi_{X_i}, U_i), \qquad \Pi_{X_i} \subseteq \mathbf{X}.$$

SCMs are structural equation models (SEMs).

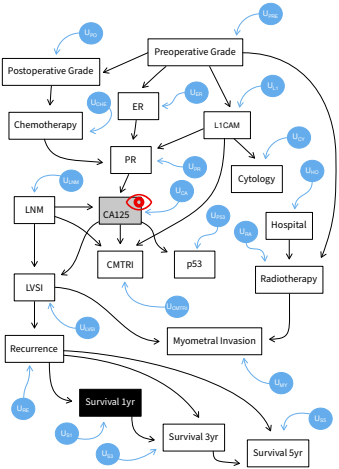


Bayesian Networks	Structural Causal Models
Probabilistic independence relationships are symmetric: if $X_i \perp\!\!\!\perp X_j$, then $X_j \perp\!\!\!\perp X_i$.	Causes and effects are not symmetric: treating symptoms does not cure the underlying disease.
X_i as CPTs, linear regressions or generalised linear models [9, 17], as in Part 1.	X_i as additive noise models [15, ANMs]: $f_i(\Pi_{X_i}, U_i) \approx f_i(\Pi_{X_i}) + U_i$, main effects only, fitted with regularised regressions.
Assume no latent variables.	Assume causal sufficiency (no unmeasured confounding) by default, but can relax it.



Network models can be used at various rungs of the Ladder of Causation [14].

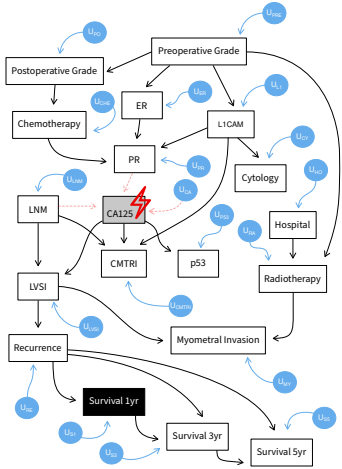
- BNs support level 1 (*Association*). If their arcs are causal effects (causal BNs), they also support level 2 (*Intervention*).
- SCMs also support level 3 (*Counterfactuals*).
- SCMs can express latent confounding by linking exogenous variables with multiple endogenous variables. BNs cannot: they need other graphs, such as the ancestral graphs learned by RFCI [8].



Just observing

$$\mathrm{P}(\text{Survival 1y} \mid \mathrm{CA125} = c) \neq \mathrm{P}(\text{Survival 1y} \mid \mathrm{do}(\mathrm{CA125} = c))$$

Actively acting

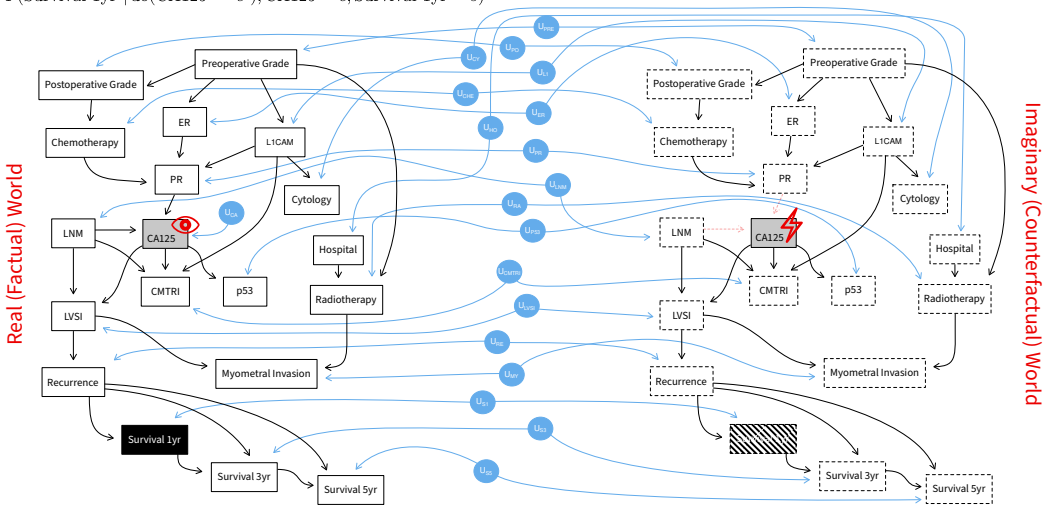


Manipulate variables, or use back-door adjustment for a $\mathbf{Z}$ that blocks all non-causal paths:

$$\mathrm{P}(Y \mid \mathrm{do}(X = x)) = \sum_{\mathbf{z}} \mathrm{P}(Y \mid X = x, \mathbf{Z} = \mathbf{z})\, \mathrm{P}(\mathbf{Z} = \mathbf{z}).$$

Counterfactuals

$$\mathrm{P}(\text{Survival 1yr}' \mid \mathrm{do}(\text{CA125}' = c'), \text{CA125} = c, \text{Survival 1yr} = s)$$



