Causal Modelling in Time and Space

State-Space Networks from Incomplete Data

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence (IDSIA)

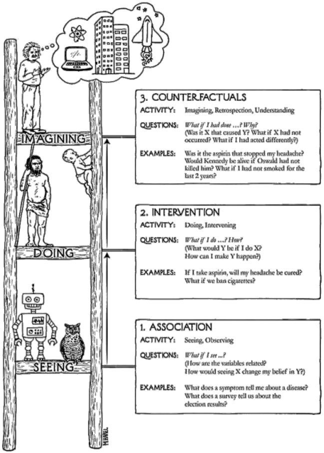
July 08, 2024

Machine learning creates black boxes that use probabilistic associations for prediction. Scientific questions are inherently causal.

Judea Pearl [10] has worked out a rigorous theory of causality that uses directed (acyclic) graphs to represent causes and effects. With it, we can reason about

- what we see,
- affecting change,
- hypothetical situations.

How can we learn them?

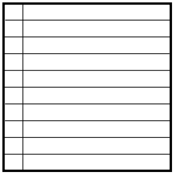


Causal Discovery Meets Data

Learning a causal network means learning its structure $\mathcal{G}$ and parameters Θ, much like Bayesian networks:

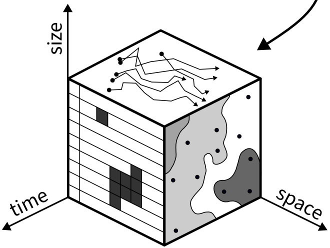
$$\underbrace{\mathrm{P}(\mathcal{G}, \Theta \mid \mathcal{D})}_{\text{learning}} = \underbrace{\mathrm{P}(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}} .$$

We used to ask domain experts for information [5, 6]; now we rely increasingly on learning algorithms and the data $\mathcal{D}$ [11].



What we assume.

The actual data.



The Context: Infodemiology

- Combinations of comorbidities are often impossible to study in a clinical trial.
- However, we have massive amounts of Internet-generated data user-contributed health-related content.
- Infodemiology (short for "information epidemiology") draws on this data to replace epidemiological data with the ultimate goal of improving public health.

We need to assume:

- a non-negligible association between the frequency of online mentions of specific diseases and their incidence;
- a broad coverage of the population.

A motivating example: understanding the effect of pollution and changing weather patterns on mental and dermatological conditions.

- Main Variables: 3 pollutants (NO_2, SO_2, PM2.5), 3 mental conditions (anxiety, depression, sleep disorders), obesity, atopic dermatitis, weather patterns (temperatures, wind speed, precipitations; 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 from a previous infodemiology study [12].

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

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.

Nuisance Parameters, Parameters of Interest

A causal network has two components: the graph $\mathcal{G}$ and the parameters Θ. Causal inference defines queries using $\mathcal{G}$:

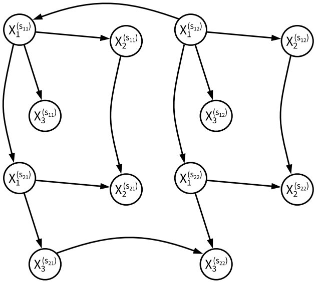
- Conditional independence, via d-separation.
- Intervention, via mutilation.
- Counterfactual, via the twin network.

Our ability to answer scientific questions using the causal network rests on having the right nodes in the network. Without them, we cannot even formulate our question.

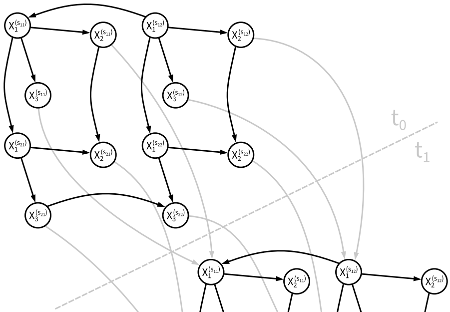
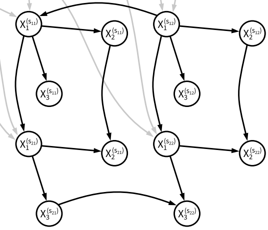
- The dimensions we use in the queries (interest) should be represented as nodes.
- The dimensions we do not (nuisance) should be represented as parameters in the local distributions.

