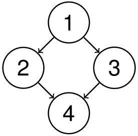
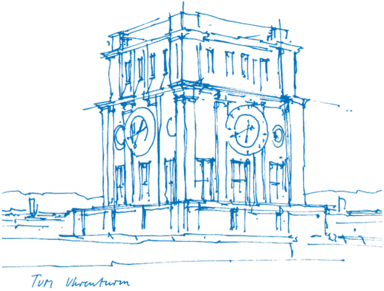
Identifiability and estimation of recursive max-linear models

Claudia Klüppelberg
with Nadine Gissible and Steffen Lauritzen

Technical University of Munich

Munich, October 2019





Ver Hoef, J.M., Peterson, E. and Theobald, D. (2006) Spatial statistical models that use flow and stream distance.

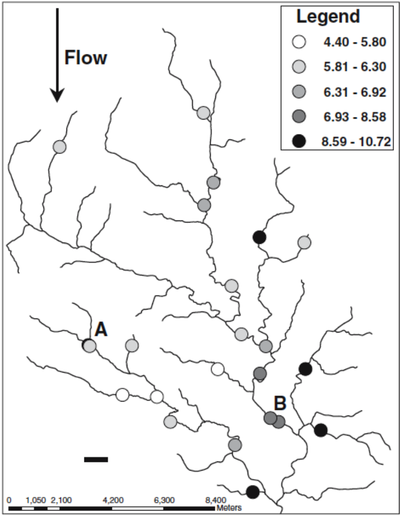


Fig. 2 An example of how flow affects stream chemistry values. Note that there are two locations in close proximity with overlapping circles near A

Asadi, P., Davison, A.C. and Engelke, S. (2015) Extremes on river networks.

2 P. ASADI, A. C. DAVISON, AND S. ENGELKE

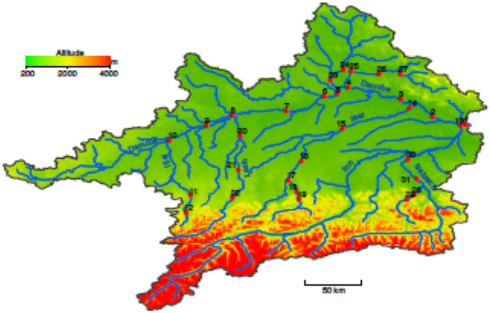
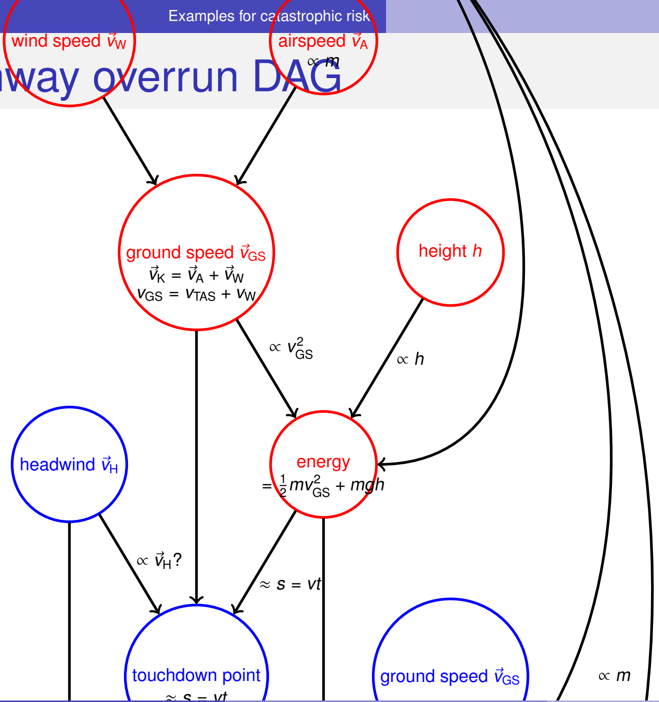


FIGURE 1. Topographic map of the upper Danube basin, showing sites of 31 gauging stations (red blobs) along the Danube and its tributaries. Water flows broadly from left to right.

Runway overrun DAG



Graphical models for causal risk modelling

Goal: Establish cause-effect relations from observational data

Causation fallacy: Correlation does not imply causation

Structural equation models: Represent the underlying causal mechanism in terms of a directed acyclic graph (DAG).

Advantage:
The edge orientations give an intuitive causal interpretation.

In the literature:
Mainly discrete models and Gaussian models, and correlation as dependence measures.