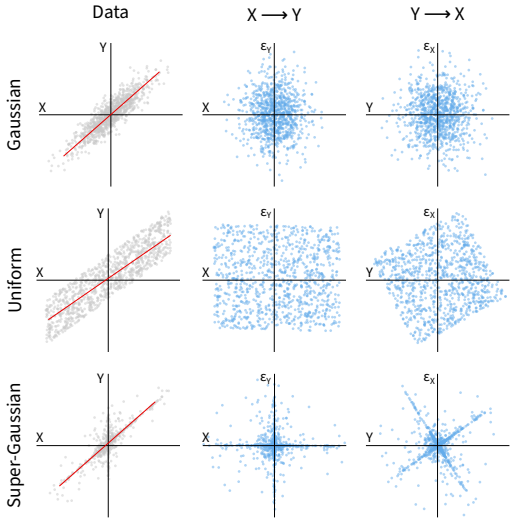
Imagine a world that did not actually occur: potential outcomes.

Class	Algorithms	Approach
Constraint-Based	PC [7] RFCI [8]	Conditional independence tests to prune the search space into an equivalence class.
Score-Based	GES [6] tabu search	Navigating the space of DAGs or equivalence classes to find the SCM with the best score.
Differentiable	NOTEARS [22] DAGMA [2]	Minimising residual variance via gradient descent, with constraints for SCM acyclicity and sparsity.
Functional	LiNGAM [19] ANMs [15]	Leveraging the asymmetries in the residuals to orient the arcs.

Data alone identify only the equivalence class (the CPDAG). Causal directions are identified with additional data or assumptions:

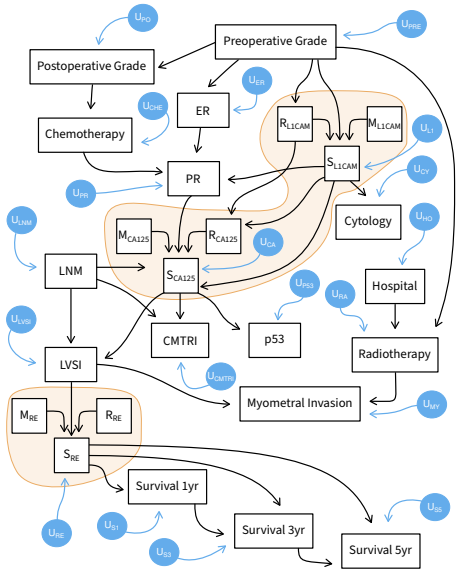
- Data from multiple environments [10].
- Data with different interventions [11].
- Non-Gaussianity [19].
- Heteroscedasticity [20].
- Spatial correlations, non-IID data [18].

Randomisation also works as usual.



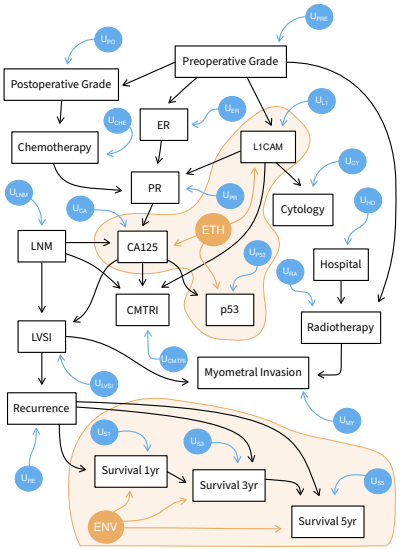
Modelling Missingness

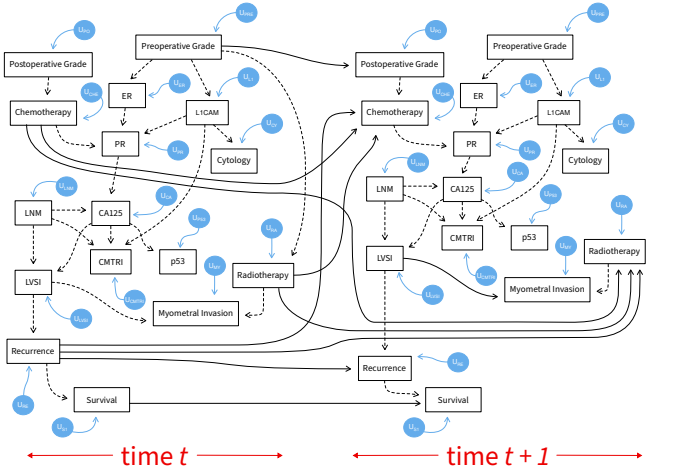
- All missing data patterns [12] have a graphical representation in missingness graphs [13].
- Each partially-observed X_i is split into three nodes, $R_{X_i} \rightarrow S_{X_i} \leftarrow M_{X_i}$: M_{X_i} (complete, latent), S_{X_i} (incomplete, observed) and R_{X_i} (missingness indicator).
- The parents of R_{X_i} determine the pattern: MCAR (no parents), MAR (only fully-observed), MNAR (partially-observed or latent).
- Expectation-Maximisation is the usual way to learn SCMs from incomplete data [16].