Network Structures: Time vs Space vs State-Space

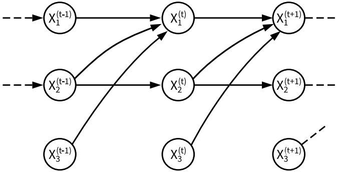
Spatial Structure

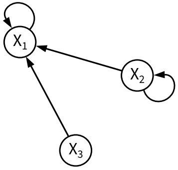


State-Space Structure

Temporal Structure (dynamic BNs)





My proposal is to use the local distributions:

$$X_i = \mu_{X_i} + \Pi_{X_i}\boldsymbol{\beta}_{X_i} + \boldsymbol{\varepsilon}_{X_i}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \Sigma_{X_i} + \mathbf{K}_{X_i})$$

where:

- $\Sigma_{X_i} = \sigma^2_{X_i} \operatorname{diag}(\mathbf{w}_{X_i})$ is heteroscedastic noise estimated by iteratively reweighted least squares (IRLS);
- $\mathbf{K}_{X_i}$ is the observations correlation structure not otherwise modelled by $\mathcal{G}$, via generalised least squares (GLS).

Score function: the penalised node-average log-likelihood (PNAL) for incomplete data [4]:

$$\mathrm{PNAL}(X_i \mid \Pi_{X_i}) = \bar{\ell}(X_i \mid \Pi_{X_i}) + \lambda_n \left|\Theta_{X_i}\right|.$$

Denoising: bagging and model averaging with data-driven threshold [14].

- I care about time, but I do not care about space.
- I need different residual variances in different states due to how the data are normalised.

I want to learn a dynamic BN that encodes a first-order vector auto-regressive process (VAR):

$$X_i^{(t)} = \mu_{X_i}^{(t)} + \Pi_{X_i}^{(t-1)} \boldsymbol{\beta}_{X_i}^{(t)} + \boldsymbol{\varepsilon}_{X_i}^{(t)},$$
$$\boldsymbol{\varepsilon}_{X_i}^{(t)} \sim N\left(\mathbf{0}, \sigma_{X_i}^{2\,(t)} \operatorname{diag}\left(\mathbf{w}_{X_i}^{(t)}\right) + \mathbf{K}_{X_i}\right)$$

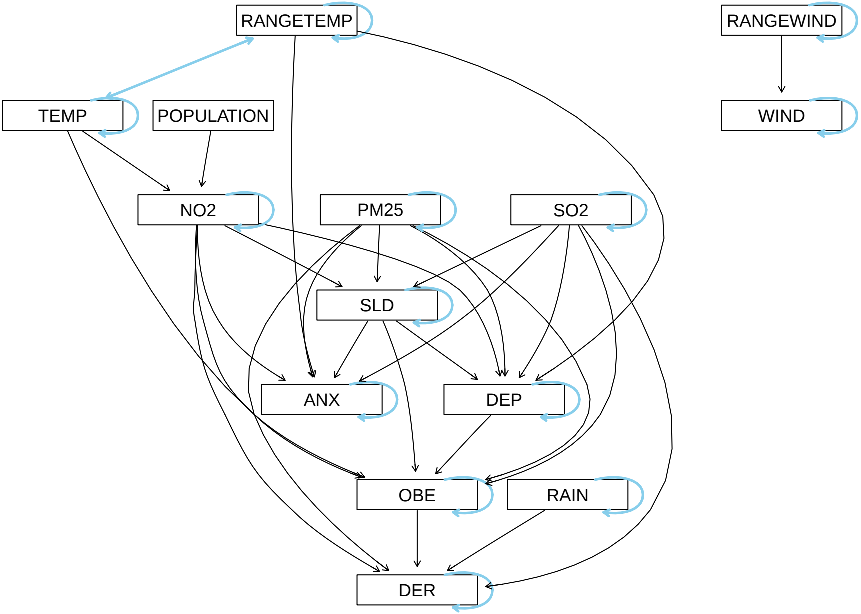
with

$$\mathbf{w}_{X_i}^{(t)} \propto 1/\operatorname{VAR}(\boldsymbol{\varepsilon}_{X_i} \mid \text{state}), \quad \mathbf{K}_{X_i} = f\left(\|\text{latitude and longitude}\|_2; \xi\right).$$