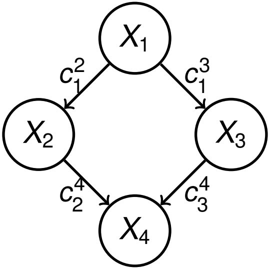
Correlation fallacy: Correlation does not imply high risk dependence
Frequency without severity fallacy: High risk is in the severities

Distribution fallacy: Normal distributions underestimate high risks

Recursive max-linear graphical models [Gissibl & K. (2018)]

For $Z_1, \ldots, Z_d > 0$ independent, continuous, unbounded support, and edge weights $c_{ik} > 0$, we define the **recursive max-linear model**[1]

$$X_i := \bigvee_{k \in \mathrm{pa}(i)} c_{ik} X_k \vee Z_i \quad i = 1, \ldots, d$$



In matrix form: $X = C \odot X \vee Z$,

with solution[a] (B is the **Kleene star matrix**)

$$X = B \odot Z$$
$$\text{and } B = \bigvee_{k=0}^{d-1} C^{\odot k} = (I_d \vee C)^{\odot(d-1)}$$

[a] Algebra: Butkovic (2010)

[1] SEM: Pearl (2009), Graphical Model: Lauritzen (1996)

Path analysis: what is the Kleene star matrix B in the DAG?

For a path $p = [j = k_0 \to k_1 \to \cdots \to k_n = i]$ define the **path-weights**

$$d_{ji}(p) := \prod_{i=0}^{n-1} c_{k_i}^{k_{i+1}}$$

$$b_{ji} := \bigvee_{p \in P_{ji}} d_{ji}(p) \quad \forall j \in \mathrm{an}(i), \quad b_{ii} = 1 \quad \text{and all other} \quad b_{ji} = 0.$$

Then[2] $\quad X_i = \bigvee_{j \in \mathrm{An}(i)} b_{ji} Z_j \quad i = 1, \ldots, d,$

- Path(s) of $\mathbb{D}$, which realize b_{ji} are called **max-weighted path.**
- We can remove an edge from $\mathbb{D}$, which is not part of a max-weighted path without changing the distribution of X.
- The DAG obtained in this way is called the **minimum max-linear DAG**, denoted by $\mathbb{D}^B$.

[2] Max-linar model: Wang and Stoev (2011), no graphs

Model identifiability: $X = C \odot X \vee Z$ has solution $X = B \odot Z$

- Only B and the corresponding DAG $\mathbb{D}^B$ are identifiable.
- Use $Y_{ji} := X_i / X_j$ to identify B from the distribution of X: The innovations Z_i have no atoms, but for $j \in \mathrm{An}(i)$,

$$X_i = \bigvee_{k \in \mathrm{An}(i)} b_{ki} Z_k \geq \bigvee_{k \in \mathrm{An}(j)} b_{ki} Z_k \geq \bigvee_{k \in \mathrm{An}(j)} b_{kj} b_{ji} Z_k = b_{ji} \bigvee_{k \in \mathrm{An}(j)} b_{kj} Z_k = b_{ji} X_j$$

$$\mathrm{supp}(Y_{ji}) = \begin{cases} [b_{ji}, \infty) & j \in \mathrm{an}(i) \\ (0, 1/b_{ij}] & j \in \mathrm{de}(i) \\ (0, \infty) & \text{otherwise} \end{cases}$$

Distributional properties of $Y_{ji} = X_i/X_j$

Relationship between i and j	(Y_{ji})	Atoms
$j \in \mathrm{an}(i)$	$[b_{ji}, \infty)$	$\{b_{\ell i}/b_{\ell j}, \ell \in \mathrm{An}(j)\}$
$i \in \mathrm{an}(j)$	$(0, 1/b_{ij}]$	$\{b_{\ell i}/b_{\ell j}, \ell \in \mathrm{An}(i)\}$
otherwise:		
if $\mathrm{an}(i) \cap \mathrm{an}(j) \neq \emptyset$	$\mathbb{R}_+$	$\{b_{\ell i}/b_{\ell j}, \ell \in \mathrm{an}(i) \cap \mathrm{an}(j)\}$
if $\mathrm{an}(i) \cap \mathrm{an}(j) = \emptyset$	$\mathbb{R}_+$	$\emptyset$

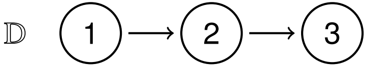
Estimation of B for known DAG: $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}_+^d$

This suggests the **minimum ratio estimator**[3]

$$\breve{b}_{ji} = \bigwedge_{t=1}^{n} y_{ji}^{(t)} = \bigwedge_{t=1}^{n} \frac{x_i^{(t)}}{x_j^{(t)}} \text{ for } j \in \mathrm{an}(i), \quad \breve{b}_{ii} = 1, \ \ \breve{b}_{ji} = 0 \text{ for } j \in V \setminus \mathrm{An}(i).$$

Problem: $\breve{B}$ not necessarily admissible for $\mathbb{D}$.