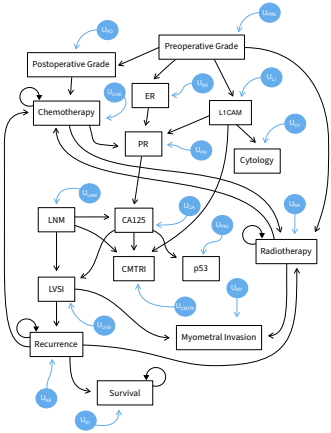


Modelling Sampling Bias

- Sampling bias has a graphical representation in augmented SCMs called selection diagrams [10].
- Treated like a latent confounder: each source of bias is represented as an exogenous variable linked to several endogenous variables.
- G-transportability establishes when study findings carry over from a source population to a different target population, such as from an RCT to observational data.
  Example: cardiovascular disease in young breast cancer survivors [4, 3].



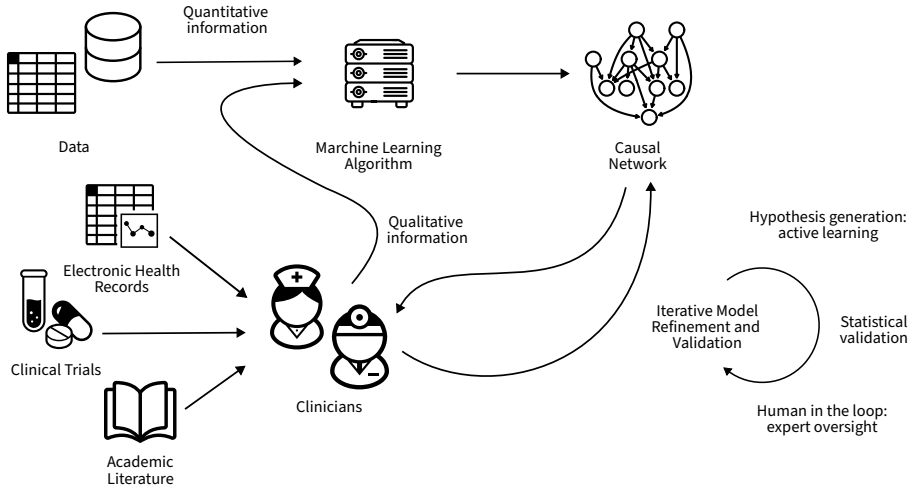




Rolled-Up

As with dynamic BNs, vector auto-regressive time series, Kalman filters and hidden Markov models are dynamic SCMs [16]. Continuous-time SCMs are also possible [5].

Inference for Refinement and Validation



Causal networks are refined and validated iteratively against data and expert knowledge, as we will see with two case studies.

That's all!

Happy to discuss in more detail.

scutari@bnlearn.com

References I

[1] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard.
On Pearl's Hierarchy and the Foundations of Causal Inference, pages 509–56.
Association for Computing Machinery, 2021.

[2] K. Bello, B. Aragam, and P. Ravikumar.
DAGMA: Learning Dags via M-Matrices and a Log-Determinant Acyclicity Characterization.
Advances in Neural Information Processing Systems, 35:8226–8239, 2022.

[3] A. Bernasconi, A. Zanga, P. J. F. Lucas, M. Scutari, …, F. Stella, A. Trama, and the Ada Working Group.
From Real-World Data to Causally Interpretable Models: A Bayesian Network to Predict Cardiovascular Diseases in Adolescents and Young Adults with Breast Cancer.
Cancers, 16(21):3634, 2024.

[4] A. Bernasconi, A. Zanga, P. J. F. Lucas, M. Scutari, A. Trama, and F. Stella.
A Causal Network Model to Estimate the Cardiotoxic Effect of Oncological Treatments in Young Breast Cancer Survivors.
Progress in Artificial Intelligence, 2024.

[5] 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.

References II

[6] D. M. Chickering. Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[7] D. Colombo and M. H. Maathuis. Order-Independent Constraint-Based Causal Structure Learning. *Journal of Machine Learning Research*, 15:3921–3962, 2014.

[8] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning High-Dimensional Directed Acyclic Graphs with Latent and Selection Variables. *The Annals of Statistics*, 40(1):294–321, 2012.

[9] G. Kratzer, F. Lewis, A. Comin, M. Pittavino, and R. Furrer. Additive Bayesian Network Modeling with the R Package Abn. *Journal of Statistical Software*, 105(8), 2023.

[10] S. Lee, J. Correa, and E. Bareinboim. General Transportability—Synthesizing Observations and Experiments from Heterogeneous Domains. *AAAI Conference on Artificial Intelligence*, 34(6):10210–10217, 2020.

References III

[11] A. Li, A. Jaber, and E. Bareinboim.
Causal Discovery from Observational and Interventional Data Across Multiple Environments.
In *Advances in Neural Information Processing Systems*, volume 36, pages 16942–16956, 2023.

[12] R. J. A. Little and D. B. Rubin.
Statistical Analysis with Missing Data.
Wiley, 3rd edition, 2019.

[13] K. Mohan and J. Pearl.
Graphical Models for Processing Missing Data.
Journal of the American Statistical Association, 116(534):1023–1037, 2018.