Code: The R Implementation

```
# provide an initial estimate.
model = nlme::gls(as.formula(f), data = full, method = "ML",
            cor = nlme::corExp(value = args$spatial[, node],
                    form = ~ LAT + LON | WEEK, nugget = TRUE, fixed = TRUE))
old.logl = as.numeric(nlme:::logLik.gls(model), REML = FALSE)

# iteratively reweighted least squares.
for (iter in 1:(args$irls.max.iter)) {

  # compute the per-state variances...
  weights = sapply(levels(full[, "STATE"]), function(s) var(resid(model)[full[, "STATE"] == s]) )
  for (i in seq(nrow(full)))
    full[i, "w"] = weights[names(weights) == full[i, "STATE"]]
  # ... and re-estimate the model.
  model = nlme::gls(as.formula(f), data = full, method = "ML",
              cor = nlme::corExp(value = args$spatial[, node],
                      form = ~ LAT + LON | WEEK, nugget = TRUE, fixed = TRUE),
              weights = nlme::varFixed(~ w))
  new.logl = as.numeric(nlme:::logLik.gls(model, REML = FALSE))

  # check convergence.
  if (isTRUE(all.equal(old.logl, new.logl)))
    break
  else
    old.logl = new.logl

}#FOR
```

Incomplete Data + Time (Looks Very Wrong)



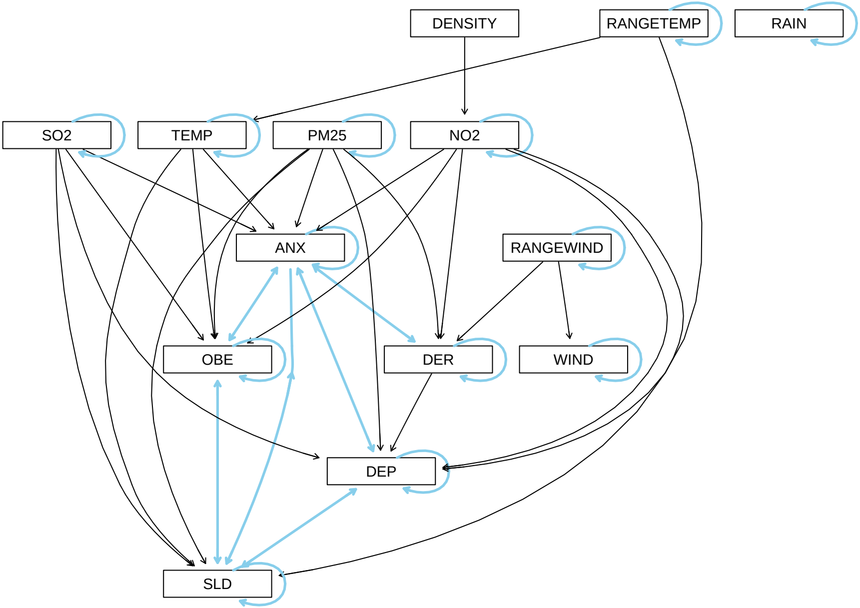
Residuals are largely free from autocorrelation! ✔

	lag 1	lag 2	lag 3	lag 4
ANX	0.024	0.000	0.000	0.048
DEP	0.016	0.000	0.000	0.000
DER	0.032	0.000	0.000	0.000
OBE	0.000	0.000	0.000	0.000
SLD	0.092	0.007	0.007	0.000

But they are full of spatial correlation! ✖

	proportion
ANX	0.460
DEP	0.325
DER	0.754
OBE	0.563
SLD	0.381

Incomplete Data + Space + Time (Looks Less Wrong)