Example:



Assume we observe $\breve{b}_{13} > \breve{b}_{12}\breve{b}_{23}$, then $\breve{B}$ is not admissible for $\mathbb{D}$ as $1 \to 3$ is estimated to be max-weighted.

[3] Davis and Resnick (1989) suggest this estimator for max-ARMA time series.

Estimation of B for known DAG

Lemma. Let B be a matrix with non-neg. entries and diagonal 1. Define $B_0 = (b_{ji}\mathbf{1}_{\mathrm{pa}(i)}(j))_{d\times d}$. Then B is admissible for $\mathbb{D}$ if and only if

$$b_{ji} > 0 \iff j \in \mathrm{an}(i) \quad \text{and} \quad B = I_d \vee (B \odot B_0).$$

The rhs has unique solution $B = (I_d \vee B_0)^{\odot(d-1)}$. □

Consequently, we estimate B_0 by $\breve{B}_0$, the minimum ratio estimator

$$\breve{b}_{ji} = \widehat{b_{ji}} = \bigwedge_{t=1}^{n} y_{ji}^{(t)} = \bigwedge_{t=1}^{n} \frac{x_i^{(t)}}{x_j^{(t)}} \text{ for } j \in \mathrm{pa}(i), \quad \breve{b}_{ii} = 1, \ \ \breve{b}_{ji} = 0 \text{ for } j \in V\backslash \mathrm{An}(i)$$

and define

$$\widehat{B} = (I_d \vee \breve{B}_0)^{\odot(d-1)}.$$

Properties of $\widehat{B}$

- $b_{ji} \le \widehat{b}_{ji} \le \breve{b}_{ji}$
- One observation may be enough to estimate B exactly.
  Example. In the diamond,
  let $1 \to 2 \to 4$ and $1 \to 3 \to 4$ be both max-weighted.
  If we observe the event

$$\left\{X_2 = b_{12}X_1\right\} \cap \left\{X_3 = b_{13}X_1\right\} \cap \left\{X_4 = b_{24}X_2\right\} \cap \left\{X_4 = b_{34}X_3\right\}$$

  (Z_1 realises all node variables $X_1, \dots, X_4$), then $\widehat{B} = B$. □

- $P(\widehat{b_{ji}} = b_{ji}) \to 1$ geometrically fast as $n \to \infty$.

• $\widehat{B}$ is a generalized MLE

Let P_B be the distribution of the recursive ML vector X.
Then $\mathcal{P} = \{P_B, B \text{ admissible for } \mathbb{D}\}$ is not dominated.

Define a **generalized MLE**[4]: Note that

$$\rho(x, P, P^*) = \frac{dP}{d(P + P^*)}(x)$$

is a density, which always exists as $P(A) + P^*(A) = 0 \Rightarrow P(A) = 0$.

Definition. $\widehat{P}$ is **GMLE of P**, if

$$\rho(x, \widehat{P}, P) \geq \rho(x, P, \widehat{P}) \quad \forall P \in \mathcal{P}.$$

$\widehat{P}$ explains x at least as well as any other distribution of $\mathcal{P}$.

Theorem. [GKL (2018)] $\widehat{B}$ is a GMLE. □

[4] Kiefer and Wolfowitz (1956)

Idea of proof

Let B, B^* be admissible for $\mathbb{D}$. Then we choose as density the measurable function $\mathbb{R}_+^d \to \{0, 1/2, 1\}$ defined as

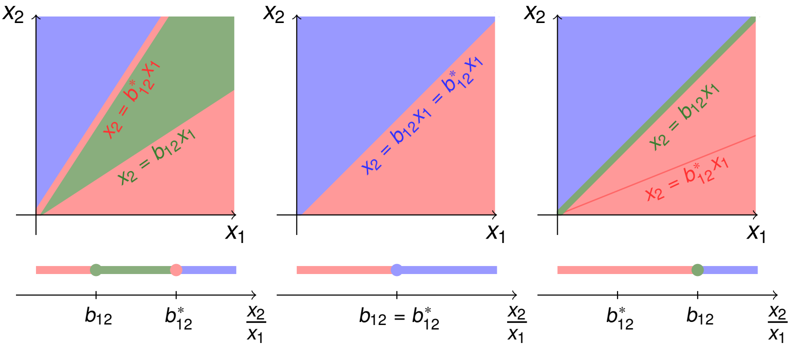
$$x \mapsto \rho(x, B, B^*) := \frac{1}{2}\mathbf{1}_{A_{1/2}(B,B^*)}(x) + \mathbf{1}_{A_1(B,B^*)}(x)$$
$$= \begin{cases} 0 & \text{if } x \in A_0(B, B^*) \\ \frac{1}{2} & \text{if } x \in A_{1/2}(B, B^*) \\ 1 & \text{if } x \in A_1(B, B^*) \end{cases}$$