[14] J. Pearl and D. Mackenzie.
The Book of Why: the New Science of Cause and Effect.
Basic Books, 2018.

[15] J. Peters, J.M. Mooij, D. Janzing, and B. Schölkopf.
Causal Discovery with Continuous Additive Noise Models.
Journal of Machine Learning Research, 15(1):2009–2053, 2014.

References IV

M. Scutari.
[16] Bayesian Network Models for Incomplete and Dynamic Data.
Statistica Neerlandica, 74(3):397–419, 2020.

M. Scutari and J.-B. Denis.
[17] *Bayesian Networks with Examples in R*.
Chapman & Hall, 2nd edition, 2021.

M. Scutari, D. Kerob, J. Krutmann, and S. Salah.
[18] Causal Networks of Infodemiological Data: Modelling Dermatitis.
In *Proceedings of the 23rd International Conference on Artificial Intelligence in Medicine (AIME25)*, pages 397–407, 2025.

S. Shimizu, P. O. Hoyer, A. Hyvärinena, and ‘ A. Kerminen.
[19] A Linear Non-Gaussian Acyclic Model for Causal Discovery.
J. Mach. Learn. Res., 7(72):2003–2030, 2006.

S. Xu, O. A. Mian, A. Marx, and J. Vreeken.
[20] Inferring Cause and Effect in the Presence of Heteroscedastic Noise.
Proceedings of Machine Learning Research (ICML), 162:24615–24630, 2022.

References V

A. Zanga, A. Bernasconi, P. J. F. Lucas, H. Pijnenborg, C. Rejinen, M. Scutari, and F. Stella.

[21] Causal Discovery with Missing Data in a Multicentric Clinical Study.

In *Proceedings of the 21st International Conference on Artificial Intelligence in Medicine (AIME23)*, pages 40–44, 2023.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing.

[22] DAGs with No Tears: Continuous Optimization for Structure Learning.

Advances in Neural Information Processing Systems, 31:0472–9483, 2018.



Case Studies

Immunodeficiencies and Dermatitis

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence

November 27, 2026

→ Adults at Risk for CID and CVID

Infodemiology of Dermatitis

npj | digital medicine

Article

https://doi.org/10.1038/s41746-025-01761-5

Large scale causal modeling to identify adults at risk for combined and common variable immunodeficiencies

Giorgos Papanastasiou[1,8] ✉, Marco Scutari[2,8], Raffi Tachdjian[3,4], Vivian Hernandez-Trujillo[5], Jason Raasch[6], Kaylyn Billmeyer[1,7], Nikolay V. Vasilyev[1] & Vladimir Ivanov[1]