The causal network fits the data much better! ✅

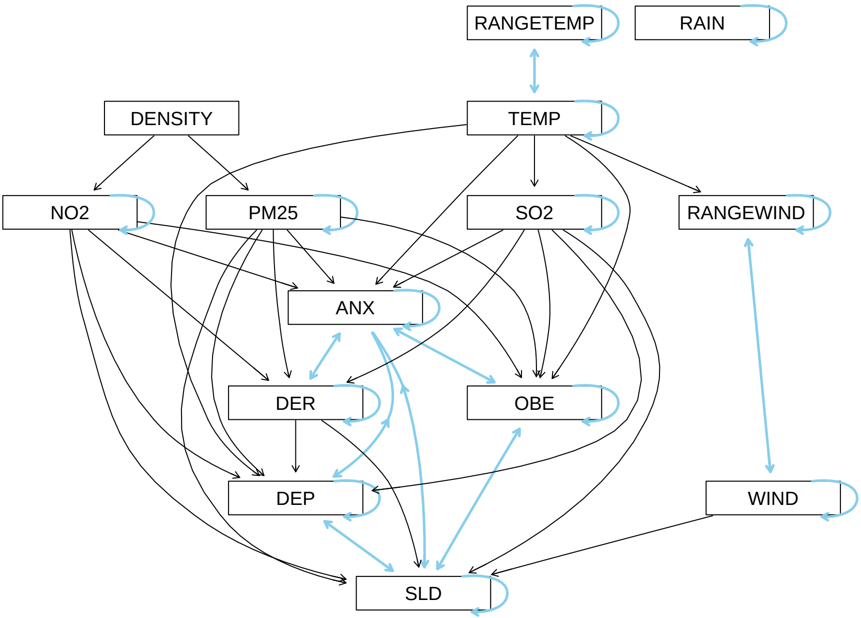
$$\log \mathrm{BF} = (-26.83) - (-31.23) = 4.4 \qquad \Longrightarrow \qquad \mathrm{BF} = 81.59.$$

But the residuals are markedly heteroscedastic! ❌

	p-value
ANX	4×10^{-169}
DEP	1×10^{-212}
DER	0
OBE	6×10^{-100}
SLD	1×10^{-154}

One more (and last) time...

Incomplete Data + Space + Time + Heteroscedasticity (Looks OK)



The causal network fits the data much better! ✅

$$\log \mathrm{BF} = (-23.6) - (-26.83) = 3.23 \qquad \Longrightarrow \qquad \mathrm{BF} = 25.31.$$

The weighted residuals are completely homoscedastic! ✅

	p-value
ANX	1
DEP	1
DER	1
OBE	1
SLD	1

- The causal network is completely identifiable because:
  - Arc directions across time points are fixed.
  - Heteroscedastic residuals + Gaussian noise [7, 17, 18].
  - Even if all $\mathbf{w}_{X_i}^{(t)} = 1$, the actual residuals $(\mathbf{K}_{X_i}^{(t)})^{-1/2}\boldsymbol{\varepsilon}_{X_i}^{(t)}$ should be heteroscedastic unless $\mathbf{K}_{X_i}^{(t)} \propto \mathbf{I}_n$.
- If we use $\mathbf{K}$ to model temporal dependencies, it can encode a full vector ARMA process [16].
- $\mathcal{G}$ requires equidistant points; $\mathbf{K}$ can accommodate irregularly spaced points in time or space.

Computational Considerations

- GLS scales $O(n^3)$, (sparse) causal discovery scales $O(|\mathbf{X}|^2)$.
- Divide and conquer works wonders:
  - The parameters ξ of $\mathbf{K}$ are (almost) independent from Π_{X_i}: we can pre-estimate them and keep them fixed during causal discovery.
  - It is much faster to estimate the $\mathbf{w}_{X_i}^{(t)}$ by wrapping GLS in IRLS than doing so directly in GLS.
- Still, imposing sparsity is critical. Subsampling within model averaging and blacklisting arcs help as well.
- PNAL is so much faster than Structural EM [8, 9] that causal discovery from incomplete data becomes feasible.

- Using GLMs is straightforward because we can estimate them with IRLS, which we already use, and allows for discrete variables.
- Bringing change point detection from the literature on VARs [1, 2].
- A more robust handling of missing values, proving that PNAL works under MAR or leveraging my students' work on causal discovery under MNAR [3, 19, 20].
- Incorporating random effects to separate global and local effects (in time/space/sub-populations) from my previous work [13, 15].

- Causal discovery makes simplifying assumptions that are too strong.
- Classical statistics gives us flexible and scalable tools to model complex structures inn the data.
- Pose the research question first: model the data dimensions you need graphically, and hide the rest in the local distributions.
- State-space data, mixed variable types, missing values, population structure, non-stationarity: we can deal with them!

Acknowledgements

Federico Maria Stefanini

Università degli Studi di Milano

Alice Bernasconi

Alessio Zanga

Fabio Stella

Università degli Studi di Milano-Bicocca

Samir Salah

Delphine Kerob

L'Oréal, La Roche-Posay

My former students: Tjebbe Bodewes (University of Oxford), Lorenzo Valleggi (Università degli Studi di Firenze).