This is a valid density for $A_0(B, B^*)$, $A_{1/2}(B, B^*)$, $A_1(B, B^*)$ chosen by the fact that $X_i \geq \bigvee_{k \in \mathrm{pa}(i)} b_{ki} X_k$ for $i \in V$.
We obtain for every Borel set $A \subseteq \mathbb{R}_+^d$,

$$\int_A \rho(x, B, B^*)(P_B + P_{B^*})(dx)$$
$$= P_B(A \cap A_{1/2}(B, B^*)) + P_B(A \cap A_1(B, B^*)) \;=\; P_B(A).$$

How to find a density and the GMLE

Let $\mathbb{D} = (\{1,2\}, 1 \to 2)$ and b_{12}, b^*_{12} be admissible for $\mathbb{D}$.
$\mathrm{supp}(X_2/X_1) = [b_{12}, \infty)$ and b_{12} is the only atom of X_2/X_1.



$\rho(\cdot, B, B^*)$ as contour plot (top line) and as function of $y_{12} = x_2/x_1$ (bottom line) for the three possible situations.
The area where it is **0**/$\frac{1}{2}$/**1** is coloured in **red**/**blue**/**green**.

Structure learning for unknown DAG

If the order of nodes/variables is known:
Estimate as before

$$\breve{b}_{ji} = \bigwedge_{t=1}^{n} y_{ji}^{(t)} = \bigwedge_{t=1}^{n} \frac{X_i^{(t)}}{X_j^{(t)}} \text{ for } j \le i, \quad \breve{b}_{ii} = 1, \;\; \breve{b}_{ji} = 0 \text{ for } j > i.$$

Then $\breve{B} = \breve{B} \odot \breve{B}$, i.e. $\breve{B}$ is idempotent, hence[5] is admissible for some DAG $\mathbb{D}$.

Moreover, $P(\breve{b}_{ji} = b_{ji}) \to 1$ geometrically fast as $n \to \infty$.

[5] Butkovic, Cor. 1.6.16

Structure learning for unknown DAG

If the order of nodes/variables is not known:

It suffices for every $i \neq j$ to decide, whether

$$\operatorname{supp}(Y_{ji}) = \operatorname{supp}\left(\frac{X_j}{X_i}\right)$$

has a positive lower bound which is b_{ji} (or upper bound $1/b_{ji}$), and to estimate this bound.

As this bound is an atom, we suggest the following naive algorithm:

Algorithm. [Find an estimate $\breve{B}$ of B from $x^{(1)}, \ldots, x^{(n)}$]

1. For all $i \in V = \{1, \ldots, d\}$: set $\breve{b}_{ii} = 1$.
2. For all $i, j \in V, i \neq j$:

$$\text{if } \#\left\{t : \bigwedge_{s=1}^{n} y_{ji}^{(s)} = y_{ji}^{(t)}\right\} \geq 2, \text{ then conclude } j \in \operatorname{an}(i),$$

$$\text{set } \breve{b}_{ji} = \bigwedge_{t=1}^{n} y_{ji}^{(t)} \quad \text{else set } \breve{b}_{ij} = 0.$$

□

Simulation study[6]

Input: Minimum ratio estimator $\widehat{b}_{ji}$ for all $i, j \in V$.
Output: Topological order

- Greedy algorithms
- Branch & Bound and clever extensions
- Dynamic programming

based e.g. on the ordering of the $\widehat{b}_{ji}$ (and computing time is given)

Simulation study
Random topological order and an Erdös Rényi graph (edge with prob. p, with edge weights $U([0, 1])$, $Z \sim$standard Fr(1).

[6] Esposito, G. (2018) Master Thesis.

Extend model to allow for observational noise

See Poster of Johannes Buck

Also a data analysis

References

- P. Butkovič (2010) *Max-linear Systems: Theory and Algorithms.* Springer.
- Gissibl, N. and Klüppelberg, C. (2018) Max-linear models on directed acyclic graphs. *Bernoulli* **24**(4A), 2693-2720.
- Gissibl, N., Klüppelberg, C. and Lauritzen, S. (2019) Identifiability and estimation of recursive max-linear models. Submitted.
- J. Kiefer and J. Wolfowitz (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, **27**(4):887-906.
- Klüppelberg, C. and Lauritzen, S. (2019) Bayesian networks for max-linear models. In: Biagini, F. and Kauermann, G. and Meyer-Brandis, T. (Ed.): *Network Science - An Aerial View from Different Perspectives.* Springer.
- Lauritzen, S.L. (1996). *Graphical Models.* Oxford University Press.