npj Digital Medicine | (2025)8:361

Notebooks: https://www.bnlearn.com/research/pfizer24

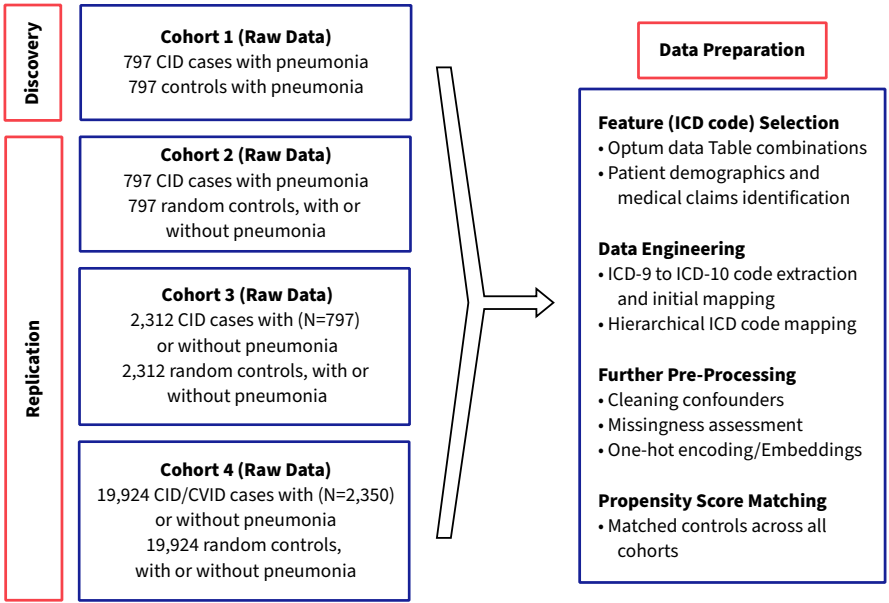
Pfizer

Primary Immunodeficiencies

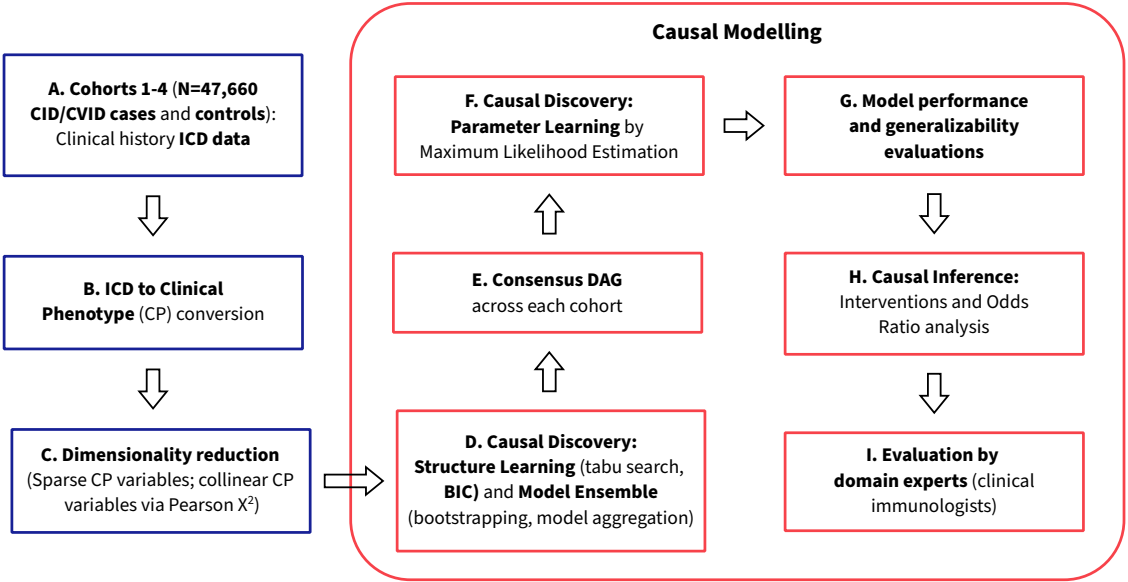
- Goal: unravelling the interplay between clinical diagnosis codes linked to combined immunodeficiencies (CID) and common variable immunodeficiencies (CVID).
- Main variables: clinical history with ICD codes from 4 different cohorts, which differ in whether cases and controls have pneumonia. ICD codes are transformed into phenotypes.
- Possible confounders: none, by assumption (untestable). We condition on all clinical history; the data are nationally representative; cases and controls are 1:1 matched by propensity score on demographics, length of medical history and number of visits.
- Size: $\approx 800/800/2.3k/20k$ cases, each with a matched control; $\approx 550/550/400/300$ phenotypes, $241/245/212/122$ after removing collinear ones, in cohorts 1–4.
- Missing values: none!

Following up on a previous effort based on a custom deep-learning architecture [3].

Preprocessing: Data and ICD Codes



Data Analysis



The Role of Clinical Immunologists

- Telling phenotypes apart.
  - With finite sample sizes and so many variables, some end up being nearly identical (pair-wise association p-value ≈ 0).
  - They explain CID/CVID equally well.
  - Which ones make the most (clinical) sense to keep?
- Validating the causal networks:
  - Reviewing the consensus causal networks from each cohort.
  - Confirming that the networks really are dense.
  - Assessing the parents of CID/CVID, the direct precursors of a diagnosis (arcs out of CID/CVID are blacklisted).

We consulted them independently, effectively handling their observations as an ensemble expert model.

- ✔ Adults at Risk for CID and CVID
- → Infodemiology of Dermatitis

Causal Networks of Infodemiological Data: Modelling Dermatitis

Marco Scutari[1(✉)], Samir Salah[2], Delphine Kerob[2,3], and Jean Krutmann[4,5]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
R. Bellazzi et al. (Eds.): AIME 2025, LNAI 15734, pp. 397–407, 2025.
https://doi.org/10.1007/978-3-031-95838-0_39

Notebooks: https://www.bnlearn.com/research/aime25

IUF
LEIBNIZ-INSTITUT FÜR UMWELT-MEDIZINISCHE FORSCHUNG

LA ROCHE-POSAY
LABORATOIRE PHARMACEUTIQUE



Dermatitis, Mental Conditions, Pollution and Climate Change

- Goal: understanding the effect of pollution and changing weather patterns on mental conditions and dermatitis, and the cascading effect of mental conditions on dermatitis, from infodemiology data (web search frequencies as a proxy for incidence).
- Main variables: 3 pollutants (NO_2, SO_2, $PM_{2.5}$), 3 mental conditions (anxiety, depression, sleep disorders), obesity, dermatitis, weather patterns (temperatures, wind speed, precipitation; both mean and spread).
- Possible confounders: education level, unemployment, income, household size and population density.
- Size: $\approx$53k observations over $\approx$500 US counties and 134 weeks.
- Missing values: between 0% (the conditions) and 55% (pollutants).

Following up on a previous infodemiology study [4].

Data Sources: Google Trends, NOAA, EPA, US Census

Google

Google COVID-19 Open Data: 400 health conditions, 4 countries (county-level in the US), weekly search frequencies for 2020-2023 normalised by NLP.

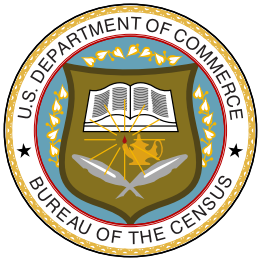
Weather stations in 1652 counties with and satellite images.