That's all!

Happy to discuss in more detail.

- P. Bai, A. Safikhani, and G. Michailidis.
  Multiple Change Points Detection in Low Rank and Sparse High Dimensional Vector Autoregressive Models.
  IEEE Transactions on Signal Processing, 68:3074–3089, 2020.
- P. Bai, A. Safikhani, and G. Michailidis.
  Multiple Change Point Detection in Reduced Rank High Dimensional Vector Autoregressive Models.
  Journal of the American Statistical Association, 118(544):2776–2792, 2023.
- A. Bernasconi, A. Zanga, P. J. F. Lucas, M. Scutari, and F. Stella.
  Towards a Transportable Causal Network Model Based on Observational Healthcare Data.
  In *AIxIA*, pages 67–82, 2023.
- T. Bodewes and M. Scutari.
  Learning Bayesian Networks from Incomplete Data with the Node-Averaged Likelihood.
  International Journal of Approximate Reasoning, 138:145–160, 2021.
- M. J. Druzdel and L. C. van der Gaag.
  Elicitation of Probabilities for Belief Networks: Combining Qualitative and Quantitative Information.
  In *UAI*, pages 141–148, 1995.

- M. J. Druzdel and L. C. van der Gaag.
  Building Probabilistic Networks: “Where Do the Numbers Come From?”.
  IEEE Transactions on Knowledge and Data Engineering, 12(4):481–486, 200.
- B. Duong and T. Nguyen.
  Heteroscedastic Causal Structure Learning, 2023.
- N. Friedman.
  Learning Belief Networks in the Presence of Missing Values and Hidden Variables.
  In *ICML*, pages 125–133, 1997.
- N. Friedman.
  The Bayesian Structural EM Algorithm.
  In *UAI*, pages 129–138, 1998.
- J. Pearl and D. Mackenzie.
  The Book of Why: the New Science of Cause and Effect.
  Basic Books, 2018.

- M. Scutari, C. E. Graafland, and J. M. Gutiérrez.
  Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms.
  International Journal of Approximate Reasoning, 115:235–253, 2019.
- M. Scutari, D. Kerob, and S. Salah.
  Inferring Skin-Brain-Skin Connections from Infodemiology Data Using Dynamic Bayesian Networks.
  Scientific Reports, 14:10266, 2024.
- M. Scutari, C. Marquis, and L. Azzimonti.
  Using Mixed-Effect Models to Learn Bayesian Networks from Related Data Sets.
  Proceedings of Machine Learning Research (PGM 2022), 186:73–84, 2022.
- M. Scutari and R. Nagarajan.
  On Identifying Significant Edges in Graphical Models of Molecular Networks.
  Artificial Intelligence in Medicine, 57(3):207–217, 2013.
- L. Valleggi, M. Scutari, and F. M. Stefanini.
  Learning Bayesian Networks with Heterogeneous Agronomic Datasets via Mixed-Effect Models and Hierarchical Clustering.
  Engineering Applications of Artificial Intelligence, 131:107867, 2024.

- J. van der Leeuw.
  The Covariance Matrix of ARMA Errors in Closed Form.
  Journal of Econometrics, 63:397–405, 1994.
- S. Xu, O. A. Mian, A. Marx, and J. Vreeken.
  Inferring Cause and Effect in the Presence of Heteroscedastic Noise.
  ICML, 162:24615–24630, 2022.
- N. Yin, T. Gao, Y. Yu, and Q. Ji.
  Effective Causal Discovery under Identifiable He,teroscedastic Noise Model.
  In *AAAI Conference on Artificial Intelligence*, volume 38, pages 16486–16494, 2024.
- A. Zanga, A. Bernasconi, P. Lucas, H. Pijnenborg, C. Reijnen, M. Scutari, and F. Stella.
  Risk Assessment of Lymph Node Metastases in Endometrial Cancer Patients: A Causal Approach.
  In *AIxIA*, pages 1–15, 2022.
- A. Zanga, A. Bernasconi, P. J. F. Lucas, H. Pijnenborg, C. Rejinen, M. Scutari, and F. Stella.
  Causal Discovery with Missing Data in a Multicentric Clinical Study.
  In *AIME*, pages 40–44, 2023.