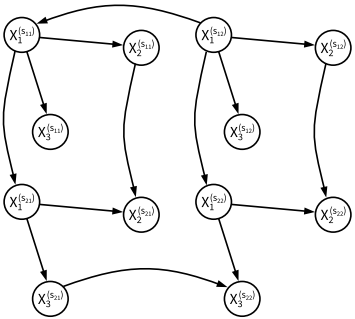


Monitoring stations in 1470 counties with hourly measurements of NOx, SOx, O3, PMx.

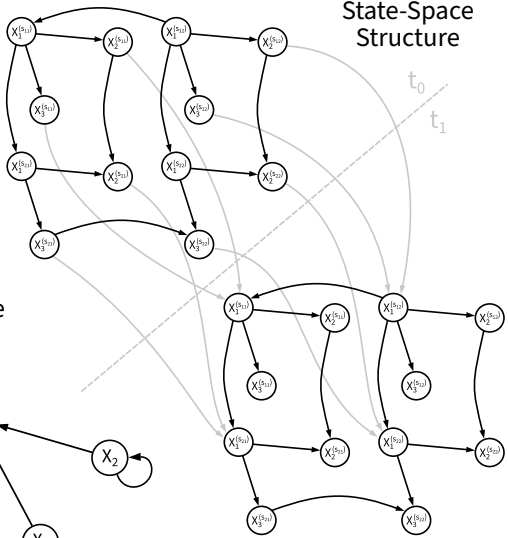
Socio-economic data at the population level to avoid confounding.



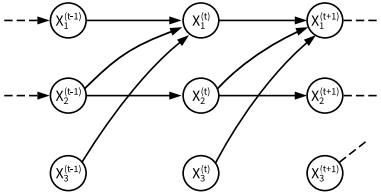
Spatial Structure



State-Space Structure



Temporal Structure (dynamic BNs)



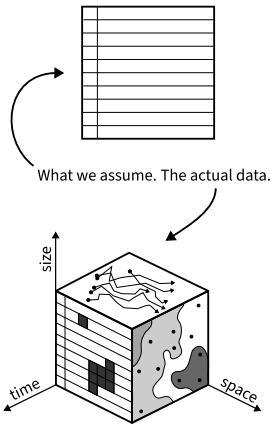
Data Analysis

1. Data fusion and preprocessing.
2. Causal discovery assuming IID data.
3. Statistical validation of the residuals.
4. Causal discovery with a spatial correlation structure.
5. Check the residuals, Bayes factors.
6. Causal discovery with spatial correlation + heterogeneity.
7. Check the residuals, Bayes factors, imposed sparsity level.
8. Predictive accuracy assessment.

Learned a dynamic BN (Part 1) encoding a first-order VAR process:

$$X_{it} = f_i(\Pi_{X_{it}}\boldsymbol{\beta}_{it}) + \varepsilon_{it}, \qquad \mathrm{COV}(\varepsilon_{it}) = \mathbf{w}_{it}^{\mathrm{T}}\boldsymbol{\Sigma}_i(\mathbf{L};\xi_i)\mathbf{w}_{it}.$$

- $\boldsymbol{\Sigma}_i(\mathbf{L};\xi_i)$ models spatial correlation via generalised least squares (GLS); location $\mathbf{L}$ and correlation decay ξ_i.
- The $\mathbf{w}_{it}$ handle
  - heteroscedasticity, via iteratively reweighted least squares (IRLS);
  - missing values, with 0-1 weights like the PNAL score [1] (if MCAR) or inverse-probability weights like HC-aIPW [2] (if MAR or MNAR).



Model averaging (Part 1): bagging with data-driven threshold [5].

Causal Inference: What Conclusions Can We Draw?

- What is the relative impact of the direct risk factors?

  ANX (0.574), NO_2 (0.339), OBE (0.077), $PM_{2.5}$, RANGETEMP, SO_2 (0.01).

- What proportion of pollution effects is mediated?

  $PM_{2.5}$, NO_2 and SO_2 change by 0.54x, 0.93x and 0.56x.

- What proportion of weather effects is mediated?

  TEMP/RANGETEMP, WIND/RANGEWIND, RAIN change by 0.29x, 0.38x, 0.02x.

- What would be the impact of tightening environmental regulations?

  $PM_{2.5}$ $12 \rightarrow 9 \mu g/m^3$ for 1 year: -18% DER. $PM_{2.5}$ $12 \rightarrow 8 \mu g/m^3$: -21% DER.

- How long must a cold spell last before dermatitis increases?

  DER $+5\%$ after 4 weeks.

That's all!

Happy to discuss in more detail.

scutari@bnlearn.com

References I

T. Bodewes and M. Scutari.
[1] Learning Bayesian Networks from Incomplete Data with the Node-Averaged Likelihood.
International Journal of Approximate Reasoning, 138:145–160, 2021.

Y. Liu and A. C. Constantinou.
[2] Greedy Structure Learning From Data That Contain Systematic Missing Values.
Machine Learning, 111(10):3867–3896, 2022.

G. Papanastasiou, G. Yang, D. I. Fotiadis, N. Dikaios, C. Wang, A. Huda, L. Sobolevsky, J. Raasch, E. Perez, G. Sidhu, and
[3] D. Palumbo.
Large-Scale Deep Learning Analysis to Identify Adult Patients at Risk for Combined and Common Variable Immunodeficiencies.
Communications Medicine, 3(1):189, 2023.

M. Scutari, D. Kerob, and S. Salah.
[4] Inferring Skin-Brain-Skin Connections from Infodemiology Data Using Dynamic Bayesian Networks.
Scientific Reports, 14:10266, 2024.

M. Scutari and R. Nagarajan.
[5] On Identifying Significant Edges in Graphical Models of Molecular Networks.
Artificial Intelligence in Medicine, 57(3):207–217, 2013.



Causality and Experimental Design

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence

November 27, 2026

Experimental Design and Causality: Many Correspondences

Experimental Design	Causality	Link
Randomisation	Intervention	Both remove the arcs into the treatment.
Blocking, ANCOVA	Back-door adjustment	Conditioning to compare like with like.
Treatment effects	Counterfactuals	Average effects against individual ones.
External validity	Transportability	Standardising over differences.
Non-compliance	Instrumental variables	Only the assignment is randomised.
Intercurrent events	Colliders, mediators	Do not condition on effects of treatment.
Factorial designs	Structural equations	Assuming negligible interactions.
Dropout	Missingness graphs	Nodes determine what we can observe.
Choosing experiments	Equivalence classes	Disambiguating causes with interventions.

→ Randomisation and Interventions

Blocking and Back-Door Adjustment

Treatment Effects and Counterfactuals

External Validity and Transportability

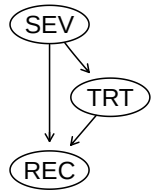
In a randomised experiment, a chance device decides which treatment each unit receives [3].

- The assignment is independent from the unit: prognosis, preferences, the physician's judgement.
- The arms are then comparable: on average, they differ only in the treatment. This holds for the characteristics we measured and for those we did not.
- The difference in the outcome between the arms estimates the effect of the treatment.
- Without randomisation, who is treated depends on those characteristics. Say, for instance, that sicker patients get the new drug; comparing treated and untreated patients mixes the effect of the drug with that of being sicker.

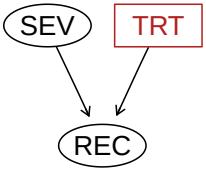
Interventions: Removing the Arcs into the Treatment

- In observational data, disease severity SEV drives the choice of treatment TRT. Observing $\mathrm{TRT} = t$ tells us something about SEV, so $\mathrm{P}(\mathrm{REC} \mid \mathrm{TRT} = t)$ mixes their effects on recovery REC.
- An intervention removes all the arcs into TRT and leaves the rest of the network as it is [4]. Then:
  - hard, $\mathrm{do}(\mathrm{TRT} = t)$: fix the treatment to the value t;
  - soft, $\mathrm{do}(\mathrm{TRT} \sim \mathrm{RND})$: draw it from a new parent RND.
- $\mathrm{P}(\mathrm{REC} \mid \mathrm{do}(\mathrm{TRT} = t)) = \mathrm{P}(\mathrm{REC} \mid \mathrm{TRT} = t)$ in the new network. In the original one, they are different.

Observational Network



After do(TRT = t)



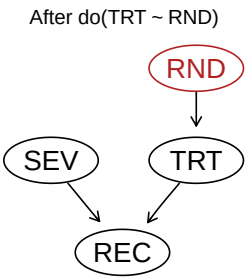
Randomisation Is an Intervention

- Randomisation is the soft intervention $\mathrm{do}(\mathrm{TRT} \sim \mathrm{RND})$ on the patients: the chance device RND replaces the parents of TRT.
- Reading one arm of the trial gives the hard intervention:

$$\mathrm{P}(\mathrm{REC} \mid \mathrm{TRT} = t; \mathrm{do}(\mathrm{TRT} \sim \mathrm{RND})) = \mathrm{P}(\mathrm{REC} \mid \mathrm{do}(\mathrm{TRT} = t)).$$

Comparing across arms now makes sense.

- We removed the arc from SEV without measuring it: this is “comparable arms” in graphical terms.
- Without randomisation, we must know and measure SEV.



- ✓ Randomisation and Interventions
- → **Blocking and Back-Door Adjustment**
- Treatment Effects and Counterfactuals
- External Validity and Transportability

Units may differ in ways that affect the outcome: centre, sex, severity. A randomised block design groups similar units into blocks and randomises the treatment within each block [3]:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}, \qquad \tau_i \text{ treatment effect, } \beta_j \text{ block effect.}$$

Block what you can, randomise what you cannot:

- Treatments are compared within blocks, so the differences between blocks do not end up in the error: more precision from the same number of units.
- Stratified randomisation is the same idea in clinical trials.
- Analysis of covariance adjusts for a covariate in the analysis instead of the design.

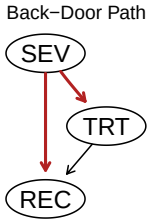
- A back-door path starts with an arc into the treatment, like $\mathrm{TRT} \leftarrow \mathrm{SEV} \rightarrow \mathrm{REC}$. It encodes an association that is not the effect of the treatment.

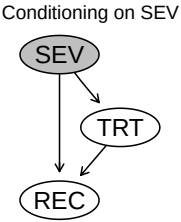
- Conditioning on SEV blocks the path by graphical separation:

$$\mathrm{P}(\mathrm{REC} \mid \mathrm{do}(\mathrm{TRT} = t)) = \sum_s \mathrm{P}(\mathrm{REC} \mid \mathrm{TRT} = t, \mathrm{SEV} = s)\, \mathrm{P}(\mathrm{SEV} = s).$$

  In general: a set of variables that blocks all back-door paths and has no descendants of the treatment [4].

- Compare treatments within each level of SEV, then average over the levels: a blocked analysis with SEV as the blocking factor.





Same Operation, Different Purpose

	Observational Study	Randomised Experiment
Why adjust?	Remove bias: the parents of the treatment affect the outcome.	Increase precision: the outcome has other causes.
Which variables?	A set of variables that blocks all back-door paths, like SEV.	The outcome's causes that are measured before randomisation.
How?	Stratification, regression, matching.	Stratified randomisation, blocking, analysis of covariance.

In both cases, the network tells us which variables we should use.

- ✓ Randomisation and Interventions
- ✓ Blocking and Back-Door Adjustment
- → **Treatment Effects and Counterfactuals**
- External Validity and Transportability

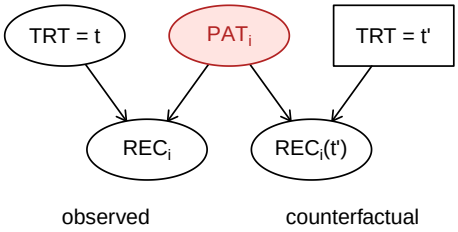
Treatment Effects

- In a parallel-group trial, each patient receives one treatment. The effect we estimate is the difference in the mean outcome between the arms: an average over patients.
- An average can hide variation: a treatment that helps most patients and harms a few has a positive average effect.
- The per-patient effect is the difference between the outcomes of the same patient under the two treatments. We can only ever measure one of them.
- In a crossover trial, each patient receives both treatments in sequence. The comparison is within patients, but it is valid only if the first treatment leaves no trace (no carry-over) and the patient does not change in between (no period effect).

The potential outcome $\mathrm{REC}_i(t) = f_{\mathrm{REC}}(t, \mathrm{PAT}_i)$ is the recovery of patient i under treatment t [6]; PAT_i is everything else about the patient (exogenous).

Having observed t and REC_i, $\mathrm{REC}_i(t')$ is a counterfactual:

1. learn PAT_i from the observed data;
2. apply $\mathrm{do}(\mathrm{TRT} = t')$;
3. recompute REC with the same PAT_i.



Average effects need only 2, from the trial. Individual effects also need f_{REC} and PAT_i [2].

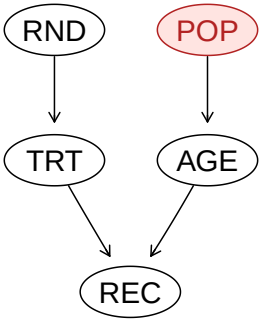
- ✓ Randomisation and Interventions
- ✓ Blocking and Back-Door Adjustment
- ✓ Treatment Effects and Counterfactuals
- → External Validity and Transportability

External Validity

- Internal validity: the estimated effect is right for the patients in the trial, thanks to randomisation.
- External validity: the estimated effect is right for the patients who will be treated. No direct effect from randomisation.
- Trial patients are selected by eligibility criteria, sites and consent. They are usually younger and with fewer comorbidities than the target population.
- It matters when the selection characteristics are effect modifiers: the effect of the treatment depends on them.
- The classical remedy is standardisation [1]: estimate the effect within the levels of the modifier, then reweight the levels to match the target population.

Transportability: Adding Selection Nodes

- A selection diagram is a single network modelling both populations: the selection node POP points at what differs between them, here AGE.



- If AGE separates POP from REC after $\mathrm{do}(\mathrm{TRT} = t)$, the trial is transportable [5] to the target population P^* by standardisation:

$$\mathrm{P}^*(\mathrm{REC} \mid \mathrm{do}(\mathrm{TRT} = t)) = \sum_a \underbrace{\mathrm{P}(\mathrm{REC} \mid \mathrm{do}(\mathrm{TRT} = t), \mathrm{AGE} = a)}_{\text{trial}} \underbrace{\mathrm{P}^*(\mathrm{AGE} = a)}_{\text{target}}.$$

- If POP points at REC, we need outcome data from the target population to infer the outcome.

That's all!

Happy to discuss in more detail.

scutari@bnlearn.com

References I

[1] M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

[2] P. W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[3] D. C. Montgomery. *Design and Analysis of Experiments*. Wiley, 10th edition, 2020.

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

[5] J. Pearl and E. Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579–595, 2014.

References II

D. B. Rubin.

[6] Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.
Journal of Educational Psychology, 66(5):688–701, 1974.